Published in final edited form as:

J Appl Res Mem Cogn. 2015 December ; 4(4): 368–373. doi:10.1016/j.jarmac.2015.07.007.

Best-classifier feedback in diagnostic classification training

Corey J. Bohil*, Andrew J. Wismer, Troy A. Schiebel, and Sarah E. Williams

Department of Psychology, University of Central Florida, United States

Abstract

Diagnostic classification training requires viewing many examples along with category membership feedback. “Objective” feedback based on category membership suggests that perfect accuracy is attainable when it may not be (e.g., with confusable categories). Previous work shows that feedback based on an “optimal” responder (that sometimes makes classification errors) leads to higher long-run reward, especially in unequal category payoff conditions. In the current study, participants learned to classify normal or cancerous mammography images, earning more points for correct “cancer” than “normal” responses. Feedback was either objective or based on performance of an empirically determined “best” classifier. This approach is necessary because theoretically optimal responses cannot be determined with complex real-world stimuli with unknown perceptual distributions. Replicating earlier work that used simple artificial stimuli, we found that best-classifier performance led to decision-criterion values (β) closer to the reward-maximizing criterion, along with higher point totals and a slight reduction (as predicted) in overall accuracy.

Keywords

Feedback; Diagnosis; Training; Classification; Optimal-classifier

In many diagnostic domains, such as radiology, there exists substantial variability among experts in decision rule adoption despite similar overall accuracy rates (Swets, 1998). Beyond individual differences, diagnostic criterion placement can vary with circumstances. For example, limiting unnecessary biopsies requires a conservative (requiring strong evidence for an “abnormal” judgment) criterion for routine mammograms but a more lenient criterion following a doctor’s referral. To maximize consistency across practitioners, and to afford policymakers some level of control over long run diagnostic accuracy rates and outcomes, it would be beneficial if “optimal” decision criterion placement could be trained.

Diagnostic classification judgments such as those made by dermatologists or radiologists require expertise gained from hundreds of hours of practice (Gunderman, Nyce, & Steele, 2002). Training for this type of classification involves viewing numerous examples from diagnostic categories along with some type of feedback (e.g., correct category membership).

*Corresponding author at: Department of Psychology, University of Central Florida, 4111 Pictor Lane Building 99, Orlando, FL 32816, United States. Tel.: +1 407 823 2755; fax: +1 407 823 5862. corey.bohil@ucf.edu (C.J. Bohil).

Conflict of interest statement

The authors declare that they have no conflict of interest.

This enables trainees to learn a classification decision-rule that maximizes either long-run accuracy or some measure of reward (Maddox, 2002). Although perfect accuracy is desirable, it is often impossible due to the confusability of diagnostic categories. Achieving the best possible long-run performance requires sensitivity to category payoffs (benefits and costs for different outcomes) and base-rates (relative prevalence of diagnostic alternatives). Although base-rates are of critical importance, the current research focuses on category payoffs. We consider base-rates further in Section 5.

Many studies have applied signal detection theoretic (SDT) analysis (or related computational model-based analysis) to examine payoff influence over classification performance (Bohil & Maddox, 2003a; Busemeyer & Myung, 1992; Busemeyer & Rappaport, 1988; Erev, 1998; Maddox & Bohil, 2000; Maddox & Bohil, 2004; Stevenson, Busemeyer, & Naylor, 1991). SDT assumes confusable categories (typically sampled from overlapping Gaussian distributions; see Fig. 1). The category payoff values determine the decision criterion (β) used by an “optimal” classifier (Green & Swets, 1966) that maximizes long-run reward (β_{rew} in Fig. 1), which serves as a benchmark for understanding human performance. Because categories overlap, classification errors are inevitable even for the optimal classifier.

1. Competition between reward and accuracy (COBRA)

With unequal category payoffs, the reward (β_{rew}) and accuracy (β_{acc}) maximizing criterion values are not the same. In fact, using the reward-maximizing criterion actually leads to a reduction in overall accuracy. If responding is biased to favor a higher payoff category then accuracy increases for that category at the expense of the lower payoff category. For example, assuming an equal number of stimuli from each category, overall accuracy is maximized using a decision criterion $\beta = 1$ (β_{acc} in Fig. 1). Given a 3:1 Category 1 to Category 2 payoff ratio, $\beta = 3$ (β_{rew} in Fig. 1) maximizes long-run reward.

