# Training for Vigilance: Using Predictive Power to Evaluate Feedback Effectiveness

James L. Szalma, Peter A. Hancock, Joel S. Warm, William N. Dember and Kelley S. Parsons

*Human Factors: The Journal of the Human Factors and Ergonomics Society* 2006 48: 682
DOI: 10.1518/001872006779166343

# Training for Vigilance: Using Predictive Power to Evaluate Feedback Effectiveness

**James L. Szalma** and **Peter A. Hancock,** University of Central Florida, Orlando, Florida, and **Joel S. Warm, William N. Dember,** and **Kelley S. Parsons,** University of Cincinnati, Cincinnati, Ohio

**Objective:** We examined the effects of knowledge of results (KR) on vigilance accuracy and report the first use of positive and negative predictive power (PPP and NPP) to assess vigilance training effectiveness. **Background:** Training individuals to detect infrequent signals among a plethora of nonsignals is critical to success in many failure-intolerant monitoring technologies. KR has been widely used for vigilance training, but the effect of the schedule of KR presentation on accuracy has been neglected. Previous research on training for vigilance has used signal detection metrics or hits and false alarms. In this study diagnosticity measures were applied to augment traditional analytic methods. **Method:** We examined the effects of continuous KR and a partial-KR regimen versus a no-KR control on decision diagnosticity. **Results:** Signal detection theory (SDT) analysis indicated that KR induced conservatism in responding but did not enhance sensitivity. However, KR in both forms equally enhanced PPP while selectively impairing NPP. **Conclusion:** There is a trade-off in the effectiveness of KR in reducing false alarms and misses. Together, SDT and PPP/NPP measures provide a more complete portrait of performance effects. **Application:** PPP and NPP together provide another assessment technique for vigilance performance, and as additional diagnostic tools, these measures are potentially useful to the human factors community.

## INTRODUCTION

Vigilance refers to the ability of observers to detect infrequent signals over prolonged periods on watch (Davies & Parasuraman, 1982). The importance of vigilance has vaulted to the forefront of current social concerns regarding detection of terrorist activities (Hancock & Hart, 2002). For example, the screening devices deployed by the Transportation Security Administration to screen 100% of checked baggage represents a major vigilance requirement for operators (Harris, 2002). Sustained attention is also central to the majority of advanced human-machine systems, including air traffic control, cockpit monitoring, industrial quality control, nuclear power generation, medical monitoring, and cytological screening (Warm, 1993). How to train individuals for vigilance and how to evaluate performance are perennial concerns. An effective training procedure is to provide information about an operator's responses via knowledge of results (KR). Research has indicated that monitors trained with KR perform better after the training aid has been withdrawn than do controls who received no KR training (see Davies & Parasuraman, 1982; Warm & Jerison, 1984). However, the mechanisms by which KR exerts its effects, and how best to employ KR for training, remain unspecified.

In vigilance, performance accuracy is assessed predominantly by evaluating the proportion of correct detections and false alarms. These scores are used to compute signal detection theory (SDT) measures of sensitivity (e.g., $d'$ or $A'$; Macmillan & Creelman, 2005; See, Howe, Warm, & Dember, 1995) and response bias (e.g., $\beta$; Davies & Parasuraman, 1982; also see See, Warm, Dember, & Howe, 1997). Sensitivity assesses operators' ability to discriminate signal from nonsignal. In vigilance, response bias, which tends to become increasingly conservative over time, is thought to reflect increasing awareness on behalf of operators that the

Address correspondence to James L. Szalma, Department of Psychology, University of Central Florida, P.O. Box 161390, Orlando, FL 32816-1390; jszalma@mail.ucf.edu. *HUMAN FACTORS*, Vol. 48, No. 4, Winter 2006, pp. 682–692. 

signals for which they are searching are relatively rare (Craig, 1978). Indeed, KR has been shown to enhance perceptual sensitivity while increasing β (e.g., see Szalma, Miller, Hitchcock, Warm, & Dember, 1998). However, although these established measures assess critical aspects of performance, they do not directly capture the diagnostic accuracy of response (i.e., the proportion of individuals' "yes" or "no" responses that are correct). SDT measures reflect diagnosticity indirectly because one might infer greater diagnostic accuracy from higher $d'$ scores. It is possible, however, to achieve high sensitivity while making poorer diagnostic decisions (Parasuraman, Hancock, & Olofinboba, 1997). This is particularly evident when the a priori probability of a signal (base rate) is very low, as in most real-world situations. Indeed, Parasuraman et al. (1997) demonstrated that under these conditions the accuracy of system response can also be low even when sensitivity is exceptionally high. Operators with high sensitivity can generate unacceptably large numbers of false alarms and reduce their diagnostic power, which inevitably results in reduced overall system effectiveness.

