First publ. in: Journal of Experimental Psychology: Applied 16 (2010), 1, pp. 1-15
doi: 10.1037/a0018758

DOI: 10.1037/a0018758

User Settings of Cue Thresholds for Binary Categorization Decisions

Assaf Botzer, Joachim Meyer, Peter Bak, and Yisrael Parmet
Ben Gurion University of the Negev

The output of binary cuing systems, such as alerts or alarms, depends on the threshold setting—a parameter that is often user-adjustable. However, it is unknown if users are able to adequately adjust thresholds and what information may help them to do so. Two experiments tested threshold settings for a binary classification task based on binary cues. During the task, participants decided whether a product was intact or faulty. Experimental conditions differed in the information participants received: all participants were informed about a product's fault probability and the payoffs associated with decision outcomes; one third also received information regarding conditional probabilities for a fault when the system indicated or did not indicate the existence of one (predictive values); and another third received information about conditional probabilities for the system indicating a fault, in the instance of the existence or lack thereof, of an actual fault (diagnostic values). Threshold settings in all experimental groups were nonoptimal, with settings closest to the optimum with predictive-values information. Results corresponded with a model describing threshold settings as a function of the conditional probabilities for the different outcomes. From a practical perspective, results indicate that predictive-values information best supports decisions about threshold settings. Consequently, for users to adjust thresholds, they should receive information about predictive-values, provided that such values can be computed.

Keywords: alerts, threshold setting, user adjustment, binary categorization

Decision makers are rarely passive respondents to available information. They actively seek information, filter it to meet their needs and attempt to design an information environment optimally configured to their needs (e.g., Hutchins, 1995). System designers often accommodate users' desires to alter system properties by allowing the adjustment of certain system properties, such as the thresholds for alerts. Although adjustability is a common feature in many systems, it remains unclear how enabling users to adjust system properties will affect system performance. The ability to change properties of the available information may lead to improved performance, as involved users may customize the system to their individual needs and specific situations. However, allowing users to adjust system settings may also result in lower performance when nonoptimal settings are chosen.

For instance, a study on users' settings of system properties in a forecasting system showed that users were more satisfied with system performance when they could alter the information display and the forecast model (Lawrence, Goodwin, & Fildes, 2002). However, overall performance did not improve when users were able to adjust system properties, because some users chose inferior models that led to increased forecasting errors. When users chose accurate models, they performed significantly better than users who received information from identical models and were unable to adjust the system. We need to understand the determinants of users' choices to be able to determine whether and under what conditions it is indeed beneficial to allow users to adjust system properties.

We deal here with the adjustment of system properties in the context of binary cueing systems, such as alarms and alerts. These systems present users with one of two possible output values, usually based on information from sensors measuring the value of a monitored variable and alerting the user when a value crosses a predefined threshold (Lehto, Papastavrou, & Giffen, 1998). Examples of such aids are smoke alarms, collision warning systems in cars or aviation, alarms in intensive care or surgical units, and alerts in automated production. In all these systems, clearly perceptible cues are issued when an event occurs that requires action or should be attended to. The absence of such an event is indicated via a different cue or the lack of a cue. Research on binary cueing systems has shown that operator responses often deviate from the normatively prescribed responses. Response prediction is complex, because responses are affected by the binary cueing system's properties, the context in which the cue is encountered, and the operator (see Meyer, 2004, for a discussion of factors affecting responses to such cues). Only highly reliable systems are clearly beneficial (e.g., Wickens & Dixon, 2007). Nonetheless, well-designed alerts may be of great importance for the safety and performance of the overall system (see, for instance, Bliss, 2003b, for a discussion of these issues in the context of aviation).

The output of a binary cueing system depends on two main factors—the discriminative ability of the system, and its threshold value. The better the system distinguishes between events and nonevents, the more the user will benefit from it. Unfortunately, a

Assaf Botzer, Joachim Meyer, Peter Bak, and Yisrael Parmet, Department of Industrial Engineering and Management, Ben Gurion University of the Negev, Beer Sheva, Israel.

This research was partly funded by an Israel Science Foundation Grant 770/03-37.0 and by the Paul Ivanier Center for Production Research. We thank Yaron Leichter and Anat Meir for their help with conducting the experiments.

Correspondence concerning this article should be addressed to Joachim Meyer, Department of Industrial Engineering and Management, Ben Gurion University of the Negev, Beer Sheva 84105, Israel. E-mail: joachim@bgu.ac.il

decision aid's discriminative ability usually depends on technological constraints, and users can do little to improve it.

However, users or manufacturers can easily adjust the system's threshold. Such adjustments do not entail technological changes, but rather are simple shifts of the value settings beyond which the occurrence of an event, such as a system failure, is declared. A low threshold value will increase detection likelihood, but will also generate frequent false alarms. In contrast, a high threshold value will lead to fewer false alarms, but will also lower the chances of detecting actual events.

In many situations, changes in the environment or the situation may cause the decision makers to prefer higher detection rates over fewer false alarms, or conversely, some reduction in false alarms over higher detection rates. For example, a cook preparing a stir-fry might choose a higher threshold setting for the smoke alarm, but a quality control team might prefer a lower threshold for product damage detection, in order to mitigate risks to the company's reputation following a recent incident in which damaged products were released to the market. Also, in pediatric intensive care units, normal heart rate ranges vary as a function of the child's age (with faster heart rates for smaller children), and applying identical thresholds for all patients may lead to excessive alarms for some. Thus, the ability to adjust a binary cueing system's threshold may lead to improved outcomes for the decision maker.

Lehto et al. (1998) conducted a pioneering study on threshold adjustments in an event detection task, which documented that threshold adjusting participants performed nearly equivalent to participants who used systems with an automatically adjusted optimal threshold. However, approximately 30% of the time, participants adjusted the thresholds in the wrong direction, compared to the normative predictions derived from an expected value maximizing model.

Signal Detection Theory (SDT)

Our study examined possible ways to facilitate adequate threshold choices in event detection tasks, using Signal Detection Theory (SDT; e.g., Green & Swets, 1966; Swets, Dawes, & Monahan, 2000) as a quantitative model to analyze system properties and user responses. SDT is a mathematical model, commonly used to describe binary categorization decisions, in which it is necessary to decide to which of two mutually exclusive categories an event belongs (see, for instance, Macmillan & Creelman, 2005, for a detailed description of SDT). These categories can be a faulty or intact product in a quality control task, a suitable or unsuitable job candidate, or a hazardous or routine situation. In SDT, the two possible states are generally referred to as signal and noise.

The distinction between "signal" and "noise" is considered analogous to deciding whether a data point was sampled from one or the other of two partly overlapping normal distributions, one for noise only and the other for noise with the addition of a signal (signal + noise). The distance between the means of the two distributions, generally measured in standard deviations and denoted as d', defines the detector's sensitivity. The detector receives discrete inputs that are sampled from either of the distributions and compares them to the value of a threshold. If the value of the input exceeds the threshold value, it is classified as a signal, and if it is smaller than the threshold value, it is classified as noise. The threshold is often denoted through the likelihood ratio (β), which is the density ratio of the two normal distributions at the threshold point, as well as the ratio between the probabilities of an event to occur in each of the two distributions (Chi & Drury, 1998; Edwards, Lindman, & Savage, 1963).

In SDT there are four possible outcomes of the decision whether an input constitutes signal or noise: "hit" when a signal is correctly detected, "miss" when a signal is mistakenly identified as noise, "false alarm" when noise is falsely announced to be a signal, or "correct rejection" (CR) when noise is correctly identified. The incorrect identifications (i.e., misses and false alarms) result from the partial overlap between the two normal distributions that the input values are sampled from. An input value may be sampled from the noise distribution but nevertheless exceed the threshold value and be classified as signal (false alarm), or the input value can be sampled from the signal distribution but nevertheless be classified as noise (miss) if its value is lower than the threshold value.

In quality control, for example, a faulty product can be referred to as "signal" and an intact product as "noise." A binary cueing system, or a decision maker, should proclaim the product as faulty if the value of some monitored variable exceeds a predefined threshold. They should proclaim the product as intact if the value of the monitored variable is lower than the threshold. Consequently, a faulty product may be successfully identified (hit), an intact product may be falsely announced as faulty (false alarm), a faulty product may be mistakenly identified as intact (miss), and an intact product may be successfully identified (correct rejection).

Each of the four outcomes in the detection task has monetary or other value for the decision maker. The maximum expected task payoff is achieved when the threshold is set to the optimal threshold setting

$$\beta^* = \frac{1 - p_S}{p_S} * \frac{V_{CR} - V_{FA}}{V_{HIT} - V_{MISS}} \quad (1)$$

where β^* is the density ratio of the two normal distributions at the optimal threshold point, p_S is the probability for a "signal" (e.g., a faulty product) and V_{CR}, V_{FA}, V_{HIT}, and V_{MISS} are the outcome values for correct rejection, false alarm, hit and miss, respectively. These four values can be referred to as the "payoff matrix" (e.g., Meyer, 2001).

Other Rules for Setting the Threshold

The computation of $\beta*$ derives from Bayes model. This model assumes that decision makers consider both probabilities and payoffs when deciding about the threshold, and their goal is to maximize the expected value (EV) of the outcomes. In addition to the Bayesian Model, the literature contains a number of additional decision rules (Chi & Drury, 1998). These rules use parts of the Bayesian rule to compute the threshold settings.