Maddox and Bohil (2003, 2004); (see also Bohil & Maddox, 2003a) tested the hypothesis that there is competition between reward and accuracy (COBRA) maximization goals. They argued that learners attempt to maximize payoff on each trial, but erroneously believe that maximizing accuracy is the way to achieve this goal. Observed β values were generally greater than one, indicating sensitivity to the payoff ratio. However, β values were typically “conservative” relative to the reward maximizing criterion. Participants consistently failed to adjust criterion values far enough to achieve maximum long-run reward (this result has been found in many similar studies, e.g., von Winterfeldt & Edwards, 1982).

Maddox and Bohil (2001) noted that the feedback used in classification training often implies a level of performance that is actually unattainable even for the optimal classifier. Due to category overlap, stimuli are often misclassified when they fall into the wrong response region relative to the reward maximizing rule. For example, a stimulus sampled from Category 1’s upper tail may look like a good example of Category 2, and if its percept falls to the right of β_{rew} it will elicit a “Category 2” response. Even though the response would be correct relative to the optimal reward-maximizing criterion, the actual category membership of this stimulus would lead to feedback indicating an incorrect response.

We refer to feedback based on actual category membership as “objective” feedback. Objective category membership is the basis for learning feedback in virtually all classification training. When categories are confusable this feedback is misleading, as it suggests a better decision rule might be found when in fact the learner may already be using the optimal rule.

Maddox and Bohil (2001, 2005); (see also Bohil & Maddox, 2003b) conducted several studies comparing category learning with “objective” versus “optimal” classifier feedback. Optimal classifier feedback following each trial revealed the optimal classifier’s response (see Fig. 3; note that “optimal” is replaced with “best” in Fig. 3 – details below). The “optimal” classifier uses the reward-maximizing decision criterion but still gets many classification responses wrong from the standpoint of objective (actual) category membership. If participants attend to the payoff ratio and attempt to maximize reward (i.e., move their criterion in the β_{rew} direction), then they must learn to sacrifice accuracy, which is maximized by β_{acc} . Feedback (e.g., objective category membership) that contradicts this goal should limit their ability to adjust in the optimal direction. Maddox and Bohil consistently found that optimal-classifier feedback led to performance that was closer to optimal (β values were larger while overall accuracy was reduced).

Optimal classifier feedback decouples the goal of maximizing reward from the strategy of attempting to achieve perfect accuracy (in effort to maximize reward). By deemphasizing accuracy on every trial, this decoupling helps learners adopt a criterion that is closer to the reward maximizing value than is typically achieved after feedback training. We believe this finding may have important practical applications.

2. Best-classifier feedback for mammography training

In Maddox and Bohil’s (2001, 2005); (Bohil & Maddox, 2003b) studies, the categorization stimulus (bar graph height) values were sampled from overlapping normal distributions whose means and variances were determined a priori. These controlled conditions make it possible to determine the optimal classifier’s response to each stimulus. Our goal in the current research was to determine whether this training feedback manipulation could translate to a more complex (i.e., high dimensional; less controlled) stimulus set based on images used in real diagnostic classification training. We examined classification learning with unequal category payoffs, and our stimuli were mammography images showing either presence or absence of cancer. Participants were trained with feedback based either on objective category membership or on performance of an empirically determined “best” classifier. This approach reflects the fact that in real-world diagnostic training the responses of a best performer (however defined) at the task can be used as a training signal for others.

Because we used stimuli drawn from populations with unknown characteristics, we could not derive the performance of an “optimal” classifier for feedback. We instead conducted a small pilot-study using objective feedback and selected the participant whose β was closest to optimal ($\beta_{opt} = 3$ in the pre-study) as a “best” classifier. The study presented in this article used this person’s responses as “best-classifier” feedback during training.

We expected to replicate Maddox and Bohil's (2001, 2005); (Bohil & Maddox, 2003b) earlier findings. Decision criterion values (β) should be closer to the reward maximizing value after training with best-classifier feedback. This should be met with a corresponding decline in overall accuracy rate but an increase in total points.