It is therefore critical in operational settings that observers learn to be highly diagnostic as well as sensitive to perceptual changes in the displays which they are monitoring. Specifically, it is vital that when an operator indicates a signal is present that the signal actually be present. Conversely, when an operator indicates that a signal is not present, it is vital that the signal is really not there. These aspects of performance are captured by the decision theory measures of positive predictive power (PPP), which is the proportion of "yes" responses that are actually correct, and negative predictive power (NPP), which is the proportion of "no" responses that are actually correct. The computational formula for PPP is $H/(H + FA)$, in which $H$ = number of signals detected (hits) and $FA$ = number of false alarms. The comparable computational formula for NPP is $CR/(CR + M)$, in which $CR$ = number of correct rejections and $M$ = number of signals missed. These indices are frequently employed in evaluating decision-making in medicine (e.g., see Linton, 1996) but have rarely been employed in the human factors literature (for an exception, see Getty, Swets, Pickett, & Gonthier, 1995). A perfectly accurate observer would yield a PPP of 1.0; a score of 0 would indicate no correct detections and no diagnosticity. Similarly, an observer who correctly rejected all nonsignals and committed no misses would achieve a NPP score of 1.0, whereas a score of 0 indicates that no correct rejections were made.

Given the importance of training personnel for monitoring and the relative costs of false alarms and misses, it is vital that any training regimen maximize both sensitivity and diagnostic power. Hence, we applied PPP and NPP to the examination of the effectiveness of KR for vigilance training, and we compared these measures to nonparametric measures of SDT. We do not propose that PPP and/or NPP measures should replace SDT, which is one of the most powerful and useful quantitative theories in behavioral science. Rather, we show how inclusion of these "new" measures complements SDT indices and provides an alternative vista into detection capability.

KR has been shown to be an effective training aid for a variety of tasks beyond vigilance, including those entailing motor skills (e.g., Salmoni, Schmidt, & Walter, 1984). However, the schedule of KR presentation influences the effectiveness of such training. Specifically, partial KR (i.e., KR provided only during portions of training) has been shown to be more effective than continuous KR in training motor skills (Salmoni et al., 1984). If these findings extend to vigilance, performance during training should be superior for monitors receiving continuous KR relative to those who receive partial KR, but the pattern should reverse during the transfer phase. Such a result would suggest that partial KR would be the more effective training mode for vigilance in operational settings.

Most vigilance studies employing KR for training have used a continuous KR schedule, but a limited number of studies have employed partial KR (i.e., McCormack, Binding, & McElheran, 1963; Warm, Hagner, & Meyer, 1971). In the latter work, Warm et al. (1971) found that observers provided with KR only 50% of the time during training showed a smaller vigilance decrement in a subsequent test period without KR than did observers who had received continuous KR during training. Warm et al. (1971) attributed this partial-KR advantage to the motivational effects of feedback. However, they employed a *speed* measure, reaction time to correct detections, to assess performance in a simple vigilance task (the onset of a small jewel light), and thus no study has examined the effect of partial KR on vigilance in which the *accuracy* of operator responses was the

critical measure of interest. Although speed is of importance in detection tasks, it is arguable that accuracy is a more important real-world imperative. Indeed, prior research has shown that accuracy and response time may not always reflect the same perceptual processes (Santee & Egeth, 1982). Thus, effects of partial KR on response accuracy in vigilance have yet to be evaluated, and this is another goal unique to the present study.

## METHOD

Sixteen observers (8 men and 8 women) were assigned at random to each of three KR groups (continuous KR, partial KR, and a no-KR control). Observers ranged in age from 18 to 40 years (mean = 20.5) and served to fulfill a university course requirement. All observers had normal or corrected-to-normal vision and were reportedly free of hearing impairments. They took part in a 45-min session divided into a training phase of three continuous 4-min periods of watch followed by a test phase of five continuous 4-min periods. A 10-min rest interval separated the training and testing phases.

The display consisted of a 1.4-cm diameter green disk flanked on each side by a 1.0-cm vertical green line appearing against a gray background. The lines were connected to the disk by a 1.0-cm horizontal green line that extended from the midpoint of each vertical line through the horizontal diameter of the disk. Neutral events, requiring no overt response, were cases in which the two vertical lines were equidistant from the disk. Critical signals were cases in which one vertical line was 0.4 cm farther from the disk than was the other line. Observers were instructed to respond by pressing the space bar on a computer keyboard whenever a critical signal appeared on the screen. Responses occurring within 1.3 s of critical signal onset were recorded by the computer as correct detections. All other responses were recorded as false alarms. The screen was mounted on a table at eye level approximately 54 cm from the seated observer. This task was selected in part because of existing comparable results readily available to the authors (e.g., Szalma et al., 1998).

In both experimental phases, the disk and line display was presented in the center of the screen for 200 ms once every 2 s (event rate = 30/min). Critical signals occurred twice per minute (signal probability = .066) at random intervals within each period on watch. Half of the signals in each period were cases in which the left vertical line was farther from the disk, and the other half were cases in which the right vertical line was farther from the disk. The order in which the two types of critical signal were presented (left vs. right) was randomly distributed over each watch period. During training, KR was provided by the computer using a prerecorded female voice to announce correct detections ("correct"), missed signals ("miss"), and errors of commission ("false alarm"). In the no-KR condition, the voice announced "saved" after each response. In the partial-KR group, participants were informed that KR would be provided by means of the voice, which announced "feedback on" at the beginning of Periods 1 and 3 and "feedback off" at the beginning of Period 2. Observers in the continuous KR condition were informed that they would receive feedback throughout the training session. Note that the difference between the continuous and partial KR conditions was in the number of periods during which KR was provided rather than the information imparted by the feedback.

