Effects of Training with Knowledge of Results on Diagnosticity in Vigilance Performance

Joel S. Warm\textsuperscript{1}, Victor Finomore\textsuperscript{1}, Tyler H. Shaw\textsuperscript{2}, Matthew E. Funke\textsuperscript{3}, Michelle J. Hausen\textsuperscript{3}, Gerald Matthews\textsuperscript{3}, Purcell Taylor\textsuperscript{3}, Michael A. Vidulich\textsuperscript{1} Daniel W. Repperger\textsuperscript{1}, James L. Szalma\textsuperscript{4}, and Peter A. Hancock\textsuperscript{4}

Air Force Research Laboratory, Wright-Patterson AFB\textsuperscript{1}; George Mason University, Fairfax, VA\textsuperscript{2}; University of Cincinnati, Cincinnati, OH\textsuperscript{2}; University of Central Florida, Orlando, FL\textsuperscript{4}

Making accurate diagnostic decisions about signal presence/absence is critical for success in many failure intolerant monitoring technologies requiring sustained attention or vigilance. This study examined the effects of training with knowledge of results (KR) on observer diagnosticity in a vigilance task. Diagnosticity was measured in terms of decision theory measures of positive predictive power (PPP) – precision in indicating when signals were actually present and negative predictive power (NPP) - precision when indicating signal absence. Initial training with KR enhanced observers’ diagnosticity on a subsequent test task in terms of PPP but not NPP. The picture of performance efficiency reflected by both diagnostic measures differed from results indexed by signal detection theory (SDT) measures of perceptual sensitivity (d’) and response bias (c). However as predicted from the computational mechanics of the decision theory and SDT measures, both diagnostic measures correlated positively with d’ while NPP correlated negatively with c. These findings indicate that combinations of perceptual ability and level of responding can influence the behavioral metrics signifying diagnosticity in vigilance performance.

Sustained attention or vigilance is a central aspect of operator performance in many human-machine systems including those involving military surveillance, air-traffic control, airport and border security, industrial quality control, and medical monitoring (Warm, Parasuraman, & Matthews, 2008). Both minor and major accidents have been attributed to vigilance failures on the part of human operators in such settings (Hawley, 2006; Molloy & Parasuraman, 1996). Therefore, the ability to understand the factors that influence vigilance performance and to train and evaluate the effectiveness of operators who are engaged in vigilance tasks are critical human engineering concerns for system reliability and for public safety and health (Nickerson, 1992).

The importance of veridical decisions in operational vigilance assignments raises the question of the diagnostic accuracy of those trained to perform vigilance tasks, i.e., the proportion of time that their decisions about signal present or absent are actually correct. Specifically, it is vital that when an operator indicates that a signal is present it is actually present and when an operator indicates that a signal is not present it is imperative that it not be there. One way to measure diagnosticity might be to use signal detection theory (SDT) measures of perceptual sensitivity (e.g., d’) and response bias (e.g., β or c). Sensitivity reflects the ability of observers to discriminate signals and non-signals while response bias reflects their willingness to emit detection responses (Macmillian & Creelman, 2005). Studies have demonstrated that knowledge of results (KR) -- an enhancement intervention that has been applied to an extensive array of situations and is the principal vehicle for preparing observers to perform vigilance tasks (Kluger & DiNisi, 1996; Warm & Jerison, 1984) -- improves perceptual sensitivity and promotes the setting of optimal levels for the response criterion (Dittmar, Warm, & Dember, 1985; Szalma, Hitchcock, Miller, Warm, & Dember, 1999). Thus, one might infer that the increased level of perceptual sensitivity and the more optimum criterion settings brought about by training with KR would enhance the diagnostic power of observers.

It is important to note that when used in this way, the SDT measures are indirect indices of diagnosticity and that the inference on which their diagnostic value is based may not always be correct. For example, while one might expect greater diagnostic accuracy from high d’ scores, Parasuraman, Hancock, and Olofinoba (1997) have demonstrated that when signal probability is low, as in most operational situations, observers with a high level of sensitivity can still make poor diagnostic decisions by generating unacceptably large numbers of false alarms. This finding led Szalma, Hancock, Warm, Dember, and Parsons (2006) to propose that diagnosticity in vigilance might be captured more directly by decision theory measures commonly used in medicine (Linton, 1996) known as positive predictive power (PPP), the proportion of signal present responses (“Yes” responses) that are actually correct and negative predictive power (NPP), the proportion of signal absent responses (“No” responses) that are actually correct.