3. Methods

3.1. Design

We manipulated the type of corrective feedback following each classification response in a mammography screening task. *Objective-classifier* feedback displayed the number of points that could have been earned had the objectively correct response been given (e.g., feedback based solely on category membership of the stimulus, implying perfect accuracy is possible). *Best-classifier* feedback displayed the number of points earned by the best performing classifier (described below) after each trial. Critically, this "best" classifier sometimes gives the incorrect response despite using the correct strategy (i.e., a single decision bound that divides the perceptual space into two response regions). Following an error the best-classifier earns 0 points. This feedback should help participants learn that 100% accuracy is not necessarily their goal and that they can sacrifice some accuracy in order to adopt a criterion that maximizes long-run reward. Because our participants were novices at mammography classification, half were shown a preview image prior to beginning the task (see Fig. 2) to help orient them to the stimulus types. This manipulation had no significant effect on any of our measures and is not considered further.

Each participant completed a single session of classification training with a 3.6:1 payoff ratio (3.6 points for correct "cancer" responses, 1 point for correct "normal" responses). This ratio was selected based on pilot results described below. Each participant completed only one experimental condition. Each session consisted of ten blocks of 50 trials. Blocks 1–9 were feedback training blocks; each response was followed by feedback. Block 10 was a test block with no corrective feedback after responding. Stimulus order was randomized for each block.

3.2. Participants

One hundred and twenty-seven undergraduate students from the University of Central Florida participated in the study for course credit. All participants reported normal or corrected-to-normal vision and were at least 18 years of age. Participants completed one experimental condition during a single 60 min session. Sample sizes were 64 and 63 in the *objective-classifier* and *best-classifier* conditions, respectively. Informed consent was obtained for experimentation with human subjects.

3.3. Stimuli

Stimuli were mammography screening images showing normal and cancer cases. Fig. 2 displays examples of cancer and normal cases (participants viewed only one mammography image per trial). The images were from a digital database of mammography screening examples ("DDSM", n.d.; Heath, Bowyer, Kopans, Moore, & Kegelmeyer, 2001). Normal cases were defined as images from a normal screening exam for which the patient had a

normal follow-up exam at least 4 years later. Cancer cases were defined as images from a screening exam in which at least one pathology proven cancer was found (“DDSM”, n.d.). Classifications used in the database were determined by experienced radiologists.

We selected 100 of these images for use as classification stimuli. These were similar to each other based on the following characteristics. All images reported a breast density value of 2, 3 or 4 on a Breast Imaging Reporting and Data System (BI-RADS) scale (cancer cases on a 5 point scale; normal cases on a 4 point scale). Density values were similar for normal and cancer cases (Normal: $M = 3.00$, $SD = 0.76$; Cancer: $M = 3.28$, $SD = 0.46$). Each mammogram displayed the breast in cranial–caudal (CC) viewpoint. A mixture of left and right breast images were used (60% left in normal cases; 68% left in cancer cases). The subtlety ratings provided by the radiologists for each abnormality present in a cancer image were distributed on a 5 point scale ($M = 3.0$, $SD = 1.4$).

Each cancer image from the database included a small red overlay indicating the region where the abnormality was classified by a radiologist. Since normal cases contained no abnormalities, there were no overlays on these images. We opted to focus participants’ attention to a region of interest in the normal images rather than for them to spend time visually searching the entire image. Therefore we copied an overlay from each cancer image and placed it in a comparable location on a similar (as determined by the experimenters) normal image (Fig. 2 shows four overlays from cancer images and the same overlays placed in a similar location in normal images). Each overlay was relatively round in shape and they were similar in size. The same images were used in both feedback conditions, neutralizing any systematic effects that may have resulted from our placement of overlays on normal images. Our complete image set contained 100 mammograms – 50 showing cancer and 50 normal images. However, only half of these were used in the current study (details below).

3.3.1. Stimulus accuracy rates—To gain insight into the psychological representation of the images, we examined response accuracy for each stimulus from a separate classification study conducted earlier in our lab (currently unpublished data) using these same 100 mammography images. Participants ($n = 92$) completed four training blocks of 100 trials with no payoff manipulation. Observed discriminability ($d' = .40$) and accuracy (57%) were very low so we concluded that the discrimination task was too difficult for novices to learn in a 1-h session.