## RESULTS

### Correct Detections

*Training phase.* The mean percentages of correct detections during both training and transfer are reported in Table 1. For training, ANOVA confirmed a significant KR effect, $F(2, 45) = 8.18$, $p < .01$, $\omega^2 = .09$. The effects for the other factors lacked statistical significance. In this and all subsequent analyses proportions were converted to arcsines to normalize the data. The means and standard errors reported are based on the untransformed percentages. Further, in all analyses, Box's correction was employed to adjust for violations of sphericity. Tukey HSD comparisons indicated that the detection scores of observers in the no-KR condition were significantly higher than those of observers who received either form of KR but that detection scores in the two KR conditions did not differ significantly from each other.

*Transfer phase.* During the transfer phase observers in the no-KR group attained higher detection scores than did those in the KR groups. An ANOVA indicated a significant main effect for KR, $F(2, 45) = 12.00$, $p < .001$, $\omega^2 = .08$, for periods, $F(3, 152) = 9.02$, $p < .001$, $\omega^2 = .12$, and for the interaction, $F(7, 152) = 2.04$, $p < .05$, $\omega^2 = .05$.

**TABLE 1:** Mean Percentage of Correct Detections and False Alarms During Training and Transfer for Three KR Conditions

| | Period on Watch | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | Mean | Cohen's *d* |
| | | | Correct Detections | | | | |
| Training | | | | | | | |
| No KR | 67 (7) | 70 (4) | 67 (6) | | | 68 (4) | |
| Continuous KR | 62 (5) | 52 (5) | 50 (3) | | | 54 (4) | –0.88 |
| Partial KR | 55 (5) | 44 (5) | 44 (4) | | | 48 (4) | –1.25 |
| Mean | 61 (3) | 56 (3) | 54 (3) | | | | |
| Transfer | | | | | | | |
| No KR | 77 (5) | 74 (5) | 65 (6) | 64 (5) | 63 (4) | 69 (4) | |
| Continuous KR | 61 (0.04) | 45 (4) | 46 (5) | 41 (5) | 38 (4) | 46 (3) | –1.53 |
| Partial KR | 62 (4) | 50 (4) | 55 (4) | 54 (4) | 58 (4) | 56 (2) | –0.98 |
| Mean | | 67 (3) | 56 (3) | 55 (3) | 53 (3) | 53 (2) | |
| | | | False Alarms | | | | |
| Training | | | | | | | |
| No KR | 4.4 (0.7) | 4.6 (0.7) | 3.5 (0.7) | | | 4.1 (0.4) | |
| Continuous KR | 2.5 (0.4) | 1.0 (0.2) | 1.2 (0.2) | | | 1.5 (0.4) | –0.94 |
| Partial KR | 2.2 (0.4) | 1.2 (0.2) | 1.5 (0.3) | | | 1.6 (0.4) | –0.90 |
| Mean | 3.0 (0.3) | 2.3 (0.3) | 2.0 (0.3) | | | | |
| Transfer | | | | | | | |
| No KR | 4.6 (0.8) | 4.6 (0.7) | 3.7 (0.7) | 3.6 (0.7) | 3.3 (0.6) | 4.0 (0.6) | |
| Continuous KR | 1.2 (0.3) | 0.5 (0.2) | 0.7 (0.2) | 0.9 (0.3) | 0.7 (0.2) | 0.8 (0.1) | –1.58 |
| Partial KR | 1.2 (0.3) | 0.9 (0.3) | 0.9 (0.2) | 0.9 (0.3) | 1.7 (0.5) | 1.1 (0.2) | –1.39 |
| Mean | 2.3 (0.3) | 2.0 (0.3) | 1.8 (0.3) | 1.8 (0.3) | 1.9 (0.3) | | |

*Note.* Values in parentheses are standard errors. Cohen's *d* was computed by comparing each experimental group with the no-KR control condition.

Simple effects tests of periods within each KR condition indicated a significant decline over time for observers in the continuous KR group, $F(4, 53) = 5.89$, $p < .01$, $\omega^2 = .20$, and the no-KR group, $F(3, 44) = 6.18$, $p < .001$, $\omega^2 = .20$. Detection scores for observers in the partial KR group did not significantly change over time.

## False Alarms

*Training phase.* The mean percentages of false alarms during the training and transfer sessions are reported in Table 1. ANOVA revealed significant effects for KR, $F(2, 45) = 9.37$, $p < .001$, $\omega^2 = .10$, and for periods, $F(2, 88) = 6.75$, $p < .01$, $\omega^2 = .07$. The interaction between these factors was not statistically significant. Tukey HSD comparisons indicated that participants in the no-KR condition committed significantly more false alarms than those in either KR condition, but scores in the two KR groups did not differ significantly from each other. Across the KR conditions false alarms declined over periods.

*Transfer phase.* During transfer, observers in the no-KR control group committed substantially more false alarms than those who received KR training. In fact, 41% of all the false alarm scores were zero for the KR groups across the five periods of watch. An ANOVA indicated a significant effect for KR, $F(2, 45) = 13.16$, $p < .001$, $\omega^2 = .09$, and a significant interaction, $F(7, 166) = 2.24$, $p < .05$, $\omega^2 = .04$. The period effect was not significant. Simple effects tests of periods within each KR condition indicated that false alarms committed by observers in the no-KR condition declined significantly over the watch period, $F(2, 33) = 3.52$, $p < .05$, $\omega^2 = .11$. False alarm scores for observers in the two KR conditions did not change significantly over time.