For instance, in the Max-P(C) (maximal probability of correct responses) rule the decision maker attempts to maximize decision accuracy (i.e., the probability for hit plus the probability for CR) by replacing the right part in (1) with 1. Consequently, the density ratio of the two normal distributions at the threshold point will be

$$\beta' = \frac{1 - P_S}{P_S}. \quad (2)$$

With Laplace's decision rule, the decision maker only considers the payoffs for each alternative and chooses the alternative that yields maximum payoff. The decision maker acts as if the possible states of the world are equally probable (Ballestero, 2002), setting a threshold of

$$\beta'' = \frac{V_{CR} - V_{FA}}{V_{HIT} - V_{MISS}}. \tag{3}$$

Wald's decision rule deviates from Laplace in that, for each alternative, it only considers the minimum payoff values and chooses the alternative with the maximum value among the minimum payoffs (Ballestero, 2002), that is, this is a "Maxi-min Strategy." The corresponding threshold will be

$$\beta''' = \frac{V_{FA}}{V_{MISS}}. \tag{4}$$

One may ask which of the decision rules is closest to the way people actually choose a threshold. Research has shown that, similar to the Bayes model, people weigh both probabilities and payoffs when setting a threshold, (e.g., Chi & Drury, 1998; Craig & Colquhoun, 1977; Fox & Haslegrave, 1969; Mobley & Goldstein, 1978). However, they tend to deviate from the Bayes model in that they set thresholds that are lower than high normative thresholds or higher than low normative thresholds. This so called "sluggish beta" behavior is a robust phenomenon in signal detection experiments (Wickens, 1992).

A number of researchers have investigated threshold settings when signal detection tasks are aided by binary cues. In these tasks, binary decisions are based on the value of an observed variable and the indication of a binary cue (e.g., Sorkin & Woods, 1985). Participants should optimally choose different threshold settings, according to the information obtained from the cue, with the difference between the thresholds increasing as a function of the validity of the cue, that is, using identical thresholds in response to both cue outputs when the cue is not valid and using very different thresholds when the cue is valid (e.g., Murrell, 1977). Experiments show that participants are sensitive to cue validities, but they tend to set thresholds that are less extreme than the optimal settings (Maltz & Meyer, 2001; Meyer, 2001; Shurtleff, 1991).

The nonoptimal threshold settings apparent in these experiments may be the result of assessment difficulties regarding the relation between the threshold setting and the expected outcomes. In most existing research on signal detection and aided signal detection, threshold settings were indirectly inferred from the probabilities p_{Hit} and p_{FA}. The decision maker received no indications about the implications of a threshold choice.

Matters are different when people adjust system thresholds. In this case it may be possible to present the conditional probabilities associated with a particular threshold setting, perhaps allowing people to adjust thresholds closer to the optimal settings.

Diagnostic and Predictive Values

Two pairs of conditional probabilities can be computed to describe the properties of a binary cuing system—its diagnostic values and its predictive values. The diagnostic-values of the system are $p_{\mathrm{C|S}}$, the conditional probability of receiving a cue, given that there is a signal (or a threat), or in SDT terms, p_{Hit} of the cueing system, and the conditional probability of receiving an cue when there is no threat ($p_{\mathrm{C|\sim S}}$), or in SDT terms, p_{FA} of the cueing system. In medical settings the diagnostic-values of a system are usually described as the system's sensitivity (p_{Hit}) and specificity ($1\text{-}p_{\mathrm{FA}}$) (Swets & Pickett, 1982). The system's diagnostic-values are stable properties of the detector, as long as the detector, the signal and the noise remain the same. They are independent of the relative frequency of signals.

The two predictive-values of a system are the positive predictive value (PPV), which is the probability for a signal when the system indicates one ($p_{\mathrm{S|C}}$), and the negative predictive value (NPV), which is the probability for there being no threat when the system indicates that there is none ($p_{\mathrm{\sim S|\sim C}}$). In contrast to diagnostic-values, predictive-values depend on relative signal frequency. The predictive-values can be calculated using Bayes's theorem

$$p_{S/C} = \frac{p_{C/S}\,p_S}{p_C} = \frac{p_{Hit}\,p_S}{p_{Hit}\,p_S + p_{FA}(1 - p_S)} \tag{5}$$

for the PPV and

$$p_{\sim S/\sim C} = \frac{(1 - p_{C/\sim S})(1 - p_S)}{1 - p_C} = \frac{(1 - p_{FA})(1 - p_S)}{(1 - p_{FA})(1 - p_S) + (1 - p_{Hit})p_S} \tag{6}$$

for the NPV.

In several studies people were better able to derive predictive-values from diagnostic- values when the latter were presented as natural frequencies, rather than probabilities (e.g., Fahey, Griffiths, & Peters, 1995; Gigerenzer & Hoffrage, 1995; Gigerenzer & Hoffrage, 1999). These findings have implications for domains such as medicine or quality control, where test results need to be interpreted correctly to choose between alternative courses of action. It is not clear, however, how information about these two types of conditional probabilities may affect people's choices of threshold settings for binary categorization decisions.

The standard information about a binary cueing system refers to its diagnostic properties (e.g., the sensitivity and specificity of a test in medicine). As noted above, these properties are the test's stable characteristics, as long as the distributions of the two states remain the same. They are unrelated to the relative frequency of the two states. It is, therefore, relatively easy to provide decision makers with this information.

However, it appears that predictive-values, more than diagnostic-values, affect people's responses to information. Getty, Swets, Rickett, & Gonthier (1995) showed that the PPV of an alerting system affects users' tendency to respond to alerts and their response latency, with responses becoming slower and less likely when the PPV was low. Sorkin (1988) reported a number of cases in which users fail to respond to alarms in response to high rates of unjustified alarms (i.e., low PPV). According to Papastavrou and Lehto (1996) operators will comply more with alerts if they are told the PPV of the system. They will not deem the system nonoperative when it generates frequent false alarms, and they may continue to comply, as they expect the system to be still correct in some cases. It has also been shown that decision makers are less likely to respond to all alarms from a reliable but not always correct alarm system when they are informed about its PPV than when they are not provided with such information (Bliss, 2003a).

The advantage of receiving information about predictive-values is even more pronounced, considering that the computation of normative decisions requires predictive-values, in lieu of diagnostic-values. The decision about a response to presented information should be based on the comparison of the expected value (EV) of each response, given the information. As Meyer (2004) demonstrated, the EV of responding that there is a signal ("S") when a cue is given is

$$\begin{aligned} EV_{\text{"S"}|C} &= p_{S|C} * V_{\text{"S"}|S} + p_{\sim S|C} * V_{\text{"S"}|\sim S} \\ &= PPV * V_{Hit} + (1 - PPV) * V_{FA} \end{aligned} \tag{7}$$

and the EV of responding that there is no signal ("~S") when a cue is given is

$$\begin{aligned} EV_{\text{"}\sim S\text{"}|C} &= p_{S|C} * V_{\text{"}\sim S\text{"}|S} + p_{\sim S|C} * V_{\text{"}\sim S\text{"}|\sim S} \\ &= PPV * V_{Miss} + (1 - PPV) * V_{CR} \end{aligned} \tag{8}$$

a decision maker should respond "S" when a cue is given if

$$EV_{\text{"S"}|C} > EV_{\text{"}\sim S\text{"}|C} \tag{9}$$

which means that

$$PPV * V_{Hit} + (1 - PPV) * V_{FA} > PPV * V_{Miss} + (1 - PPV) * V_{CR} \tag{10}$$

Thus, given a specific payoff matrix, the decision whether to respond when a cue was given depends on the system's PPV. A parallel computation applies to the response in the no-cue situation. In this case, the decision is based on the NPV. Thus, both theoretical considerations and empirical findings suggest that predictive-values, rather than diagnostic-values, are the more useful information for human decision makers when working with binary cueing systems. We therefore hypothesize that information about predictive-values should enable decision makers to set closer to optimal thresholds.

To test our hypothesis, we conducted two experiments in which participants performed a simulated quality control task with the aid of binary cues. Participants had to decide whether to produce or not to produce an item based on whether it was cued red (faulty) or green (intact). Participants' sole information source was the binary cues. Participants set the cue thresholds and we evaluated the determinants and consequences of their settings.

Experiment 1

In Experiment 1, participants were asked to decide whether to produce a product or discard it as faulty, based on the indications from a binary cue. At the beginning of each product batch, the participants were informed of the payoff matrix for the specific batch (i.e., the costs of false alarms and missed detections and the benefits from hits and correct rejections) and the probability of malfunctions in the batch. They then set the initial threshold for this batch and could readjust it after every 10 products in a batch.

To assess the effects of different information types on performance, we divided the participants into three groups. All participants received information about malfunction probability and the payoff matrix. For the first group, this information was all they received. We provided the second group with diagnostic-values information ($p_{A|S}$ or p_{Hit}, $p_{A|\sim S}$ or p_{FA}, 1-$p_{A|S}$ or p_{MISS}, and 1-$p_{A|\sim S}$ or p_{CR}), and the third group with predictive-values information ($p_{S|A}$ or PPV, 1-$p_{S|A}$ or 1-PPV, $p_{\sim S|\sim A}$, or NPV and $p_{S|\sim A}$ or 1-NPV).

This experiment strived to answer a series of questions: First, what is the proximity of participants' threshold settings to optimal settings, and do the participants' settings differ as a function of available information? Second, if systematic deviations from the optimal settings exist, do such deviations correspond to any of the decision rules described above? Finally, how difficult is the threshold setting process for the participants? Conceivably, all participants may set similar thresholds, but particular information may ease the process.