For the experiment reported below we selected from the pre-study a subset of 50 stimuli (25 from each category) – those whose response accuracy was highest (although many were still difficult to classify). Accuracy for this subset ranged from 53% to 91% correct. The average proportion correct for normal images ($M = 0.68$, $SD = 0.10$) and cancer images ($M = 0.70$, $SD = 0.10$) from this subset were not reliably different from each other, ($t(48) = .798$, $p > .05$).

3.4. Feedback design

Training feedback consisted of points earned for each classification trial and were presented after each response. Correct responses earned either 3.6 or 1 point depending on category, and incorrect responses earned 0 points. Thus, points earned on each trial simultaneously

provided an indication of response accuracy. Total points accumulated were also displayed on each trial. The total-points score was reset to 0 at the start of each block. Fig. 3 displays sample feedback for each condition. Potential Gain (for a single trial) and Potential Point Total (over trials) were the only differences between objective and best-classifier feedback conditions.

3.4.1. Objective feedback—In the objective-classifier condition, feedback was based on actual category membership (cancer or normal). Because an “objective-classifier” knows the correct category on each trial, Potential Point Total in this condition increased on every trial.

3.4.2. Best-classifier feedback—To derive responses of a “best-classifier” to be used as training feedback, a small pilot study was conducted ($n = 8$) using objective feedback and a 3:1 payoff manipulation (correct cancer diagnoses earned 3 points, normal diagnoses earned 1 point). Pilot data were analyzed in terms of decision criterion placement (β) and discriminability (d'). The responses of the “best” performer in this pilot study were to be used as the feedback signal in the main study. The best performer was found to use a criterion very sensitive to the payoff ratio ($\beta = 3.6$, hit rate = 96, false alarm rate = 24) and also had the highest level of discriminability ($d' = 2.46$) of all participants in the pilot study. In other words, this participant was very good at discriminating between cancer and normal images, and used a criterion that sacrificed overall accuracy (due to reduced accuracy on “normal” images) in order to maximize long-run reward (i.e., did not use β_{acc}). Thus, in the best-classifier feedback condition of our main experiment (reported below), feedback for each response was based on the response given by this best-classifier to that stimulus during the pilot study. Responses from the best-classifier’s single best block of trials were used to prevent inconsistencies over stimulus repetitions.

Because the best-classifier feedback was based on the performance of an individual using a criterion value of 3.6, we awarded 3.6 points for a correct cancer classification and 1 point for a correct normal classification in our primary study regardless of whether best-classifier or objective feedback was presented. As a result, the “optimal” criterion was $\beta_{opt} = 3.6$ in all conditions.

3.5. Procedure

Depending on experimental condition, participants received instructions (shown below) explaining how they should think about the feedback. All participants were instructed to maximize their point total.

Best-classifier Feedback Condition—Because these mammograms are not perfect predictors of cancer, perfect performance is impossible. However, the “optimal” strategy is one that maximizes your point total. The feedback you receive after responding will tell you how an “optimal responder” would have done. You should compare your performance with that of the “optimal responder” on every trial. Pay attention to your “actual” and “potential” points on each trial.

Objective Feedback Condition—Because these mammograms are not perfect predictors of cancer, perfect performance is impossible. However, the “optimal” strategy is one that

maximizes your point total. The feedback you receive after responding will tell you not only how many points you earned, but also the points you COULD HAVE earned (whether right or wrong). Pay attention to your “actual” and “potential” points on each trial.

Stimuli were displayed in the center of the screen on a high resolution monitor with a black background (Stimulus size varied somewhat but was close to 400×700 pixels). On each trial, the participant viewed a stimulus (response-terminated; maximum allowed response time = 10 s) and then made a classification response by pressing a key labeled either ‘N’ for normal or ‘C’ for cancer (the q and p keys, respectively).¹ After a classification response, the stimulus was removed and the point feedback appeared (remaining onscreen until a button was pressed). Point totals reset to 0 each block and final point totals were not displayed between blocks. During the 10th (test) block, point feedback was replaced by the words ‘Response recorded’ after each response.