## Signal Detection Analysis

The substantial number of zero false alarm rates in the KR groups rendered application of parametric SDT analysis inadvisable because of the problems associated with computation of such indices under these conditions (see Davies & Parasuraman, 1982; Macmillan & Creelman, 2005).

To circumvent these concerns, Craig (1979) suggested that nonparametric sensitivity indices (specifically, $A'$) should be used when parametric measures are inappropriate. Work by See et al. (1997) established that among the available, nonparametric indices of response bias, $\beta''_D$ was the most effective for vigilance. The correct detection and false alarm scores for each participant were therefore used to compute $A'$ and $\beta''_D$.

### Perceptual Sensitivity ($A'$)

*Training phase.* An ANOVA of the $A'$ scores during training indicated that sensitivity did not significantly differ across the three KR conditions and that $A'$ scores for observers remained stable with time on watch. The interaction between these factors also lacked significance.

*Transfer phase.* Sensitivity scores during transfer for the three KR conditions are plotted as a function of periods of watch in the top panel of Figure 1. An ANOVA indicated a significant effect for KR, $F(2, 45) = 6.82$, $p < .01$, $\omega^2 = .05$, and a significant effect for period, $F(4, 104) = 4.20$, $p < .05$, $\omega^2 = .05$. Sensitivity of observers declined over time in all conditions, confirming the classic "vigilance decrement" (See et al., 1995). The interaction was not statistically significant. Tukey HSD tests indicated that observers in the continuous KR group ($M = .85$, $d = -.78$) were significantly less sensitive than those in the no-KR condition ($M = .91$). $A'$ scores for observers in the partial KR condition ($M = .88$, $d = -.53$) did not differ significantly from those in the other two groups.

### Response Bias ($\beta''_D$)

*Training phase.* An ANOVA on $\beta''_D$ scores during training indicated a significant effect for KR, $F(2, 45) = 13.93$, $p < .001$ $\omega^2 = .15$. However, there

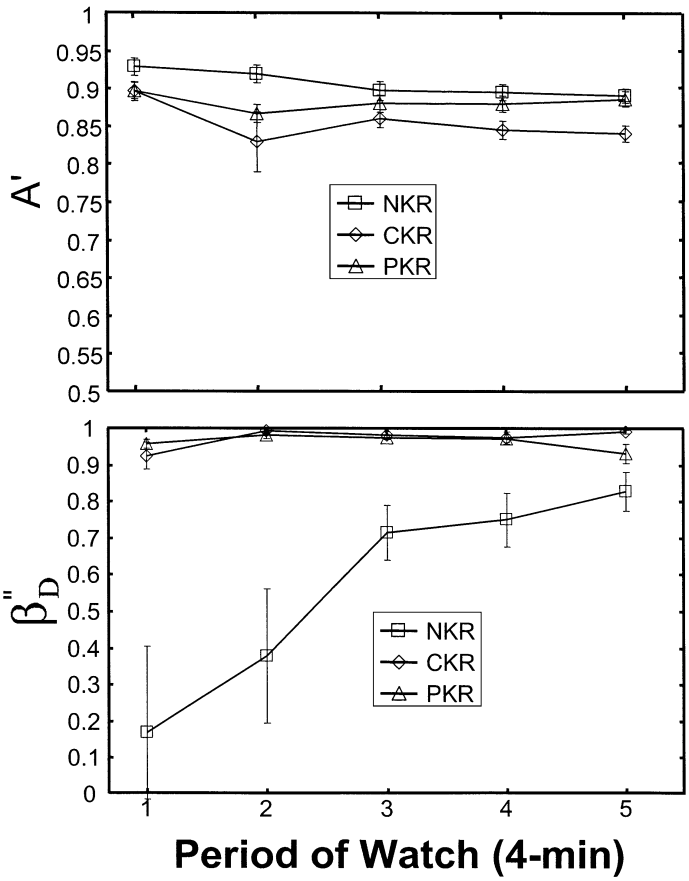


*Figure 1.* Top: Mean sensitivity ($A'$) scores as a function of periods of watch during transfer. Bottom: Mean response bias ($\beta''_D$) scores as a function of periods of watch during transfer. CKR = continuous KR; PKR = partial KR; NKR = no KR. Error bars are standard errors.

were no significant differences in $\beta''_D$ scores across periods, and the interaction was also not statistically significant. Tukey HSD tests revealed that observers in the continuous KR ($M = .94$, $d = .84$) and partial KR ($M = .95$, $d = .87$) groups did not significantly differ from one another, but both groups were significantly more conservative than those in the no-KR condition ($M = .54$).

*Transfer phase.* Mean response bias scores for observers in the three KR conditions during transfer are plotted as a function of periods of watch in the bottom panel of Figure 1. Observers who received either form of KR achieved a conservative level of responding early in the watch and remained at that level, whereas observers who did not receive KR were more lenient in responding and became more conservative over time. ANOVA revealed a significant effect for KR, $F(2, 45) = 13.21$, $p < .001$, $\omega^2 = .09$, for period, $F(4, 77) = 7.92$, $p < .01$, $\omega^2 = .10$, and for the interaction, $F(8, 77) = 6.85$, $p < .001$, $\omega^2 = .16$. Tests for the simple effects of period within each KR condition confirmed that observers in the no-KR group became more conservative over time, $F(4, 77) = 21.49$, $p < .01$, $\omega^2 = .51$, but that $\beta''_D$ scores of observers in the two KR groups did not change significantly over time.