The computation formula for PPP is $H/(H + FA)$ in which $H = \text{the number of signals detected (hits)}$ and $FA = \text{the number of false alarms}$. A PPP score of 1.0 indicates a perfectly accurate observer; a score of 0 would indicate no correct detections and no diagnosticity. The comparable computational formula for NPP is $CR/(CR + M)$ in which $CR = \text{the number of correct rejections}$ and $M = \text{the number of signals missed}$. An observer who correctly rejected all nonsignals and committed no misses would achieve an NPP score of 1, while a score of 0 would indicate that no correct rejections were made and no diagnosticity.

To compare the diagnostic utility of the SDT and the decision theory measures, Szalma and his associates (2006) trained observers with or without composite KR (information about hits, misses, and false alarms) and then tested them in the absence of KR. They found that KR-training induced
conservatism in responding but did not enhance perceptual sensitivity and that KR-training enhanced PPP but that it lowered the level of NPP. Clearly, the Szalma et al. (2006) study showed disengagement between the SDT measures and the decision theory measures. In considering that result, Szalma and his associates affirmed that a goal for any training regimen is that it maximize diagnostic power and that the SDT and decision theory measures can augment one another in terms of diagnosticity since they provide alternate views into the character of detection capability. The SDT measures reflect sensory processing and decision making while the decision theory measures provide a direct metric of the accuracy of “Yes” and “No” responses derived from those psychological processes. Szalma and his associates also pointed out that while the SDT and decision theory measures do not correspond completely, there should be some overlap between them. More specifically, since the diagnostic measures are determined by the relative frequencies of correct and incorrect responses the observer’s perceptual ability and level of responding can influence diagnosticity. With the computational mechanics of the decision theory measures in mind, consider the following examples. Given a fixed level of conservatism, one might anticipate that the higher the d’, the greater the hit rate and the lower the miss rate, a combination that would result in higher PPP and NPP scores. Given a fixed non-zero level of perceptual sensitivity, one might anticipate that the higher the conservatism score, the lower the false alarm rate and the higher the miss rate, a combination that would result in an increased PPP score but a decreased NPP score. The present study was designed to test these possibilities.

Method

Twenty-eight undergraduates from the University of Cincinnati, 14 women and 14 men, served as observers for course credit. All observers had normal or corrected-to-normal vision and were free of any known hearing impairments. None of the students had participated previously in vigilance experiments. Fourteen observers equated for sex were assigned at random to each of two training conditions, one in which composite KR was provided and one in which it was not.

All observers took part in a 50-minute session which consisted of a training phase divided into four continuous 5-min periods of watch followed by a test phase of six continuous 5-min periods of watch. They assumed the role of airport baggage inspectors looking for explosives on a video display terminal. As shown in Figure 1, the display consisted of a pair of adjacent gray circles arrayed along a vertical vector that appeared against a gray background. In the normal case (no explosive present), both circles in a pair could either be medium gray (Michaelson contrast ratio with the background = .50) or dark gray (Michaelson contrast ratio with the background = .70) and the circle pairs could appear at one of five locations on the screen, center, upper left, upper right, lower left, and lower right. Location was determined at random for each observer with the restriction that all five positions occurred equally often during each watchkeeping period in each phase of the experiment.

In all conditions, the circle pairs were exposed at a rate of 30 events/min with a dwell time of 300 msec. Critical signals for detection (explosive present) were cases in which one of the circles in a pair was medium gray and the other was dark gray. Ten critical signals were presented per period of watch in all experimental conditions. The spatial location of the signals was varied at random for each observer with the restriction that two signals appeared at each location and that at each location, the critical signal combinations, dark gray/light gray and light gray/dark gray occurred equally often. Observers indicated their detection of critical signals by pressing a computer space bar. Responses occurring within 1000 msec after the offset of a critical signal were recorded automatically as correct detections. All other responses were recorded as false alarms. During the training phase of the study, KR was provided by a prerecorded computerized female voice (62.8 dB) announcing “Hit”, “False Alarm,” or “Miss”. To control for accessory stimulation, observers in the non-KR training condition heard the message “Save” after each response. The KR and acknowledgement messages were not provided during the test phase of the study. Training and test phases were separated by a 10 min rest interval.

Results

Signal Detection Theory Measures. The percentages of correct detections and false alarms were used to determine SDT measures of perceptual sensitivity (d’) and response bias (c) in all experimental conditions. The measure c was employed instead of the more familiar index, β, because of evidence indicating that c is a more effective indicator of response bias in vigilance studies (See, Warm, Dember, & Howe, 1997).

Mean d’ scores during the training and testing phases of the experiment are present in Figure 2. The data for the KR and NKR groups are plotted as a function of periods of watch within each phase. Error bars are standard errors.
An analysis of variance (ANOVA) of the data for the training phase revealed no significant overall effects for groups or periods of watch, $p > .05$ in both cases. However, the interaction between these factors was significant, $F(2.55, 66.30) = 4.70, p < .007, \eta^2_p = .153$. It is evident in the figure that the perceptual sensitivity scores declined over time in the NKR group while they increased over time in the KR group. In this and all subsequent ANOVAs, the Box correction was applied to compensate for violations of the sphericity assumption (Maxwell & Delaney, 2004).