4. Results

We evaluated learning in terms of signal detection measures (hit rate, false alarm rate, discriminability (d'), and decision criterion (β)), point totals and accuracy rates. Average values for these measures are displayed in Table 1. In cases where hit or false alarm rate were 0 or 1 (preventing calculation of meaningful d' and β values), we used a correction method described by Wixted and Lee (n.d.). For $N = 25$ targets (cancer cases), values of 0 were replaced by $1/2N = .02$ and values of 1 were replaced by $1 - (1/2N) = .98$.

Points [$F(9,123) = 32.13, p < .001, \eta_p^2 = .207$], accuracy [$F(9,123) = 29.62, p < .001, \eta_p^2 = .194$], d' [$F(9,123) = 36.63, p < .001, \eta_p^2 = .229$] and β [$F(9,123) = 8.63, p < .001, \eta_p^2 = .066$] values all increased significantly over blocks. There were no significant interactions between feedback type and block (all F s < 1.40).

Of primary interest is the influence of feedback type. As in previous studies with simple artificial stimuli, we expected β values to be larger with best-classifier feedback than with objective feedback (optimal $\beta = 3.6$ given our 3.6:1 payoff ratio). Higher point totals and lower overall accuracy should also result from use of larger values of β .

There was a main effect of feedback type on β values, $F(1,123) = 17.81, p < .005, \eta_p^2 = .066$. Average β with best-classifier feedback ($M = 2.65$) was closer to optimal than with objective feedback ($M = 1.96$). Underlying the difference in β by feedback type are main effects of feedback on hit rates, $F(1,123) = 19.49, p < .005, \eta_p^2 = .137$ and false alarm rates, $F(1,123) = 7.13, p = .009, \eta_p^2 = .116$. Average best-classifier feedback hit rate ($M = .89$) and false alarm rate ($M = .39$) were higher than objective feedback hit ($M = .84$) and false alarm rates ($M = .33$).

There was also a reliable effect of feedback type on point totals, $F(1,123) = 5.35, p = .022, \eta_p^2 = .042$. Average point totals were higher in conditions with best-classifier feedback ($M =$

¹Real diagnostic decisions would not be “normal” or “cancer”, but rather something like “not likely cancerous” or “potentially cancerous”, which could be followed up with a tissue biopsy. Because our stimuli actually displayed cancerous or non-cancerous mammograms, we simplified the task by using “normal” and “cancer” labels.

94.96) than with objective feedback ($M = 92.63$). And although the effect of feedback type on accuracy was not large enough to reach statistical significance, $F(1,123) = .24$, $p = .626$, $\eta_p^2 = .002$, the consistent trend over blocks seems to support our prediction that overall accuracy should be higher with objective feedback. Mean accuracy was higher with objective feedback than best-classifier feedback in 8 of 10 blocks (see Table 1).

The transition from block 9 (classification with feedback) to block 10 (no feedback after each response) caused participants to become more conservative (i.e., focused on accuracy) in their responding. There was a significant main effect of block on measures of β : $F(1,125) = 4.95$, $p < .05$, d' : $F(1,125) = 27.07$, $p < .001$, and accuracy, $F(1,125) = 42.67$, $p < .001$. Nevertheless, the differences in performance based on feedback type still held across blocks 9 and 10 for β , $F(1,125) = 5.57$, $p < .05$. There were no significant interactions between block and feedback type over blocks 9 and 10.

5. Discussion

In this study participants learned to classify normal or cancerous mammography images. Training feedback was based either on objective category membership or on the response of a “best” classifier whose responses matched those of a participant from a pilot study using the same stimulus set (and objective feedback). Our “best” classifier used $\beta = 3.6$, and so our payoff ratio in the current study was 3.6:1.

As in previous studies using simple artificial stimuli and well understood category distributions, observed decision criterion (β) values were closer to optimal in the best-classifier feedback condition (optimal was defined in terms of maximizing long-run reward). Feedback type had no effect on category discriminability (d'). In accord with prior COBRA studies, point totals were higher in the best-classifier condition while overall accuracy was slightly lower relative to the objective feedback condition. These results support the hypothesis that dissociating accuracy- and reward-maximizing goals can foster more-optimal criterion placement.