### Diagnosticity Measures

The PPP and NPP were computed for each participant in each period of watch during both training and transfer sessions. It is important to remember that in vigilance the number of nonsignals far exceeds the number of signals, so opportunities for errors of omission are much rarer than opportunities for false alarms. Thus, in cases in which there is a low signal probability and many correct rejections (i.e., the observer adopts a conservative criterion), NPP values would be expected to be artificially high (above 90%). Therefore, any differences in NPP among KR conditions will likely be driven by differences in errors of omission. In contrast, PPP is influenced by frequencies of correct detections and false alarms, and these responses are relatively few in number compared with the total number of events during a vigil.

### Positive Predictive Power (PPP)

*Training phase.* An ANOVA of the PPP scores revealed a significant period effect, $F(2, 83) = 5.50$, $p < .01$, $\omega^2 = .06$, with scores increasing over time ($M$s for Periods 1–3 = .62, .68, and .71, respectively). No significant differences among KR groups were observed, and the interaction also lacked statistical significance.

*Transfer phase.* Mean PPP scores for the three groups are displayed as a function of period on watch in the top panel of Figure 2. ANOVA revealed a significant effect for KR, $F(2, 45) = 6.92$, $p < .01$, $\omega^2 = .05$. No significant differences were observed for period or the KR by period interaction. Tukey HSD tests revealed that the two KR groups did not significantly differ from each other (continuous KR $M = .83$, $d = .96$; partial KR $M = .81$, $d = .88$), but the PPP scores of both these groups exceeded those of the no-KR group ($M = .63$).

### Negative Predictive Power (NPP)

*Training phase.* An ANOVA revealed a significant effect for KR group, $F(2, 45) = 11.69$, $p < .01$, $\omega^2 = .13$. No significant effects were observed for periods or the interaction between these factors. Tukey HSD tests revealed that the no-KR group ($M = .98$) achieved NPP scores that were significantly higher than those of the continuous KR ($M = .97$, $d = –.72$) and partial KR groups ($M = .96$; $d = –.99$). The NPP scores of the two KR conditions did not differ significantly from each other. Although these absolute mean differences are small, the effect sizes are large as a result of low within-group variability.

*Transfer phase.* Mean NPP scores for the three groups are displayed as a function of period on watch in the bottom panel of Figure 2. Observers in the no-KR condition achieved higher NPP scores during transfer than did those in the two KR conditions. ANOVA indicated significant effects for KR, $F(2, 45) = 12.20$, $p < .001$, $\omega^2 = .09$, periods, $F(3, 141) = 10.93$, $p < .001$, $\omega^2 = .14$, and the interaction, $F(6, 141) = 2.82$, $p < .05$, $\omega^2 = .06$. Simple effects tests of periods within each KR condition indicated that NPP scores significantly declined over the watch period in the no-KR condition, $F(3, 141) = 11.07$, $p < .01$, $\omega^2 = .14$, and the continuous KR condition, $F(3, 141) = 4.30$, $p < .01$, $\omega^2 = .05$. NPP scores for participants in the partial KR condition did not change significantly over time.

## DISCUSSION

### Signal Detection

In this experiment we found no evidence that KR enhanced perceptual sensitivity. This was surprising, given that KR typically enhances signal detection in vigilance (Warm & Jerison, 1984).

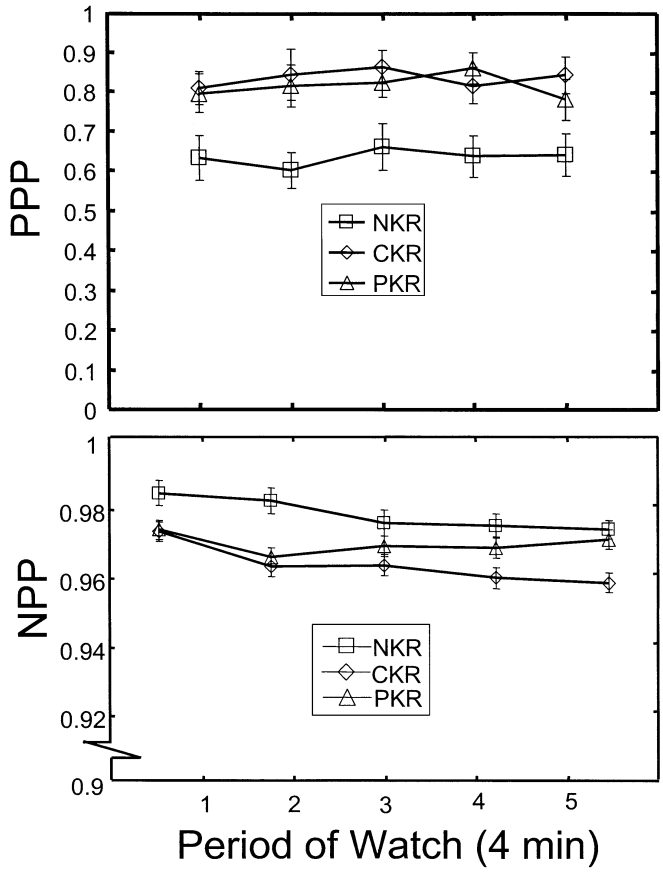


*Figure 2.* Top: Mean positive predictive power (PPP) as a function of periods of watch during transfer. Bottom: Mean negative predictive power (NPP) as a function of periods of watch during transfer. CKR = continuous KR; PKR = partial KR; NKR = no KR. Error bars are standard errors.