Method

Participants. Participants were 72 undergraduate industrial engineering students (approximate age distribution in the range of 20 to 35 with a mean of 26 and a standard deviation of 2.06, approximately 40% were women) who received a course bonus point for participating in the experiment. In addition, participants could gain monetary rewards up to 32 NIS (New Israeli Shekels) (~US $7), according to their total experimental score. Participants were industrial engineering students familiar with quality control, so the domain was not entirely unfamiliar to them, although the experimental task was a simulation.

Experimental design. This experiment tested four independent variables. The presented information (predictive-values, diagnostic-values, none) was a between-subjects variable, and the three variables—the probability of a malfunction, the payoff matrix and the block number were within-subjects variables. The two levels—the payoff matrix and the probability for a faulty product (i.e., a signal)— were designed so that one would prompt the setting of a high threshold value, while the other would prompt the setting of a low threshold value.

In both levels of the payoff matrix $V_{CR} = V_{Hit} = 0$. V_{Miss} and V_{FA} were either -50 and -120 or -120 and -50, respectively. Consequently there were two different payoff matrices:

$$V1 = \frac{V_{CR} - V_{FA}}{V_{HIT} - V_{MISS}} = \frac{0 - (-50)}{0 - (-120)} = 0.416$$

and

$$V2 = \frac{V_{CR} - V_{FA}}{V_{HIT} - V_{MISS}} = \frac{0 - (-120)}{0 - (-50)} = 2.4. \tag{11}$$

We will refer to the payoff matrix V1 as the "costly misses" payoff matrix and to the payoff matrix V2 as the "costly false alarms" payoff matrix. A practical example of a costly misses payoff matrix in the quality control domain is when the cost of dealing with a faulty product, if delivered to the market, is much greater than the financial loss of discarding an intact product. In the costly false alarms payoff matrix, discarding an intact product is much more costly than delivering a faulty product to the market. The zero values for the other two outcomes (i.e., hit and correct rejection) are quite common if the payoff matrix is constructed according to the responsibilities assigned to quality control teams. They are often not rewarded for detecting faults (hits) or for successful production (CR) but are responsible for negative outcomes (i.e., false alarms and misses).

The two different probability levels for a faulty product (i.e., signal) were $p_{S1} = 0.15$ and $p_{S2} = 0.85$. Four combinations resulted from crossing the two levels of the payoff matrix with the two levels of the faults probability. Each combination generated one batch of 40 products, for which participants decided whether to produce or not, based on a binary cue indicating the quality of each product. Participants received no additional information about the quality of each product, besides the binary cue. Participants could adjust the threshold of the binary cue after they classified 10 products. Therefore, 10 products constituted a block in the experiment, and the independent variable "block number" had four levels (4×10). Table 1 presents the optimal threshold setting for each of the four combinations. As β is a ratio, we used the natural logarithm $\ln\beta$ to compute negative and positive threshold values, relative to the origin.

Instruments and procedure. A MSVisualBasic program was developed for the experiment. Experimental sessions were conducted with groups of up to 20 participants, each assigned to a personal computer. Participants were randomly assigned to one of the three experimental conditions, and they completed an on-screen basic demographic questionnaire and read the instructions before beginning the experiment.

The experiment consisted of a binary categorization task (described as the decision whether or not to manufacture) based on indications from a cueing system. The system supposedly sampled raw material density, comparing it to a threshold value. The system then indicated the product as possibly faulty by presenting a cue if the density was higher than the threshold value, or indicated possible product intactness by presenting a different cue when the density was lower than the threshold value. Participants needed to set the threshold value for the cue. The payoff matrix and probability for a signal for each batch of 40 products appeared at the top of the screen (see Figure 1).

Participants could set the threshold by pressing the right mouse-button to mark a blue line on a white bar, and then using the mouse again to press the screen button "set position." This white bar was 17.7 cm wide and 0.7 cm high and was composed of 800 pixels for 800 values of $\ln\beta$ between -4 and 4 in steps of 0.01. The word *high* at the right end of the bar indicated high density values, while the word *low* at the left end of the bar indicated low density values.

After the participant set a threshold, the system supposedly began to sample products and measure their density. The binary cue was displayed in a 3.7 cm wide and 1.2 cm high rectangular field, which was red to indicate a fault when the threshold was exceeded and green to indicate an intact product when the threshold was not exceeded. Participants decided whether to continue production by pressing the "continue" button or to cancel production by pressing the "stop" button. After a decision, the cue disappeared, the participant received feedback on the correctness of the latest decision, and the number of points gained or lost from the last decision was shown, together with the cumulative payoff for the current block. The notice remained visible for 4 s, after which the next trial began.

Table 1
Optimal Threshold Values for the Four Combinations of the Payoff Matrix and the Probability for a Malfunction (p_S)

Probabilities	Costly misses	Costly FA
$p_S = 0.15$	0.9	2.6
$p_S = 0.85$	−2.6	−0.9

Note. All threshold values are in $\ln\beta$ units. FA = false alarms.

Participants could select a new threshold value after a block of 10 products. The three experimental conditions differed in the information participants received when evaluating threshold values, with participants either receiving no additional information, seeing the predictive-values or seeing the diagnostic-values associated with the thresholds they were considering (see Figure 2). There was no time limit to choose a threshold, so participants could consider their choices and even make calculations prior to setting the threshold.

The experiment consisted of five batches of 40 products each. The settings of the first batch were identical for all participants with $p_S = 0.5$ and $V = 1$ (i.e., the same costs for misses and false alarms). This batch served as practice, and its data were not analyzed. The following four batches were the four combinations of p_S and V. For each information condition, each participant received one of the 24 possible sequences of the four combinations.

Researchers instructed participants to strive to maximize the total score and emphasized that they should ask if they suspect they did not fully understand any detail in the instructions. All actions, displayed values and events were recorded in a time-stamped log file, from which the different dependent variables for the data analyses were computed.

Results

The analysis of the results is divided into two parts. First, we present the participants' results in terms of scores and threshold settings. We then present the degree to which the threshold settings correspond with the different models. We provide the value of the statistic, the mean square error (*MSE*), the *p* value and the effect size (partial η^2) of every statistical test. For within-subject effects we applied Greenhouse-Geisser corrections for violations of sphericity. Prior to the analysis of the results we identified outliers, defined as participants for whom at least one mean threshold in one of the batches deviated by more than 2.5 standard deviations (*SD*) from the mean threshold value for this batch in the specific information condition. In practice, such a deviation indicated that the participant had set the threshold for the entire batch at one of the ends of the scale and failed to adjust it during the batch. We view such an extreme choice as indicative of low task involvement. Three outliers, one from each experimental group, were identified and excluded from the analyses.

Score. A four-way analysis of variance was performed on the average score for a product. All scores are negative, since participants could only minimize their losses to maximize their payoff. The average score for a product across all experimental conditions was -10.01 ($SD = 10.56$) and there were differences in the average score for a product between the different information conditions. The average score with predictive-values information was -8.67 ($SD = 9.11$), while the average scores with the none- and diagnostic-values information were -11.75 ($SD = 12.26$) and -9.87 ($SD = 9.78$) respectively, $F(2, 66) = 5.15$, $MSE = 171.82$,

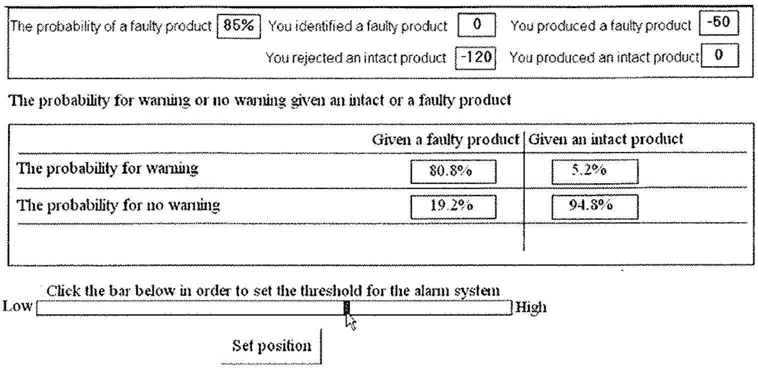

Figure 1. A full screen layout. The top panel specifies the payoff structure and the probability for a faulty product for a current batch of 40 inspected products. The middle panel presents the conditional probabilities associated with the currently chosen threshold. Participants selected thresholds on the lower panel's bar until they chose one by clicking on the "set position" button below the bar.

p = .009, partial η^2 = 0.13. Using a Ryan correction, only the difference between the predictive-values-display and the none-display was significant, $p < .05$.

Threshold values. Threshold settings were analyzed with a four-way analysis of variance. Participants chose higher threshold values when false alarms were more costly than when misses were more costly, as required by the normative decision making model, $F(1, 66) = 4.68$, $MSE = 12.65$, $p = .03$, partial η^2 = 02. As evident from Table 2, participants set higher threshold values when $p_S = 0.15$ than when $p_S = 0.85$, $F(1, 66) = 66.06$, $MSE = 15.30$, $p < .001$, partial $\eta^2 = 0.5$. Figure 3 shows that the mean differences between the threshold values for different p_S values increased slightly over time, in accordance with the normative model, $F(1.755, 115.831) = 3.60$, $MSE = 2.67$, $p = .01$, partial $\eta^2 = .05$, for the interaction between block and p_S. We tend to treat this finding as an informal manipulation check for the understandability of our task. Such a consistent change in threshold choices could not be evident if participants did not understand the task.