A similar ANOVA of the $d'$-data for the testing phase revealed significant main effects for KR, $F(1, 26) = 15.28, p < .01, \eta^2_p = .370$ and for periods of watch, $F(4.16, 108.20) = 2.65, p < .05, \eta^2_p = .092$. The interaction between these factors lacked significance, $p > .05$. It can be seen in the figure that overall perceptual sensitivity during the testing phase of the study was greater for observers trained with KR than for those who did not receive KR training and that in both groups, perceptual sensitivity declined over the course of the testing phase.

Mean $c$ scores during the training and testing phases of the experiment are presented in Figure 3. As in Figure 2, the data for the KR and NKR groups are plotted as a function of periods of watch within each phase and the error bars are standard errors.

Inspection of the figure will reveal that during training, the overall level of conservatism was greater for observers who received KR than for those who did not and that for both groups, the levels of conservatism in responding increased over time. These impressions were confirmed by an ANOVA of the $c$-scores in which the main effect for periods of watch was statistically significant, $F(2.89, 75.1) = 6.90, p < .001, \eta^2_p = .210$, while that for the training condition closely approached significance, $F(1, 26) = 4.05, p = .055, \eta^2_p = .135$. The KR x Periods interaction was not significant, $p > .05$.

The trends with regard to the response criterion noted during training, continued into the testing phase of the study. Once again, a higher overall level of conservatism was noted for the KR as compared to the NKR group, $F(1, 26) = 6.73, p < .05, \eta^2_p = .206$, and the level of conservatism across groups increased significantly over the course of the testing phase, $F(3.90, 101.29) = 2.95, p < .05 \eta^2_p = .102$. As in the training phase, the interaction between these factors was not significant, $p > .05$.

**Decision Theory Measures.** Mean PPP values during the training and testing phases of the experiment are presented in Figure 4. The data for the KR and NKR groups are plotted as a function of periods of watch within each phase. Error bars are standard errors.

Perusal of Figure 4 will reveal that at the outset of the training phase of the study, the diagnostic accuracy of both the KR and NKR groups was quite low. In both groups, the diagnostic accuracy of their “signal present” or “Yes” responses was only 50%. However, while participants in the NKR group remained at that level of diagnostic accuracy throughout the training session, the PPP scores for participants receiving KR grew dramatically over the course of the training session reaching a diagnostic accuracy level of approximately 80% by the end of training. An ANOVA of the PPP data during training revealed significant main effects for training condition $F(1, 26) = 4.17, p = .05, \eta^2_p = .138$ and for periods of watch, $F(2.34, 60.84) = 7.37, p < .01, \eta^2_p = .221$ and most critical for changes in the rate of gain in PPP over time, a significant interaction between KR and periods of watch, $F(2.34, 60.84) = 3.81, p < .05, \eta^2_p = .128$.

Figure 4 also reveals that the diagnostic advantage given to the KR group during training continued into the testing phase of the study. The scores for the NKR group remained at the level observed continuously during training, while those for the KR group remained at the final level obtained during training. An ANOVA of the testing data for the PPP measure revealed a significant main effect for KR, $F(1, 26) = 19.63, p < .001, \eta^2_p = .430$. The other sources of variance in this analysis were not significant, $p > .05$ in each case.
Mean NPP values during the training and testing phases of the experiment are presented in Figure 5. The data for the KR and NKR groups are plotted as a function of periods of watch within each phase. Error bars are standard errors.

Contrary to the early results with the PPP measure, Figure 5 reveals that at the outset of the experiment, the diagnostic accuracy of both experimental groups with the NPP measure was quite high. The diagnostic accuracy of their “signal absent” or “NO” responses approximated 99%. The only significant source of variance in an ANOVA of the NPP training data was the KR x Periods interaction, $F(2.68, 69.56) = 3.06, p < .05, \eta^2 = .105$. It appears in the figure that the scores for the KR group remained relatively stable across the four training periods while those for the NKR group declined over time. However, even with this decline, the diagnostic accuracy of the NKR group remained quite high. Figure 5 also shows that high levels of diagnostic accuracy for the two experimental groups were also evident during the test phase of the study. An ANOVA revealed a statistically significant decline in NPP scores over time, $F(3.85, 100.07) = 4.61, p < .05, \eta^2 = .151$. The other sources of variance in this analysis were not significant, $p > .05$ in each case.