However, KR did increase conservatism in responding relative to a no-KR condition, a finding that has been observed previously in vigilance experiments (See et al., 1997). The increase in conservatism over time on watch in the no-KR condition is also consistent with prior research and may be attributable to the growing awareness over the vigil that critical signals are relatively rare (Craig, 1978). The provision of KR likely produces an immediate and explicit awareness of the rarity of signals and so causes observers to set a high criterion very early in the vigil.

There was a trade-off between PPP and NPP such that the enhanced diagnosticity provided by KR for "yes" responses (PPP) is achieved at the cost of lower diagnosticity for "no" responses (NPP). Consideration of both the SDT and diagnosticity metrics indicates that KR increases conservatism and improves the reliability of "yes" responses but that this occurs at the expense of lower accuracy in "no" decisions and lower perceptual sensitivity. That is, KR facilitated learning to avoid false alarms, reflected in enhanced PPP and increased conservatism, but did not facilitate learning to avoid misses, manifested in lower levels of NPP and sensitivity.

The differential efficacy of KR for reducing false alarms versus misses may have resulted in part from the perceptual discrimination required. As in most vigilance experiments, stimuli were presented on the screen only for a brief duration. These brief exposure times, coupled with the spatial discrimination required, induced gamma motion in that the two vertical lines appeared to move from the center dot when they appeared on the screen. This apparent motion may have induced leniency in responding, which was reduced by the feedback. However, this task feature may have distracted observers in the KR conditions from learning the distances to be discriminated. Apparent

motion did not affect vigilance per se, given that the no-KR control results replicate those typically observed. Rather, the apparent motion may have interacted with the informational value of each form of KR, which subsequently interfered with the learning process in regard to sensitivity.

Evidence that different forms of KR are not equivalent in their informational value or in their performance effects comes from the work of Dittmar, Warm, and Dember (1985). They provided observers with feedback regarding correct detections, false alarms, or misses. Whereas hit and false alarm KR enhanced detection accuracy, miss KR failed to enhance observers' detection ability. These effects occurred despite the fact that hit and false alarm KR should, in principle, provide observers with the same informational feedback.

The relative effectiveness of KR for reducing false alarms versus misses may also depend on whether the feedback occurs after overt responding. In most vigilance experiments overt responses were required only when the observer believed a signal was present (for one exception see Parasuraman & Davies, 1976). Nonsignal events therefore require the "response of not responding," and no KR is provided for correct rejections. It may be that the KR typically used in vigilance is effective only when observers make an overt response and thereby link the feedback to a specific self-generated event. KR that cannot be so associated with a discrete action (miss KR) may therefore be less effective. Indeed, Dittmar et al. (1985) argued that miss KR does not adequately provide sufficient information to be perceptually useful. Requiring overt responses for nonsignals may therefore enhance the efficacy of KR, especially for improving NPP and sensitivity. Further exploration of this potentiality is required to establish what elements of KR produce these differential performance effects.

## Criterion Setting or Diagnosticity?

The results of the SDT analysis might tempt one to conclude that the effects of KR on PPP and NPP are driven solely by criterion setting. However, reduction in responding per se does not need to result in a change in a particular direction of either PPP or NPP. This can be seen by the example in Table 2, which shows hit and false alarm rates that produce common levels of response bias but different values of PPP and NPP, and a comparable example analysis in which $A'$ is held constant. The diagnosticity measures are determined by the relative frequencies of correct and incorrect decisions, rather than the absolute level of "yes" responses. Therefore, both the operator's perceptual ability and level of responding can influence diagnosticity.

**TABLE 2:** Diagnosticity Measures Corresponding to Hit/False Alarm Pairs With the Same Response Bias Level

| $A'$ | $\beta''_D$ | $p$(H) | $p$(FA) | PPP | NPP |
|---|---|---|---|---|---|
| .50 | .99 | .077 | .053 | .09 | .93 |
| .50 | .99 | .120 | .031 | .22 | .94 |
| .51 | .99 | .178 | .017 | .43 | .94 |
| .52 | .99 | .250 | .009 | .66 | .95 |
| .59 | .99 | .500 | .004 | .89 | .97 |
| .70 | .99 | .688 | .002 | .96 | .98 |
| .80 | .99 | .812 | $8.93 \times 10^{-4}$ | .98 | .99 |
| .86 | .99 | .875 | $8.93 \times 10^{-4}$ | .99 | .99 |
| .96 | .99 | .969 | $1.19 \times 10^{-4}$ | 1.00 | 1.00 |
| .71 | .08 | .838 | .143 | .30 | .99 |
| .71 | .27 | .812 | .116 | .33 | .98 |
| .71 | .35 | .800 | .107 | .35 | .98 |
| .71 | .60 | .762 | .071 | .43 | .98 |
| .71 | .66 | .762 | .061 | .47 | .98 |
| .71 | .76 | .738 | .046 | .54 | .98 |
| .71 | .84 | .736 | .031 | .63 | .98 |
| .71 | .95 | .712 | .010 | .84 | .98 |
| .71 | 1.00 | .700 | $8.93 \times 10^{-6}$ | 1.00 | .98 |

*Note.* Values in this table assume base rates used in this study (signal rate of .067 and a nonsignal rate of .933). PPP = positive predictive power; NPP = negative predictive power.