Figure 4 shows that regardless of the payoff matrix, the difference in threshold values between the two probabilities for fault was largest with predictive-values information and was close to the difference between the optimal threshold settings. Threshold settings with the none- and the diagnostic-values information were much less affected by the prior probabilities for a fault, $F(2, 66) = 13.36$, $MSE = 15.30$, $p < .001$, partial $\eta^2 = 0.29$, for the two-way interaction p_S * Display. The three-way interaction p_S * Display * Payoff Matrix was not significant.

Analysis according to the models. In the introduction we presented the Bayes model and three alternative decision rules. For reasons of simplicity, we will refer to decision rules and models as models. To test the degree to which participants' threshold settings corresponded to each of the models, we computed Pearson correlations between a participant's four average lnβ values and the four lnβ values computed according to each model, and we transformed the correlations to z-scores with the Fischer transformation. These z-scores served as dependent variables for an analysis

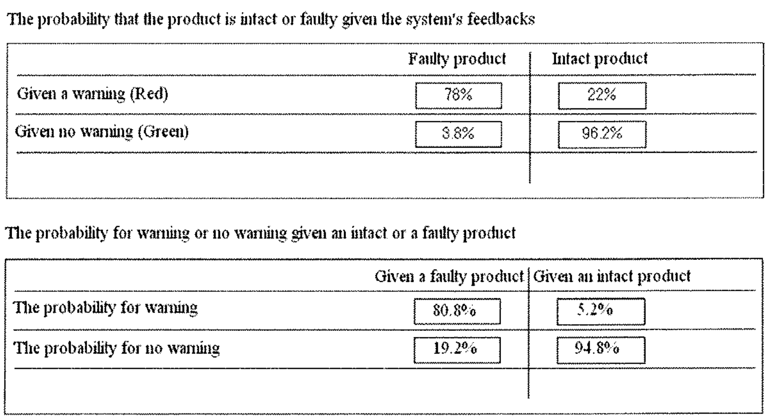

Figure 2. The predictive values information (upper panel of the figure) and the diagnostic values information (lower panel of the figure). Participants saw only one type of information.

Table 2
Mean Threshold Values and (Standard Deviations) for the Two Different Probabilities and the Two Different Payoff Matrices in Experiment 1

Probability		Payoff matrix	
$p_S = 0.85$	$p_S = 0.15$	Costly misses	Costly FA
0.72 (2.3)	1.19 (1.92)	0.0016 (2.36)	0.46 (2.21)

Note. All threshold values are in lnβ units. FA = false alarms.

of variance with the model and the block as within-subject variables and the information as a between-subjects variable.

Because in this experiment the rewards for hit and correct rejection were always 0, the predictions from the Wald model and the Laplace model are identical. So we computed only one correlation, hereinafter referred to as the *Wald-Laplace* model.

The degree of fit to the models increased within each batch, over time, with mean z-scores of 0.46 (SD = 1.12), 0.59 (SD = 1.06), 0.82 (SD = 1.71), and 0.82 (SD = 1.69) for the four blocks, respectively, $F(1.594, 105.206) = 6.49$, $MSE = 1.98$, $p = .004$, partial $\eta^2 = .09$. This finding also serves as an indication for participants' understanding of the task.

The information affected the correspondence of threshold choices with the different models, as is evident in the significant interaction Information * Model, shown in Figure 5, F (2.752, 90.801) = 5.86, $MSE = 4.90$, $p = .005$, partial $\eta^2 = 0.15$. With the none- and the diagnostic-values information participants did not behave according to a specific model. With the predictive-values information, behavior was most compatible with the Max-P(C) model, less compatible with the Bayes model and unrelated to the Wald-Laplace model. Thus, only in the predictive-values condition did participants act according to one model, namely the Max-P(C) model, in which only probabilities affect threshold settings.

Time and considerations. Differences between the information conditions may possibly arise because participants devote more time to the task when receiving certain types of information. Alternatively, while performance levels may be the same, more or less effort may be required to attain these performance levels with different types of information. The time required to set the thresholds and the number of examined threshold settings prior to threshold choice may indicate the level of effort participants invested in the task. We analyzed these variables with four-way analyses of variance.

In the analysis of the time, only the main effect of the block was significant, $F(1.582, 104.395) = 26.94$, $MSE = 155.95$, $p < .001$, partial $\eta^2 = 0.29$. The time participants invested in threshold decisions gradually decreased, with 16.66 s (SD = 14.17), 12.89 s (SD = 7.93), 10.84 s (SD = 5.19), and 10.45 s (SD = 5.32), for the four blocks, respectively. Participants apparently developed their response strategy during the first block, in which they devoted much time to deciding how to set the threshold. In later blocks, they used the strategies and devoted less effort to deliberations. A notable finding is the lack of an effect of the information condition—the time was the same for the none-condition as for the two other conditions. This finding indicates that performance differences between the information conditions were not due to participants in the none-condition investing less effort in the task.

The analysis examining the number of considerations, that is, the number of threshold changes prior to setting a threshold for a certain block, showed fewer considerations when no conditional probabilities were presented than in the two other conditions, and this difference was largest in the first block, as indicated by main effects of the information, $F(2, 66) = 21.63$, $MSE = 81.87$, $p < .001$, partial $\eta^2 = 0.4$, and of the block, $F(1.305, 86.11) = 99.71$, $MSE = 84.81$, $p < .001$, partial $\eta^2 = 0.6$, and a significant interaction Block * Information, shown in Figure 6, F (2.609, 86.11) = 24.76, $MSE = 84.81$, $p < .001$, partial $\eta^2 = 0.43$. This finding is reasonable, because participants only gained information from considering different thresholds when they received predictive- or diagnostic-values information, which could help form a strategy.

The number of considerations with the predictive-values information was significantly larger than with the diagnostic-values information ($p < .05$ with a Ryan correction). Possibly, the

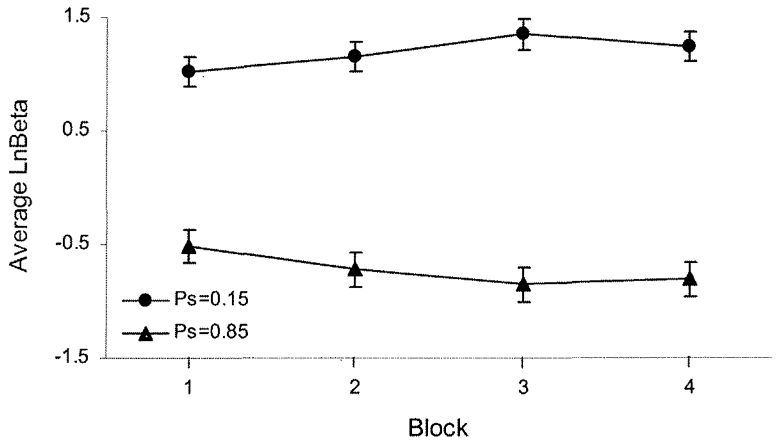

Figure 3. Threshold values in Experiment 1 for different probabilities in the four experimental blocks. Vertical lines depict standard errors of the means.

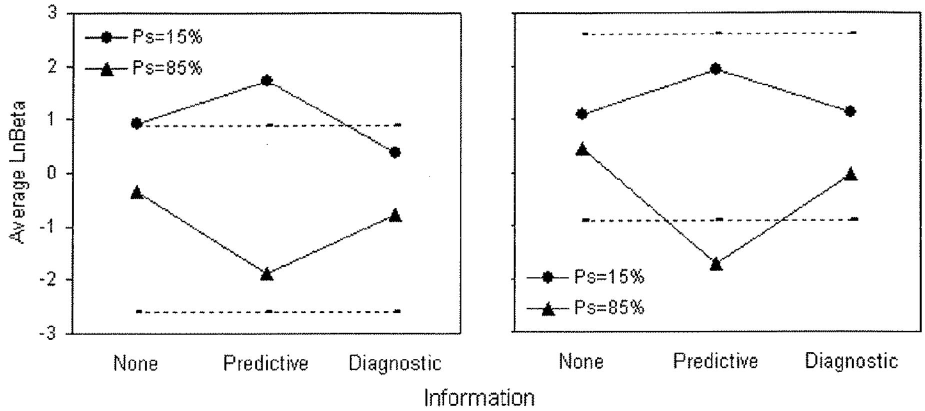

Figure 4. Threshold values for different probabilities in Experiment 1 when misses were costly (left panel) and when false alarms were costly (right panel) for the three information conditions. The dashed lines represent optimal threshold values.

predictive-values information was easier to understand than diagnostic-values information, and participants therefore examined a larger number of alternative threshold values when they received information on predictive values.

Agreement. Other than the binary cue, participants had no additional information on which to base their decision about a product. Therefore, in essence, they should have always followed the cue to maximize their score, unless the payoffs or probabilities are so extreme that one should always respond identically, regardless of the cue (e.g., always discard a product or always approve it). As the probabilities we used were far from 1 or 0 and the between-outcome value ratios in the payoff matrices were far from 0 or infinity, participants in the experiment should have always followed the cues.

To test whether participants' agreement with the cues depended on the different independent variables, we analyzed their agreement percentages (i.e., the percent of times participants complied with the binary cue). We combined the results from all four blocks into one mean value to avoid missing values in cases where one kind of feedback was not provided in a specific block (e.g., no green cues or no red cues).