Although the accuracy and point differences between feedback conditions were small, this may have been due partly to the relative ease of classification in our task. Because this is preliminary work, we had to guess at the level of perceptual discriminability of our stimulus set (see details in Section 3). In the current study observed d' levels increased over blocks to values over 2. This large empirical d' value indicates that the categories were highly discriminable.

A question for follow-up study is the extent to which a more difficult discrimination task will moderate the feedback effect reported here. We suspect that when classification is more difficult (e.g., d' values near 1) the effect of best-classifier feedback may be more pronounced. In fact, the current data are consistent with this prediction. Based on response accuracy for each stimulus (from the unbiased payoff study described in Section 3), we divided our 50 stimuli into “easy” and “difficult” sets. Category payoffs generally had a larger impact on β values for the difficult stimuli ($M = 1.60$) than for the easy stimuli ($M = 1.06$), $F(1,125) = 20.57$, $p < .001$, $\eta_p^2 = .141$. For easy stimuli, there was no significant

difference in average β based on feedback type (objective $M = 1.13$; best-classifier $M = .99$, $t(125) = 1.01$, $p = .158$). But for the difficult stimuli, β values with best-classifier feedback ($M = 1.73$) were reliably larger than with objective feedback ($M = 1.48$), $t(125) = 1.67$, $p < .05$. However, the interaction between feedback type and stimulus difficulty failed to reach statistical significance, $F(1,125) = 2.529$, $p = .11$, $\eta_p^2 = .114$.

Another target for research is to explore other methods for determining best-classifier performance, rather than based on the actual performance of a single best observer. For example, one could determine best-classifier responses based on average accuracy for each stimulus derived from a separate classification study. Such an approach could be used in concert with another recent line of research using idealized training sets to improve diagnostic performance (Giguère & Love, 2013; Hornsby & Love, 2014; Patil, Zhu, Kope , & Love, 2014). The approach used by Love and colleagues reduces the salience of ambiguous cases by removing them from the training set, thereby strengthening learning for the most critical aspects of each category. This would be an interesting approach to explore given that the payoff effects we report here seem actually to be largely contingent upon the ambiguity of some cases.

Another important question concerns the durability of best-classifier feedback training. In the current study, participants appeared to re-focus on accuracy (i.e., β values became less biased) in the final block of trials (when no feedback was presented). Nevertheless, the difference between feedback conditions held, as β values remained larger in the best-classifier condition. Future work needs to explore transfer in more detail, perhaps with longer durations between training and transfer, as well as with more confusable stimuli.

5.1. Interactions between base-rates and payoffs

Although the current research focuses on feedback effects in payoff conditions, another determinant of decision criterion placement is category base-rates (relative prevalence of each category). For example, in the general population cancer is rare compared to the number of normal mammograms viewed each year (Gloeckler Ries et al., 1997). Base-rates actually determine the location of the accuracy-maximizing decision criterion, while base-rates and payoffs combine to determine the location of the reward-maximizing category bound (Bohil & Maddox, 2003a; Maddox & Bohil, 2004).

To utilize best-classifier training it would be sensible to take base-rates into consideration as well as payoffs. Failure to do so could, for example, actually lead to reduced long-run payoff if the error rate were to increase for more-frequent normal cases. However, this possibility seems remote given the many sources of information that reduce uncertainty in medical diagnosis (e.g., the results of a physician's physical exam). And radiologists and physicians no doubt already have a basic appreciation of the costs and benefits of different diagnoses, whether correct or incorrect. Best-classifier feedback might, however, be used in early training for novices, or as a method to adjust or make more uniform the decision criteria used by experienced radiologists (who are known to show variability in criterion placement (Swets, 1998)).

An impetus for the current research is the often-reported difficulty in training observers to adopt a more extreme (i.e., closer to the reward-maximizing) decision criterion (e.g., Healy & Kubovy, 1978; von Winterfeldt & Edwards, 1982). We have argued here and previously that the cause for this is competition between the goals of maximizing reward and accuracy. We believe the current work demonstrates a way to decouple these competing goals and gain increased control over criterion placement across observers. With increased control over diagnostic outcomes, the criterion (and thus tradeoff between accuracy and the outcomes of correct and incorrect responses) becomes an issue of policy.