## KR Schedule

One goal for this study was to investigate the possibility that observers who receive partial KR training perform better during a test vigil than those who experience continuous KR training. This hypothesis was not supported, an outcome that differs from that obtained by Warm et al. (1971). They reported that partial KR training facilitated performance during a test vigil. However, they evaluated reaction time, whereas this study evaluated response accuracy. These aspects of performance are thought to reflect different perceptual processes (Santee & Egeth, 1982). Reaction time to easily detectable stimuli, such as those employed by Warm et al. (1971), likely reflects resource-limited processing. This in turn may be influenced by motivational processes. By contrast, the present detection task required a much more difficult discrimination and is much more likely to be data limited and thus not influenced by increased effort (see Norman & Bobrow, 1975). Partial KR may be more effective than continuous KR when performance is in the resource-limited range, as participants are able to modulate their effort. They may reduce their effort during the KR portions of the experiment and increase effort during the portions in which no feedback is provided. In the data-limited range, effort modulation will not influence performance, rendering the two forms of KR equally efficacious. This interpretation is consistent with an effort regulation model in which effort level is modulated according to the level of demand (see Hockey, 1997). Modulation of effort may be effective when the task is resource limited, but in the data-limited range such regulation is less flexible and therefore less beneficial.

## SDT and PPP/NPP

The effects of KR and time on task in this study were contingent upon the dependent measures employed. $A'$ showed a decline over time and a negative effect of KR during transfer, and $\beta''_D$ showed a KR effect during training and an interaction during transfer. PPP showed no time effects during either session but did show a facilitative effect for KR during transfer. The NPP analysis indicated a debilitative effect of KR during training and an interaction between KR and time during transfer. Thus, the two sets of measures did not produce the same patterns of outcome. The question is not which set of metrics is better but, rather, how the metrics may be used together to provide a more complete performance portraiture. Sensitivity provides information regarding stimulus discriminability, whereas response bias reflects the decision criterion. These measures correspond more directly to psychological processes, although they do not imply specific mechanisms. Indeed, Swets (1977) argued that there is more to vigilance than discrimination and criterion setting and that SDT cannot provide a total explanation for its performance effects. Further, sensitivity measures often reflect more than pure sensory processing ability (Pastore, Crawley, Berens, & Skelly, 2003). So, although SDT indices may tend to reflect psychological processes, they do not, for this reason alone, supersede other useful measures. The PPP/NPP approach provides a direct assessment of the accuracy of yes/no responses given particular levels of sensitivity and bias, and thereby it indicates how often the individual's decision will be correct. Given that poor diagnostic decisions can occur when signal base rates are low (Parasuraman et al., 1997), measures that directly reflect diagnostic power can be especially useful in evaluating real-world detection systems.

## Practical Applications

The Transportation Security Administration has placed over $1 billion of detection equipment into operation at over 400 airports within the United States (U.S. Government Accountability Office, 2006). The efficiency of such systems depends crucially upon the capability of operators who use them. To enhance operator efficiency, technologies such as the Threat Image Projection System (Neiderman & Fobes, 1997) overlay computer-generated images onto inspected items to provide feedback for performance improvement. Here we demonstrate that the role of feedback in vigilance can be interpreted differently, contingent on the nature of the dependent measure employed. The relative importance of diagnosticity for signal and nonsignal events is therefore critical in evaluating the effectiveness of KR manipulations as training strategies for monitoring tasks. Our results show that KR effectiveness depends in part on whether emphasis in training is on false alarm reduction or on the reduction of missed signals. For the latter, it might be advisable to require operators to make overt "no" responses during training to attenuate the decline in accuracy.

Aviation security is only one example among

many domains in which training for vigilance has practical utility. Another obvious realm of concern is detection and decision making in medicine. Although PPP and NPP are relatively new to human factors professionals, these measures are often used by health professionals to whom the human factors community is now reaching out (Bogner, 1994). Therefore, we must assure those professionals with whom we wish to interact that we can embrace and integrate their measurement approaches with those with which we are already comfortable and familiar. This study is the first direct comparison between PPP/NPP and the classic measures of signal detection in sustained attention. Our results have shown that PPP and NPP provide another window on performance and are especially useful when false alarm rate is naturally low. We hope that others will see the advantage of these differing methods and, given the simple calculations needed for their derivation, will include PPP and NPP in their own efforts to understand operators' capabilities to detect environmental signals.

## ACKNOWLEDGMENTS

Dr. William Dember passed away September 4, 2006. His co-authors wish to dedicate this paper to Professor Dember for his many years of outstanding scholarship and friendship.

Completion of this work was supported in part by the Department of Defense Multidisciplinary University Research Initiative (MURI) program administered by the Army Research Office under Grant DAAD19-01-1-0621, P. A. Hancock, principal investigator. The views expressed in this work are those of the authors and do not necessarily reflect official U.S. Army policy. The authors wish to thank Dr. Sherry Tove, Dr. Elmar Schmeisser, and Dr. Mike Drillings for providing administrative and technical direction for the grant. We are also grateful to Eric Warm, M.D., for helpful discussions regarding the application of diagnosticity measures in medicine.