**Intermeasure Correlations.** Table 1 presents the correlations between the SDT and decision theory measures for the KR and NKR groups during the training and testing phases of the study. Also presented are the intermeasure correlations within each phase of the experiment when the data are combined across the training groups. The table reveals a consistent pattern of correlations in both phases of the experiment with and without KR-training. As anticipated, $d'$ correlated positively with the PPP and the NPP measures, while $c$ correlated positively with PPP and negatively with NPP. This pattern shows up quite clearly when the data are combined across the KR and NKR groups.

**Discussion**

The present study was designed to follow the lead of Szalma and his associates (Szalma et al., 2006) and provide further information about the manner in which decision theory metrics and traditional SDT measures can augment each other in indexing diagnostic accuracy in vigilance following training with KR. The results indicate that the measures are indeed related but that they also provide different views into detection capability resulting from the training regimen.

As anticipated from the computational mechanics of the decision theory measures, PPP correlated positively with $d'$ and $c$ while NPP correlated positively with $d'$ and negatively with $c$, thus indicating that the observer’s level of perceptual ability and level of responding can influence diagnosticity.

Consistent with earlier results (Dittmar et al., 1985; Szalma et al., 1999), KR enhanced perceptual sensitivity and increased conservatism in responding during training and these effects continued into the testing phase of the study. Of critical importance, were the temporal changes that occurred with the SDT measures during the test phase. Training with KR did not counter the vigilance decrement, perceptual sensitivity declined over time in both the KR and NKR groups, and KR did not alter the growth in conservatism over time seen in both the KR and NKR groups. A temporal decline in perceptual sensitivity and a temporal increase in response conservatism are common findings with SDT measures in vigilance (Matthews, Davies, Westerman, & Stammars, 2000; See, Howe, Warm, & Dember, 1995).

In accord with the initial findings by Szalma et al. (2006), training with KR enhanced the overall level of PPP during the test phase of the study. As in the case of the perceptual sensitivity measure, KR did not produce a differential effect in the training groups over time in the test phase. However, contrary to the results for perceptual sensitivity, the PPP scores for both the KR and NKR groups remained stable over time during the testing phase of the study. Since the SDT measures and the PPP measure are determined by the frequencies of correct detections and false alarms, the temporal stability of the PPP measure during the test phase can be explained by the SDT dynamics — a temporal decline in the frequency of hits reflected in the decline in $d'$, is compensated by a reduction in the frequency of false alarms,
reflected in the rise in c. Moreover, the SDT dynamics can also explain the temporal decline in NPP—an increase in conservatism increases the likelihood of missed signals and renders observers’ decisions about signal absence less precise over time.

Viewed in this way, the complementary relation between the decision theory index and the SDT measures becomes apparent. Given a particular level of sensitivity and bias, the PPP/NPP approach provides direct metrics of the accuracy of the observer’s “Yes” and “No” responses and therefore, how often the observer’s decisions about the presence and absence of critical signals will be correct over the course of the watch. The SDT measures provide insight into the perceptual and response bias mechanisms that underlie those decisions. Both types of measures can be useful in evaluating real-world detection judgments. PPP/NPP afford immediate metrics of diagnostic power while the SDT measures show the dimensions in which experimental manipulations can preserve that power or alter it.

The ability of KR to enhance the overall level of PPP coupled with its ability to enhance the overall level of perceptual sensitivity and the level of caution in responding implies that observers in the present study learned to identify critical signals or dangerous events (“explosives present”) and to avoid false alarms. However, the lack of KR effects on the NPP measure implies that the beneficial effects of KR training did not also extend to the ability to identify the presence of nonsignals or safe events (“explosives absent”) or to avoid misses. As in most vigilance studies and in the majority operational situations as well, the number of nonsignals in this investigation far exceeded the number of signals. In situations of this sort, observers have much more familiarity with nonsignals than they do with signals and opportunities for errors of omission or misses are much rarer than the opportunities for false alarms. As a result, the NPP measure which encompasses correct rejections and misses may be less responsive to the effects of KR and other experimental interventions than the PPP measure in which hits and false alarms are featured. Thus, due to the low signal base rate, NPP is not likely to be as useful a measure as PPP as an index of diagnostic power in vigilance.

It is noteworthy, that the correlations between the SDT and decision theory measures were observed in both the KR and No-KR conditions. Consequently, these correlations reflect general patterns that may have important implications for operational situations in which the SDT and decision theory measures are jointly employed. The positive correlations between d’ and PPP and NPP imply that individuals who are high in perceptual sensitivity will be more accurate in making both signal present and signal absent decisions. The positive correlation between c and PPP implies the same thing. However, the negative correlation between c and NPP suggests that those who approach the vigilance task with a more stringent response criterion will be more prone to errors in deciding that a signal is truly absent. This suggests that thought needs to be given to the observer’s response level in situations in which the importance of diagnostic powers in the PPP an NPP realms may differ.

References