Figure 7 shows that when the cue indicated a fault (red), participants complied with the cue more when $p_S = 0.85$ than when $p_S = 0.15$, and when the cue indicated an intact product (green), participants complied with it more when $p_S = 0.15$ than when $p_S = 0.85$. The fact that this interaction did not depend on available information shows that participants did not need to be directly informed of the PPV and NPV values to realize that PPV will be greater when $p_S = 0.85$ and NPV will be greater when $p_S = 0.15$. Still, this pattern does not reflect normative decision making, because participants should have always followed the cue, regardless of fault probability.

A four-way analysis of variance (Cue $\times$ V $\times$ p_S $\times$ Information) showed that the interaction Cue $\times$ p_S, depicted in Figure 7, was significant, $F(1, 59) = 55.5$, $MSE = 0.02$, $p < .001$, partial $\eta^2 = 0.49$. Notably, there was no significant effect of information on participants' agreement with the cue. Participants' agreement with the system was very high across all displays (0.90–0.95), and they apparently understood quite well that they needed to follow the cue.

Discussion

The experiment tested participants' binary cue threshold settings in response to different fault probabilities and payoff matrices. As predicted from the normative model, there was a general tendency to use higher thresholds when p_S were smaller and false alarms

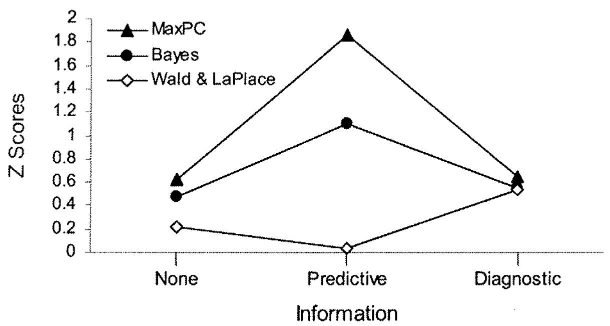

Figure 5. The degree of fit between participants' responses in Experiment 1 and the three models for the three information conditions.

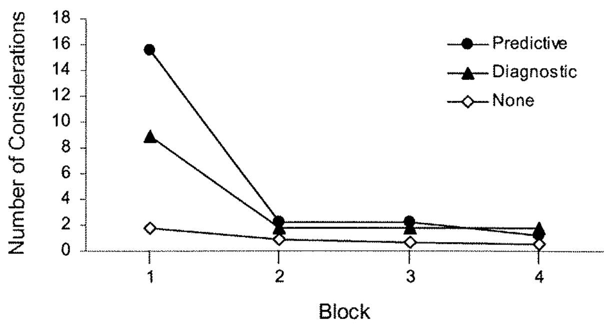

Figure 6. The number of considerations for each information condition in the four blocks.

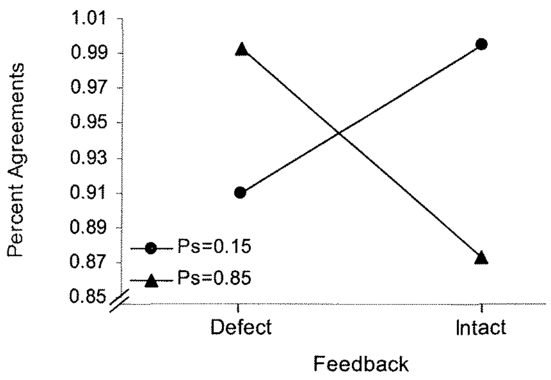

Figure 7. Percent agreements with the different feedbacks (red or green) as a function of the probability for a fault.

were more costly than misses. This finding is in line with previous studies, which found both probabilities and payoffs to affect people's threshold settings (e.g., Chi & Drury, 1998; Craig & Colquhoun, 1977; Fox & Haslegrave, 1969; Mobley & Goldstein, 1978).

However, participants were not sufficiently responsive to the changes in the payoff matrix. The thresholds for the different payoffs differed by less than 0.5, while the optimal thresholds should have shifted from 0.9 to 2.6 and from −2.6 to −0.9 for $p_S = .15$ and $p_S = .85$, respectively. This tendency was not affected by the available information. Thus, in line with previous literature, there was some indication for a "sluggish beta" phenomenon with respect to payoffs (Wickens, 1992). It seems that participants failed to assign sufficient weight to the payoffs when adjusting the thresholds, and experience did not help them converge toward the optimal weight.

The average adjustment as a function of p_S was also smaller than the optimal adjustment, but here, a difference was apparent among the three information conditions. With predictive-values information, the adjustment approached the optimal adjustment, while with the other two types of information the adjustment was much smaller, indicating that with them it was difficult to determine the appropriate threshold for different probabilities. Thus our hypothesis about the benefit from receiving predictive-values information was partly supported. Participants who received such information approached optimal adjustments in response to changes in probabilities, but they were not close to optimal adjustments in response to changes in payoffs.

When comparing the results with the predictions from the different models, the only strong correlation was between the responses in the predictive-values condition and the Max-P(C) model, that is, the model according to which people respond only to the probabilities and ignore the payoff values. None of the models were clearly correlated with the responses in the two other information conditions.

The differences between the conditions cannot be attributed to differences in the effort invested in making the decisions, because the times required for setting thresholds were similar in all three information conditions. We did find that participants who received information about conditional probabilities considered more thresholds in the first experimental block than participants who received no such information. This finding implies that when a new combination of payoff and probability was shown, participants who received information about conditional probabilities attempted to use this information to form a strategy for threshold selection. This strategy was then applied to the following three blocks with the same payoff and probabilities.

Both the analysis of the mean threshold in the different conditions and the analysis of individual correlations with the different models yielded similar results. In the predictive-values condition, participants relied strongly on the prior probabilities when choosing the thresholds, and their responses were close to optimal. In the other two conditions, there was clear evidence of a "sluggish beta" phenomenon. Participants adjusted the thresholds in the correct directions, but failed to do so sufficiently.

Experiment 2

In Experiment 2 we reexamined the determinants of participants' threshold choices as a function of the available information, and we aimed to determine whether participants indeed considered probabilities and payoffs in their threshold setting. To do so, we selected a somewhat different approach than the one taken in Experiment 1. Here we defined a "basic combination" of p_S and the payoff ratio V, which optimally required a high threshold setting. We then defined two altered combinations by either changing p_S or V, so that the optimal threshold setting for both would be the same (and, of course, different from the optimal setting in the basic combination).

There are three extreme possibilities of results. First, participants may be able to set their thresholds according to the optimal threshold settings. In this case, the threshold for the basic combination should be different from the thresholds for the two other combinations, which should not differ from each other. Alternatively, threshold settings may depend only on the probabilities. In this case, the threshold settings for the basic combination and the altered payoff combination should be identical (because they both have the same p_S) and the threshold for the altered probability combination should differ. Finally, threshold settings may depend only on the payoffs. In this case, the basic and altered probability combinations should lead to identical threshold settings and the altered payoff combination should show a different threshold. Of course, intermediate results are also possible, and they are actually expected, considering the results from Experiment 1.

Method

Participants. Fifty-seven undergraduate industrial engineering students (approximate age distribution in the range of 21 to 43 with a mean of 26 and a standard deviation of 2.29, approximately 33% were women) participated in the experiment in partial fulfillment of an Ergonomics course requirement. As motivation, they received monetary rewards up to 18 NIS (~US $4), according to their total experimental score.

Experimental design. This experiment tested three independent variables. The type of information (predictive-values, diagnostic-values, none) was a between-subjects variable. The two other independent variables were: (1) the combination of the fault probability and the payoff matrix, and (2) the block number, where 10 products constituted a block. Both were within-subjects variables.

Three probabilities and payoffs combinations existed in this experiment: the basic combination of payoff and probability had $p_S = .3$ and $V = 2.4$, and it required an optimal threshold of $\ln\beta^* = 1.722$. The two other combinations were chosen by either altering the probability ($p_S = .85$; $V = 2.4$) or the payoff ($p_S = .3$; $V = 0.18$). The latter payoff ratio was constructed by making misses more costly than false alarms (i.e., -50 and -9, respectively) and assigning a zero value to the two other possible outcomes. $V = 2.4$ was constructed in the same way as in Experiment 1. The optimal threshold setting for both combinations was $\ln\beta^* = -0.86$. We will refer to the three combinations from this point on as the "basic combination" the "different probability combination" and the "different payoff combination." The order in which the combinations were presented was counterbalanced, yielding six possible orders to which participants were randomly assigned.

Instruments and procedure. The experimental software and procedure were identical to those utilized in the first experiment, except that after the experiment, we also asked participants how they decided to set the threshold. This question was intended to provide some information on the way participants approached the threshold setting task. The relation between the number of points gained and the payment was the same as in Experiment 1.

Results

We again analyzed the score, the threshold values settings, the degree to which the results correspond with the different models, the time required for setting a threshold, the number of considerations, and the agreement with the cues. We used the same technique and criteria to identify outliers as in the first experiment, omitting the results of two participants in the predictive-values group and one participant in the diagnostic-values group from the analyses of all dependent variables. The independent variables in the analyses were (1) the three combinations of p_S and V, which constituted three levels of the independent variable "combination"; (2) the three information conditions—none, predictive and diagnostic; and (3) the block. We applied Greenhouse-Geisser corrections to all within-subject effects.

Score. Participants' scores differed among the three types of information, $F(2, 51) = 5.85$, $MSE = 264.52$, $p < .01$, partial $\eta^2 = 0.19$. As in Experiment 1, the predictive-values information led to the best performance (-7.48, $SD = 8.44$), followed by the diagnostic-values information (-9.5, $SD = 11.43$) and the none information (-12.76, $SD = 12.84$). A Ryan test indicated a significant difference between the predictive- and diagnostic-values conditions and the none condition ($p < .05$). The information conditions differed most in the basic $p_S = 0.3$, $V = 2.4$ combination. The interaction between the combination and the information, depicted in Figure 8, was significant, $F(3.295, 84.01) = 2.86$, $MSE = 217.77$, $p = .04$, partial $\eta^2 = 0.1$.