5.2. Practical applications

Diagnostic classification judgment is the primary task in professions such as mammography, sonography, dermatology, and forensic evaluation. Similar classification judgments are also made in medical and psychological diagnosis. Perfect accuracy is, of course, the most highly desired outcome in these areas. But classification decisions are often made under uncertainty and perfect accuracy is not always attainable given time, financial and other constraints. We make trade-offs between accuracy and outcome payoffs for available choices.

The technique described here – using the performance of a best classifier to train others – could prove useful for moving people to either a lax or conservative criterion depending on the desired long run performance goal (e.g., the desired tradeoff between accuracy and reward). In the current study we defined optimal in terms of long-run reward. The best-classifier feedback was intended to move β values in the direction of the reward maximizing criterion (i.e., to address the persistent problem of conservative cutoff placement that results from objective feedback training). This requires a sacrifice of overall accuracy. If there were a reason to desire a more conservative criterion, best-classifier feedback could be defined in such a way as to achieve this goal as well.

In sum, our results show that basing training feedback on the performance of a single best observer has potential to improve criterion learning. This improvement is based on a definition of “best” chosen by the trainer, rather than on the actual category membership of stimuli. If perfect accuracy is possible given time and cost constraints, then this should certainly be the goal. However, the best-classifier approach takes into account the confusability of categories and the fact that perfect accuracy is often unattainable. This approach may lead to increased control over the outcome of training, however the desired outcome may be defined.

References

- Bohil CJ, Maddox WT. A test of the optimal classifier’s independence assumption in perceptual categorization. *Perception and Psychophysics*. 2003a; 65:478–493. [PubMed: 12785076]
- Bohil CJ, Maddox WT. On the generality of objective vs. optimal classifier feedback effects on decision criterion learning in perceptual categorization. *Memory & Cognition*. 2003b; 31:181–198. [PubMed: 12749461]
- Bussemeyer JR, Myung IJ. An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General*. 1992; 121:177–194.
- Bussemeyer JR, Rappaport A. Psychological models of deferred decision making. *Journal of Mathematical Psychology*. 1988; 32:91–134.

- DDSM: Digital Database for Screening Mammography. n.d.Retrieved from <http://marathon.csee.usf.edu/Mammography/Database.html>
- Erev I. Signal detection by human observers: A cutoff reinforcement learning model of categorization decisions under uncertainty. *Psychological Review*. 1998; 105:280–298. [PubMed: 9669925]
- Giguère G, Love BC. Limits in decision making arise from limits in memory retrieval. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*. 2013; 110(19):7613–7618. [PubMed: 23610402]
- Gloeckler Ries, LA.; Kosary, CL.; Hankey, BF.; Miller, BA.; Harras, A.; Edwards, BK. SEER cancer statistics review, 1973–1994. 1997. NIH publication no. 97
- Green, DM.; Swets, JA. Signal detection theory and psychophysics. Wiley; New York: 1966.
- Gunderman RB, Nyce JM, Steele J. Radiologic research: The residents' perspective. *Radiology*. 2002; 223(2):308–310. [PubMed: 11997529]
- Healy AF, Kubovy M. The effects of payoffs and prior probabilities on indices of performance and cutoff location in recognition memory. *Memory & Cognition*. 1978; 6(5):544–553. [PubMed: 24203388]
- Heath, M.; Bowyer, K.; Kopans, D.; Moore, R.; Kegelmeyer, WP. The digital database for screening mammography; *Proceedings of the Fifth International Workshop on Digital Mammography*; 2001; p. 212-218.
- Hornsby AN, Love BC. Improved classification of mammograms following idealized training. *Journal of Applied Research in Memory and Cognition*. 2014; 3:72–76. [PubMed: 24955325]
- Maddox WT. Toward a unified theory of decision criterion learning in perceptual categorization. *Journal of the Experimental Analysis of Behavior*. 2002; 78:567–595. [PubMed: 12507020]
- Maddox WT, Bohil CJ. Costs and benefits in perceptual categorization. *Memory & Cognition*. 2000; 28:597–615. [PubMed: 10946543]
- Maddox WT, Bohil CJ. Feedback effects on cost-benefit learning in perceptual categorization. *Memory & Cognition*. 2001; 29:598–615. [PubMed: 11504008]
- Maddox WT, Bohil CJ. A theoretical framework for understanding the effects of simultaneous base-rate and payoff manipulations on decision criterion learning in perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 2003; 29:307–320.
- Maddox WT, Bohil CJ. Probability matching, accuracy maximization, and a test of the optimal classifier's independence assumption in perceptual categorization. *Perception & Psychophysics*. 2004; 66:104–118. [PubMed: 15095944]
- Maddox WT, Bohil CJ. Optimal classifier feedback improves cost-benefit but not base-rate decision criterion learning in perceptual categorization. *Memory & Cognition*. 2005; 33(2):303–319. [PubMed: 16028585]
- Patil, KR.; Zhu, X.; Kope, L.; Love, BC. *Advances in Neural Information Processing Systems*. 2014. Optimal teaching for limited-capacity human learners; p. 2465-2473.
- Stevenson, MK.; Busemeyer, JR.; Naylor, JC. Judgment and decision-making theory. In: Dunnette, MD.; Hough, LM., editors. *Handbook of industrial and organizational psychology*. 2nd ed.. Vol. 1. Consulting Psychologists Press; Palo Alto, CA: 1991. p. 283-374. Vol. 1
- Swets, JA. Separating discrimination and decision in detection, recognition, and matters of life and death. In: Scarborough, D.; Sternberg, S., editors. *An invitation to cognitive science. Methods, models, and conceptual issues*. Vol. 4. MIT Press; Cambridge, MA: 1998. p. 635-702.
- von Winterfeldt D, Edwards W. Costs and payoffs in perceptual research. *Psychological Bulletin*. 1982; 91:609–622.
- Wixted, J.; Lee, K. Signal detection theory. n.d.Retrieved from <http://www.kangleelab.com/signal%20detection%20theory.html>