## REFERENCES

Bogner, M. S. (Ed.). (1994). *Human error in medicine.* Hillsdale, NJ: Erlbaum.

Craig, A. (1978). Is the vigilance decrement simply a response toward probability matching? *Human Factors, 20,* 441–446.

Craig, A. (1979). Nonparametric measures of sensory efficiency for sustained monitoring tasks. *Human Factors, 21,* 69–78.

Davies, D. R., & Parasuraman, R. (1982). *The psychology of vigilance.* London: Academic Press.

Dittmar, M. L., Warm, J. S., & Dember, W. N. (1985). Effects of knowledge of results on performance in successive and simultaneous vigilance tasks: A signal detection analysis. In R. E. Eberts & C. G. Eberts (Eds.), *Trends in ergonomics/human factors II* (pp. 195–202). Amsterdam: Elsevier.

Getty, D. J., Swets, J. A., Pickett, 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.

Hancock, P. A., & Hart, S. G. (2002). Defeating terrorism: What can human factors/ergonomics offer? *Ergonomics in Design, 10*(1), 6–16.

Harris, D. (2002). How to really improve airport security. *Ergonomics in Design*, *10*(1), 17–22.

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology, 45,* 73–93.

Linton, C. S. (1996). General internal medicine. In U. B. S. Prakash (Ed.), *Mayo Internal Medicine Board review 1996–1997* (pp. 333–352). Rochester, MN: Mayo Foundation for Medical Education and Research.

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

McCormack, P. D., Binding, F. R. S., & McElerhan, W. G. (1963). Effects on reaction time of partial knowledge of results of performance. *Perceptual and Motor Skills, 17,* 279–281.

Neiderman, E. C., & Fobes, J. L. (1997). Operational test and evaluation of human factors interventions for airport security screeners. In *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting* (pp. 1094–1097). Santa Monica, CA: Human Factors and Ergonomics Society.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology, 7,* 44–64.

Parasuraman, R., & Davies, D. R. (1976). Decision theory analysis of response latencies in vigilance. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 578–590.

Parasuraman, R., Hancock, P. A., & Olofinboba, O. (1997). Alarm effectiveness in driver-centered collision warning systems. *Ergonomics, 39,* 390–399.

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric" A′ and other modern misconceptions about signal detection theory. *Psychonomic Bulletin and Review, 10,* 556–569.

Salmoni, A. W., Schmidt, R. A., & Walter, C. B. (1984). Knowledge of results and motor learning: A review and critical reappraisal. *Psychological Bulletin, 95,* 355–386.

Santee, J. L., & Egeth, H. E. (1982). Do reaction time and accuracy measure the same aspects of letter recognition? *Journal of Experimental Psychology: Human Perception and Performance, 8,* 489–501.

See, J. E., Howe, S. R., Warm, J. S., & Dember, W. N. (1995). Meta-analysis of the sensitivity decrement in vigilance. *Psychological Bulletin, 117,* 230–249.

See, J. E., Warm, J. S., Dember, W. N., & Howe, S. R. (1997). Vigilance and signal detection theory: An empirical evaluation of five measures of response bias. *Human Factors, 39,* 14–29.

Swets, J. A. (1977). Signal detection theory applied to vigilance. In R. R. Mackie (Ed.), *Vigilance: Operational performance and physiological correlates* (pp. 705–718). New York: Plenum Press.

Szalma, J. L, Miller, L. C., Hitchcock, E. M., Warm, J. S., & Dember, W. N. (1998). Intraclass and interclass transfer of training for vigilance. In M. W. Scerbo & M. Mouloua (Eds.), *Automation technology and human performance: Current research and trends* (pp. 183–187). Mahwah, NJ: Erlbaum.

U.S. Government Accountability Office. (2006). *Transportation security administration: Oversight of explosive detection systems maintenance contracts could be strengthened* (Report to Congressional Committees, GAO-06-795). Washington, DC: Author.

Warm, J. S. (1993). Vigilance and target detection. In B. M. Huey & C. D. Wickens (Eds.), *Workload transition: Implications for individual and team performance* (pp. 139–170). Washington, DC: National Academy Press.

Warm, J. S., Hagner, G. L., & Meyer, D. (1971). The partial reinforcement effect in a vigilance task. *Perceptual and Motor Skills, 32,* 987–993.

Warm, J. S., & Jerison, H. J. (1984). The psychophysics of vigilance. In J. S. Warm (Ed.), *Sustained attention in human performance* (pp. 15–59). Chichester, UK: Wiley.

James L. Szalma is an assistant professor of psychology at the University of Central Florida. He received his Ph.D. in experimental psychology/human factors from the University of Cincinnati in 1999.

Peter A. Hancock is the Provost Distinguished Professor of Psychology at the University of Central Florida. He received his Ph.D. in motor performance from the University of Illinois at Urbana-Champaign in 1983.

Joel S. Warm is a professor of psychology at the University of Cincinnati. He received his Ph.D. in experimental psychology from the University of Alabama in 1966.

William N. Dember was dean emeritus and professor emeritus at the University of Cincinnati. He received his Ph.D. in experimental psychology from the University of Michigan in 1955.

Kelley S. Parsons is a graduate student in the doctoral program in experimental psychology/human factors at the University of Cincinnati, where she received an M.A. in experimental psychology/human factors in 2001.

*Date received: February 17, 2004*
*Date accepted: August 15, 2005*