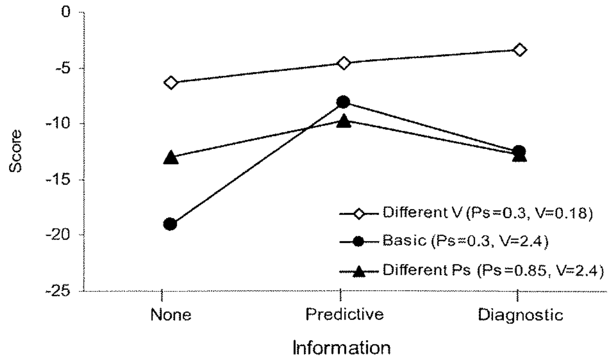

Figure 8. Average score for a product in Experiment 2 in the different combinations of p_S and V for the three information conditions.

Threshold values. Participants in all three conditions set higher thresholds for the basic combination than for the other combinations (see Figure 9), $F(2, 102) = 10.86$, $MSE = 11.74$, $p < .001$, partial $\eta^2 = 0.18$. The difference between the combinations was affected by the information, as indicated by a significant Information * Combination interaction, $F(4, 102) = 2.72$, $MSE = 11.74$, $p = .03$, partial $\eta^2 = 0.1$. In the none and the diagnostic information conditions, participants chose similar thresholds for both the altered-probability and the altered-payoff combinations, and these settings were higher than the optimal setting, that is, their adjustment of threshold settings, compared to the setting in the basic condition, was insufficient. In the predictive information condition, participants set a higher threshold in the basic and the altered-payoff conditions, although the latter was lower than the former. The altered-probability condition led to much lower threshold settings, even below the optimal threshold. Thus, participants in this condition were clearly sensitive to the probability.

Analysis according to the models. We again computed the correlations between participants' individual threshold settings and the thresholds computed with each of the three models. The correlation coefficients were transformed to z-scores using the Fischer transformation. The z-scores then served as dependent variables in a three-way analysis of variance (Model * Information * Block) of the fit between participants' responses and each of the models. Figure 10 shows that the fit between threshold settings and models depended on the information, $F(3.09, 78.805) = 3.79$, $MSE = 4.21$, $p = .006$, partial $\eta^2 = 0.13$. As in Experiment 1, the thresholds in the predictive-values condition were relatively strongly correlated with the Max-P(C) model, less correlated with the Bayes model and unrelated to the Wald-Laplace model. Threshold settings in the two other information conditions were not strongly correlated with any of the models.

Time and considerations. A three-way analysis of variance was run on the time required to set a threshold value. As in Experiment 1, only the main effect of the block was significant, $F(2.146, 109.439) = 8.02$, $MSE = 111.35$, $p < .001$, partial $\eta^2 = 0.14$, with 16.20 s ($SD = 12.15$), 12.96 s ($SD = 7.919$), 12.66 s ($SD = 10.25$), and 11.54 s ($SD = 8.14$) for the four blocks, respectively. Again, during Block 1, participants spent considerable time determining how to perform the experimental task and employed the strategies developed in the first block in subsequent blocks.

A three-way analysis of variance on the number of considerations yielded main effects of the block, $F(1.445, 73.697) = 47.17$, $MSE = 63.69$, $p < .001$, partial $\eta^2 = 0.48$, and the information, $F(2, 51) = 23.53$, $MSE = 23.56$, $p < .001$, partial $\eta^2 = 0.48$, and a significant interaction Block * Information, $F(2.89, 73.697) = 10.85$, $MSE = 63.69$, $p < .001$, partial $\eta^2 = 0.3$. Figure 11 shows that, as in Experiment 1, in the first block participants used the

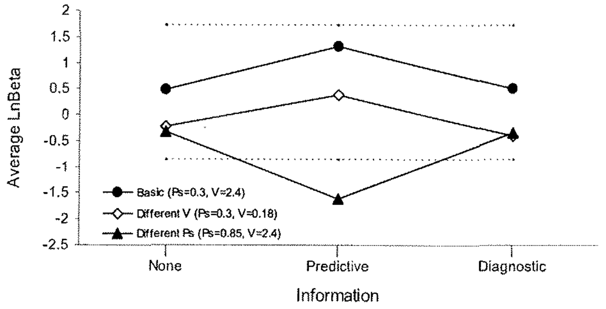

Figure 9. Threshold values in Experiment 2 for different combinations of p_S and V for the three information conditions. The dashed lines represent optimal threshold values.

predictive- and diagnostic-values information for strategy formation, and employed the developed strategy in the later blocks.

Agreements. To analyze participants' agreement with the cues we combined the results from all four blocks into one mean value to avoid missing values. Agreement with the cues was generally very high, but it differed among the information conditions, with 0.97 (SD = 0.05), 0.95 (SD = 0.05), and 0.92 (SD = 0.05) for the predictive, diagnostic, and none-information, respectively, $F(2, 49) = 3.37$, $MSE = 0.02$, $p = .04$, partial $\eta^2 = 0.12$. A post hoc comparison with a Ryan correction indicated that the only significant difference was between the predictive-values and the none-information ($p < .05$). This finding is in line with Papastavrou and Lehto's (1996) research which suggested greater compliance with warning cues when the cues' PPV is provided than when such information is lacking.

Similar to the findings in Experiment 1 (shown in Figure 7), participants agreed more with a green cue than with a red cue when the probability for a fault was relatively low (i.e., $p_S = 0.3$), but agreed more with a red rather than a green cue when the fault probability was relatively high (i.e., $p_S = 0.85$), as indicated by the significant interaction between the combination and the cue (red or green), $F(1.73, 84.98) = 11.7$, $MSE = 0.01$, $p < .001$, partial $\eta^2 = 0.19$. This is additional evidence for participants realizing that the PPV of a red cue is greater when the fault

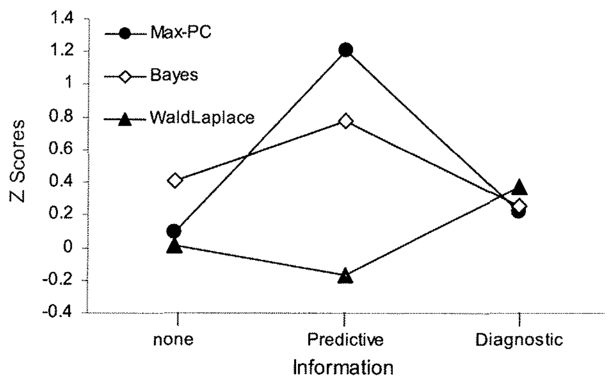

Figure 10. The degree of fit between participants' responses in Experiment 2 and the three models for the three information conditions.

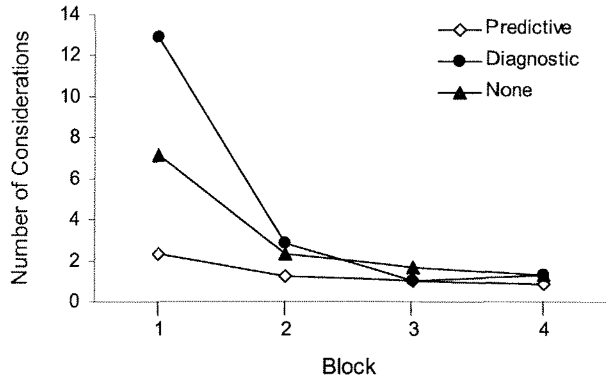

Figure 11. The number of considerations in the four blocks in Experiment 2 for each information condition.

probability is higher and the NPV of a green cue is greater when the fault probability is lower. Still, as previously noted, to maximize their scores participants should have followed the cues, regardless of the probability for a fault.

Discussion

The results of Experiment 2 are similar to those of Experiment 1. As in Experiment 1, the predictive-values condition yielded the highest score, and this score was significantly different from the score in the none-condition but not from the score in the diagnostic-values condition. In Experiment 2, however, the score in the diagnostic-values condition was also significantly higher than the score in the none-condition.

In Experiment 2, participants were again sensitive to changes in V and p_S, and they adjusted the threshold settings in the correct directions. They chose higher threshold values if the probabilities for faulty products were low or if the penalty for a false alarm was high. For V, the adjustment did not depend on the available information, and it was smaller than that required by the normative model. Thus, again, a sluggish beta effect was evident. The adjustment for p_S was smaller than required in the none and the diagnostic-values conditions, and it was close to the optimal adjustment in the predictive-values condition.

When testing the fit of the individual threshold settings to the different models, the only relatively high correlation was between the settings in the predictive-values condition and the Max-P(C) model. In the other two information conditions, there were no significant differences in the degree of fit to the models.

Similar to Experiment 1, the available information did not influence the time spent on choosing a threshold setting. Thus, the improved performance with predictive-values displays did not result from participants investing greater effort in strategy formation. Nonetheless, in Block 1 the number of considerations differed between the none-condition and the two other conditions thus indicating that participants did consider the presented conditional probability information. In the next section, drawing on the results of the two experiments and on participants' self reports, we describe how participants might have used conditional probabilities related information in choosing a threshold.