Fig. 1.

β_{acc} = accuracy maximizing criterion; β_{rew} = reward maximizing criterion; x = perceptual effect created by presentation of a stimulus; $l(x|stimulus)$ = likelihood of perceptual effect x given the stimulus.

Fig. 2.

Examples of normal and cancer-containing mammograms. During each trial of the experiment, only a single mammogram was shown in the center of the screen.

You said:		Normal	You said:		Normal
Actual Gain:		0	Actual Gain:		0
Potential Gain:		3.6	Potential Gain:		0
Your Total Points:		26.6	Your Total Points:		30.2
Potential Point Total:		44	Potential Point Total:		34.8

Objective Classifier

Best-Classifier

Fig. 3.

Hypothetical feedback displays for the objective classifier and best-classifier feedback conditions.

Table 1

Signal detection measures, average accuracy, and points by block for objective versus best-classifier feedback conditions.

Block	Hit rate		False alarm rate		β		d'		Accuracy		Points	
	Obj	Best	Obj	Best	Obj	Best	Obj	Best	Obj	Best	Obj	Best
1	.753	.817	.370	.399	1.204	1.660	1.067	1.247	.689	.709	82.58	88.58
2	.828	.880	.316	.389	1.641	2.886	1.523	1.612	.754	.747	90.64	94.50
3	.835	.864	.343	.419	1.669	2.291	1.476	1.417	.745	.723	90.55	92.26
4	.857	.897	.349	.413	2.164	2.812	1.604	1.594	.753	.743	92.38	95.40
5	.846	.892	.337	.413	1.938	2.754	1.581	1.579	.751	.741	91.67	94.98
6	.865	.896	.346	.404	2.138	2.659	1.635	1.639	.758	.747	93.12	95.53
7	.853	.909	.354	.401	2.392	3.318	1.595	1.729	.747	.756	91.89	96.75
8	.860	.901	.329	.392	2.438	2.778	1.720	1.719	.764	.756	93.15	96.31
9	.875	.903	.318	.377	2.234	2.846	1.813	1.761	.775	.764	94.67	96.80
10	.851	.893	.208	.276	1.809	2.547	2.070	2.053	.822	.810	95.33	98.45

Note: Optimal $\beta = 3.6$; *Obj* = objective feedback; *Best* = best-classifier feedback.