A Quantitative Model of Threshold Choices

We questioned participants about the strategies employed for threshold setting decisions. A commonly described strategy was setting the threshold at a point where about 95% confidence can be achieved for one type of event, while at the same time about 85% confidence can be achieved for a different type of event. For example, in the predictive-values condition, a setting can be chosen to reach about 95% PPV and about 85% NPV. In the diagnostic-values condition, the threshold can be set to reach about 95% p_{Hit} and about 85% p_{CR} and vice versa.

Table 3 presents the average lnβ values participants set in the two experiments with their corresponding PPV, NPV, p_{Hit}, and p_{CR} values. It shows that the PPV, NPV, p_{Hit}, and p_{CR} values in both experiments correspond quite closely with the model described in the previous paragraph.

Figures 12 and 13 provide additional support for the suggested model. In Figure 12, p_{Hit} rapidly decreases as lnβ moves toward the positive side of the scale, and p_{CR} rapidly decreases as lnβ moves toward the negative side of the scale. When lnβ is greater than 1.02 or smaller than −1.02, p_{Hit} and p_{CR} become smaller than 0.8, respectively. Thus, Figure 12 explains why participants who received diagnostic-values information chose average lnβ values in the range between −0.408 and 1.17.

At the same time, the three panels of Figure 13 explain why participants who viewed the predictive-values display chose more extreme threshold values than participants who viewed the diagnostic-values display. A comparison between Figures 12 and 13 shows that predictive-values condition thresholds need to be more extreme than diagnostic-values condition thresholds before one of the desirable conditional probabilities falls below 0.8.

The three panels of Figure 13 also demonstrate the strong effect probabilities have in the predictive-values condition. Participants set very different thresholds when $p_S = 0.15$ and $p_S = 0.85$ but very similar thresholds for the two different payoff matrices. It seems that participants tried to maintain an NPV of 95% when the probability of a malfunction was low ($p_S = 0.15$) and a PPV of 95% when the probability of a malfunction was high ($p_S = 0.85$), regardless of the payoffs.

Only when $p_S = 0.3$ did participants consider payoffs, as required by the normative model. The average thresholds that participants set for the $p_S = 0.3$, V = 0.18 and $p_S = 0.3$, V = 2.4 combinations were 0.9 units of lnβ apart, and their positions relative to one another corresponded with the normative model (see Table 3). The middle panel of Figure 13 shows that participants were ready to accept lower than 95% NPV values for the $p_S = 0.3$, V = 2.4 combination, which brought about the difference between the thresholds for the two combinations with $p_S =$ 0.3 and the different payoffs. Possibly, because $p_S = 0.3$ is less extreme than $p_S = 0.15$ or $p_S = 0.85$, participants were ready to set a threshold for the $p_S = 0.3$, V = 2.4 combination that corresponds to a less extreme conditional probability, but is more in line with the values in the payoff matrix.

Participants in the diagnostic-values conditions were more responsive to the values in the payoff matrices. The thresholds they set for the same p_S but different payoff combinations were further apart than the thresholds set in the predictive-values condition (see Table 3). Participants apparently sought to maintain either the value of p_{Hit} at about 0.95 and the value of p_{CR} above 0.80 or vice versa, and they chose the conditional probabilities for which they desired the greater value according to both p_S and the payoffs. Therefore, the thresholds they set for the combinations $p_S = 0.3$, V = 0.18 and $p_S = 0.3$, V = 2.4 were on both sides of the origin, in accordance with the normative model.

Thus, it seems that decision makers understand the required pattern of changes in p_{Hit} and p_{CR} for different combinations of p_S and payoffs better than they understand the required pattern of changes in PPV and NPV. However, predictive-values information leads to better threshold adjustments in response to changes in p_S than the diagnostic-values information. Therefore, as long as the payoffs are not very dominant in determining the optimal threshold value (see formula 1), we can expect better threshold choices with predictive-values information.

General Discussion

In two experiments, we assessed participants' threshold settings for a binary cue in a binary categorization task as a function of the possible outcomes values, the categories' probabilities, and the information available to participants about the conditional probabilities associated with different thresholds. Participants' scores in both experiments were significantly higher when they received

Table 3
Average Threshold Values (lnβ Units) in the Two Experiments With Their Corresponding Positive Predictive Values and Negative Predictive Values and With Their Corresponding Probability for Hit and Probability for False Alarms

			Predictive values condition			Diagnostic values condition		
Experiment	p_S	V	lnβ	PPV	NPV	lnβ	p_{Hit}	p_{CR}
1	0.15	0.416	1.722	0.82	0.95	0.38	0.86	0.92
1	0.15	2.4	1.92	0.84	0.95	1.17	0.78	0.95
1	0.85	0.416	−1.88	0.94	0.84	−0.77	0.94	0.82
1	0.85	2.4	−1.73	0.95	0.82	−0.04	0.89	0.89
2	0.3	2.4	1.3	0.89	0.9	0.508	0.85	0.92
2	0.3	0.18	0.39	0.82	0.94	−0.408	0.92	0.86
2	0.85	2.4	−1.61	0.95	0.85	−0.34	0.92	0.86

Note. In Experiment 1, V = 0.416 is the costly misses payoff matrix and V = 2.4 is the costly false alarms payoff matrix. In Experiment 2, $p_S = 0.3$, V = 2.4 is the basic combination, $p_S = 0.85$, V = 2.4 is the different probability combination and $p_S = 0.3$, V = 0.18 is the different payoff combination. PPV = positive predictive values; NPV = negative predictive values; Phit = probability for hit; PFA = probability for false alarms.

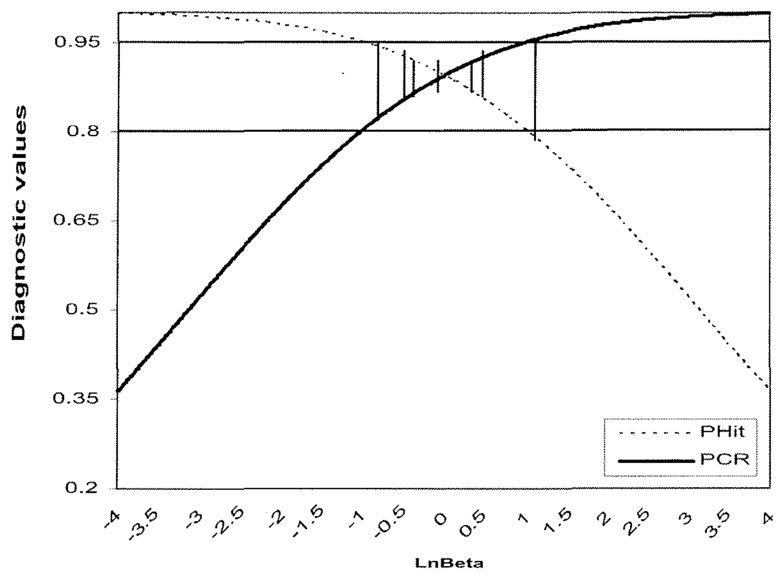

Figure 12. Diagnostic values as a function of lnβ. The 0.95 and 0.8 conditional probabilities are designated by the two horizontal lines in the figure. Each of the seven vertical lines in the figure designates the average threshold value for one of the seven combinations in the two experiments.

predictive-values information than when they received no information. When given diagnostic-values information, participants achieved intermediate scores, which were not significantly different from the scores in the two other information conditions in Experiment 1 but were significantly higher than the score in the none condition in Experiment 2.

The general tendency in all three information conditions was to use higher thresholds when p_S was smaller and false alarms were costly; a finding which adheres to the normative model. Participants were somewhat sensitive to changes in the payoffs and p_S, and they adjusted threshold settings in the correct directions. For payoffs, the adjustment was unaffected by available information, and it was smaller than required by the normative model. Thus, participants were not sufficiently responsive to changes in the payoffs.

For p_S, the adjustment in the none and the diagnostic-values conditions was smaller than required. For the predictive-values condition the change in the threshold setting was close to the

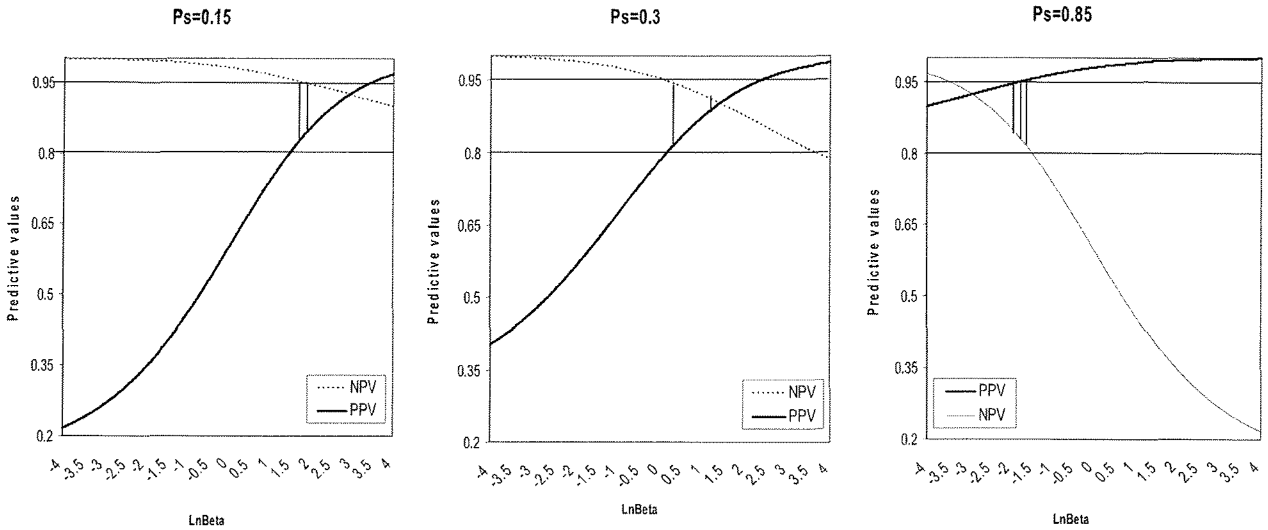

Figure 13. Predictive values as a function of lnβ for p_S = .15 (left panel), p_S = .3 (middle panel), and p_S = .85 (right panel). The 0.95 and 0.8 conditional probabilities are indicated by the two horizontal lines across the figure. The smaller vertical lines designate the threshold values set for the different combinations of V and p_S.

optimal change. Individual threshold settings in the predictive-values condition corresponded quite well with the Max-P(C) model, a finding which indicates that participants mainly considered probabilities when choosing threshold settings. No significant differences were found in the degree to which individual threshold settings in the none and the diagnostic-values conditions corresponded with the different models.

We propose a simple model to explain participants' threshold settings in the diagnostic- and predictive-values conditions. According to the proposed model, participants set their thresholds so that the probability for one event that they strived to maximize was close to 0.95 and the probability for the other event was around 0.8. The probabilities for events participants considered were the PPV and the NPV values in the predictive-values condition and the p_{Hit} and p_{CR} values in the diagnostic-values condition.

The reason for the participants' above described behavior is unclear. One obvious explanation might be that our participants, who had all attended statistics courses, considered the .05 criterion for significance, although none of our participants explicitly mentioned such a consideration. Alternatively, the .95 value may be special in some way. Ayres, Ayres-Brown, and Ayres-Brown (2007) suggest that the .95 probability may indeed have an intuitively unique status, which may be related to the properties of the normal distribution (i.e., the convex curvature of the normal probability density function is maximal at 1.73 standard deviations from the mean which corresponds to $p = .96$). We believe that further studies with participants who are unaware of the special status of the .05 probability in many scientific domains may help evaluate the applicability of our model to other populations.

Use of this model led to close-to-optimal threshold settings in the predictive-values condition, as long as the probability for a faulty product was the main determinant of the optimal threshold values. When payoffs were also important, participants chose erroneous thresholds because it was apparently difficult for them to accept the idea that good threshold choices may be associated with intermediate NPV or PPV values.

Although participants' behavior matched the predictions of the Max-P(C) model, it seems that they did not actually act according to the model, that is, neglected payoffs and considered only prior probabilities. Rather, by trying to keep the PPV and NPV above certain values, they acted as if they considered mainly the prior probabilities. When the same rule was applied to p_{Hit} and p_{CR}, none of the models was correlated with choices.

Therefore, in the diagnostic-values condition, participants' threshold choices were correct, in terms of their position, relative to the origin (i.e., they were positive when indeed they should have been positive). However, participants could not accept that good threshold choices may be associated with p_{Hit} or p_{CR} values lower than 0.8. As a result, their thresholds were conservative and close to the origin when they should have been more negative or more positive.

It appears, then, that the superior threshold choices by participants who received predictive-values information were not the result of considering only prior probabilities. Rather, they were the result of setting more extreme thresholds that better fit changes in prior probabilities. We already showed that when payoffs were dominant in determining optimal threshold values, thresholds set by participants who received predictive-values information were worse than those set with the other displays.

The question arises whether our findings are limited to the experimental settings we employed. In particular, one may ask whether these results can be generalized to other payoff matrices and probabilities for faults. Of course, the only way to test these questions is to conduct further research on threshold settings in different settings.

Still, the particular settings we used here are important and resemble commonly encountered settings. For instance, in many domains, such as quality control or safety related decisions, including air traffic control, payoff matrices tend to have only negative values, with penalties for missed detections and for false alarms, but no explicit bonuses for correct rejections or hits. Also, the model we describe above leads to some nontrivial predictions when fault probabilities are very small (for instance less than 1/1000). In these cases the model would predict the setting of extremely high thresholds for which PPV would be above .8, leading to very few detections of faults. In contrast, when diagnostic information is available, the model predicts the same threshold settings as in our study, which will greatly differ from the optimal settings. Further research should examine the question whether for these low probabilities the model still predicts threshold settings, or whether participants adopt a different strategy for choosing thresholds when the probability for faults is very low.

Conclusions

In both experiments, participants were not particularly successful at setting thresholds for binary categorization decisions. They tended to shift the threshold in the correct direction in response to changes in probabilities and payoffs, but the shift size was usually smaller than the optimal adjustment derived from a maximum utility model.

When optimal thresholds mainly depended on probabilities, participants in the predictive-values condition set more accurate thresholds and performed tasks better than participants in the other two information conditions. However, when optimal thresholds strongly depended on payoffs, predictive-values information led to inferior threshold settings, compared to the other information conditions.

In the diagnostic-values condition, threshold settings accuracy did not change as a function of the importance of the payoffs in determining optimal threshold values. Participants chose very conservative thresholds, similar to the thresholds set by participants lacking information about conditional probabilities.

We, therefore, conclude that as long as threshold settings mainly depend on probabilities, they may improve if decision makers receive information about the predictive-values associated with different thresholds. Currently, however, this conclusion should be qualified to cases where prior probabilities are not very low. This issue, together with the structure of payoff matrices and the target population should be addressed in future investigations.

At a practical level we suggest that whenever possible, thresholds should be derived from a normative model and set automatically. If automatic threshold setting is impossible, people may indeed adjust thresholds as they are capable of shifting thresholds in the correct direction, in response to dynamic environmental changes. However, manually set thresholds are likely to diverge greatly from optimal settings.

References

Ayres, I., Ayres-Brown, A. R., & Ayres-Brown, H. J. (2007). Seeing significance: Is the 95% probability range easier to perceive? *Chance, 20,* 11–16.

Ballestero, E. (2002). Strict uncertainty: A criterion for moderately pessimistic decision makers. *Decision Sciences, 33,* 87–107.

Bliss, J. P. (2003a). An archival investigation of extreme alarm response patterns in laboratory experiments. Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society. October 13–17: Denver, CO.

Bliss, J. P. (2003b). Investigation of alarm-related accidents and incidents in aviation. *International Journal of Aviation Psychology, 13,* 249–268.

Chi, C. F., & Drury, C. G. (1998). Do people choose optimal response criterion in an inspection task? *IIE Transactions, 30,* 257–266.

Craig, A., & Colquhoun, W. P. (1977). Vigilance effects in complex inspection. In *vigilance-theory. Operational performance and physiological correlates.* Mackie. R. R. (Ed.). Plenum Press. New York. pp. 19–29.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193–242.

Fahey, T., Griffiths, S., & Peters, T. J. (1995). Evidence based purchasing: Understanding results of clinical trials and systematic reviews. *British Medical Journal, 311,* 1056–1059.

Fox, J. G., & Haslegrave, C. H. (1969). Industrial inspection efficiency and the probability of a defect occurring. *Ergonomics, 12,* 713–721.

Getty, D. J., Swets, J. A., Rickett, R. M., & Gonthier, D. (1995). System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. *Journal of Experimental Psychology: Applied, 1,* 19–33.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instructions: Frequency formats. *Psychological Review, 102,* 684–704.

Gigerenzer, G., & Hoffrage, U. (1999). Helping people overcome difficulties in Bayesian reasoning: A reply to Lewis and Keren (1999). and Mellers and McGraw (1999). *Psychological Review, 106,* 425–430.

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics.* New York: Wiley.

Hutchins, E. (1995). *Cognition in the wild.* Cambridge, MA: MIT Press.

Lawrence, M., Goodwin, P., & Fildes, R. (2002). Influence of user participation on DSS use and decision accuracy. *Omega, 30,* 381–392.

Lehto, M. R., Papastavrou, J. D., & Giffen, W. J. (1998). An empirical study of adaptive warnings: Human- versus computer-adjusted warning thresholds. *International Journal of Cognitive Ergonomics, 2,* 19–33.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd Edition). Mahwah, NJ: Erlbaum.

Maltz, M., & Meyer, J. (2001). Use of warnings in an attentionally demanding detection task. *Human Factors, 43,* 217–226.

Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human Factors, 43,* 563–572.

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors, 46,* 196–204.

Mobley, W. H., & Goldstein, I. L. (1978). The effects of payoff on the visual processing of dental radiographs. *Human Factors, 20,* 385–390.

Murrell, G. (1977). Combination of evidence in a probabilistic visual search and detection task. *Organizational Behavior and Human Performance, 18,* 3–18.

Papastavrou, J. D., & Lehto, M. R. (1996). Improving the effectiveness of warnings by increasing the appropriateness of their information content: Some hypotheses about human compliance. *Safety Science, 21,* 175–189.

Shurtleff, M. S. (1991). Effects of specificity of probability information on human performance in a signal detection task. *Ergonomics, 34,* 469–486.

Sorkin, R. D. (1988). Why are people turning off our alarms? *Journal of the Acoustical Society of America, 84,* 1107–1108.

Sorkin, R. D., & Woods, D. D. (1985). Systems with human monitors: A signal detection analysis. *Human Computer Interaction, 1,* 49–75.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Better decisions through science. *Scientific American, 283,* 82–87.

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems.* New York: Academic.

Wickens, C. D. (1992). *Engineering psychology and human performance,* 2nd edn., Harper Collins: New York.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science, 8,* 201–212.