

**SUPPORTING SITUATION ASSESSMENT THROUGH ATTENTION  
GUIDANCE: A COST-BENEFIT AND DEPTH OF PROCESSING ANALYSIS**

William J. Horrey & Christopher D. Wickens  
University of Illinois at Urbana-Champaign

Automated support systems may be useful tools for aiding situation assessment in complex environments such as the military battlefield. These environments are marked by large amounts of information which often must be weighted and integrated into a meaningful judgment or assessment. The present research examines the effects of attention cueing on information integration tasks in static battlefield situations. Sixteen participants completed a resource allocation task for 56 battlefield scenarios (based on perceived threats). For half the trials, an automated system guided their attention to high-threat units. On 2 trials a memory probe was administered to assess the depth of processing of information, and on the final trial an automation failure was presented. Results demonstrated an overall allocation performance advantage for automation but poorer recall for automation-enhanced units. Half of the participants failed to attend to the system failure. Those participants who detected the failure were inferred to have processed the cues more deeply on the memory trials. The costs and benefits of automated cueing are discussed.

Battlefield commanders' situation awareness often involves the integration of large amounts of information from a number of sources in order to form a situation assessment (Graham & Matthews, 1999). This weighted information includes the location and strength of other friendly and opposing forces, the surrounding terrain, and a large number of other METT-T (Mission, Enemy, Terrain, Troops, and Time) operational factors. Previous work has shown that people do not always integrate multiple pieces of information optimally (when making a judgment or decision), especially under conditions of high workload, time pressure, or when the information is unreliable in nature, conditions which are characteristic of the battlefield environment.

Employing the taxonomy proposed by Parasuraman, et al. (2000), automation can be provided to assist the battlefield commander in this task at various stages of information processing, for example in guiding attention to the most valuable cues (stage 1), in integrating cues and diagnosing what automation infers to be the most likely state of intent (stage 2), or in recommending the most appropriate course of action (stage 3). However limitations of automatic diagnosis and choice have been found in operator over-reliance upon imperfect automation (Parasuraman & Riley, 1997; Metzger and Parasuraman, 2001, Mosier, et al., 1998). Thus we focus our interest on automation at the first stage, to assist the operator by highlighting the most relevant cues for situation awareness (SA) or assessment. Unlike automated situation assessment and choice, which allow the operator to perform without necessarily attending to the cues, stage 1 automation requires the operator to consider at least some of the cues upon which the diagnosis is based. Research on target cueing (a form of attention guidance) has reliably demonstrated the benefits of automation. Nevertheless such highlighting or attention cueing has been found to produce unwanted effects on attentional tunneling (e.g., Yeh, et al., 1999; Yeh and Wickens, 2001, in press; Metzger & Parsuraman, 2001; Davison & Wickens, 2001), and over-reliance.

While past research on automation attention guidance has focused on target detection tasks (e.g., Yeh, et al., 1999;

Davison & Wickens, 2001), the current research examines stage one attention cueing in an information integration task (i.e., Endsley's (1995) level 2 SA) where all the raw data are available (the cues highlight the most relevant information). Specifically, we assessed the effects of an automated cueing aid in a static battlefield map display on (a) the assessed threat of enemy attack from the east and west (we operationally measure this assessment by the participants' subsequent deployment of defensive resources), (b) the depth of processing of raw data (for high and low relevant information, cued and uncued), and (c) over-reliance on imperfect automation (the participant's reaction to the automation's failure to cue a high relevant piece of information).

Participants under time pressure observed map displays which contained large amounts of information (regarding the type, location, strength, and accessibility of other military units, as well as the reliability of the information source). On some trials, the cueing aid highlighted the enemy units that were most relevant (i.e., had the highest information value) to the participant's threat assessment and was intended to help the observers filter out the less relevant information (e.g., neutral or other friendly units). We hypothesized that the filtering effects of the automated aid would allow participants to make more optimal defensive allocations.

Memory probes were used on some trials to assess differential effects of automated cueing on the depth of information processing ( Craik & Lockhart, 1972) for a particular unit (i.e., whether cueing would increase or decrease the memory for separate attributes of the cued target; Yeh and Wickens, 2001, in press). It was also predicted that the failure of automation to highlight a relevant cue would result in a high number of misses and hence an inappropriate allocation of resources.

Finally, we were interested in whether certain information cue types would be intrinsically given more weight in the threat assessment, independent of the level of automation and of their information value. In particular, we wished to see whether the abstract cue of reliability would receive less processing, and therefore contribute less weight to the

judgments than the more concrete cues of unit size, distance and terrain.

## Methods

Eight upper level Army ROTC students and eight non-ROTC (graduate) students from the University of Illinois volunteered for this study. Participants were presented with 56 digital battlefield scenarios based on topographical maps of Fort Irwin and standard military symbology for enemy, neutral, and friendly units (see Figure 1). These units were embedded within these maps and varied in size (e.g., platoon), type (e.g., enemy combat mechanized), location (distance and terrain separation from own forces), and the reliability of the intelligence estimate of their identity as coded by line type of the symbol (solid, dashed, dotted). For non-ROTC students, a numerical digit replaced the standard symbology for unit strength. The participant's own unit was always located near the center of the map.

On each trial, participants had 20 defensive resources to deploy either to the east or west of their position. Participants were required to evaluate the integrated threat of units in the east versus those in the west and allocate these resources accordingly. Optimally, a large threat from the east would receive a larger proportion of these resources than would a lower perceived threat from the west. The relative threat of each unit was based on weighted evidence from multiple attributes of each unit (unit type and size, separation distance (relative to their own position), difficulty of the terrain between the unit and themselves, and the reliability of the intelligence assessment of the unit's identity). For each trial, the map was displayed for up to 25 seconds.

*Automation.* On half of the trials, an automation feature was incorporated which guided attention to the most relevant (highest threat) symbols on the map by flashing them. The relevance of a symbol was based on its information value (units having higher information value were deemed to be more of a threat) and this information value (Barnett & Wickens, 1988) was based on the 5 attributes characterizing

neutral forces. This is a simplified trial - typical scenarios had 20 other units on the display.

each unit and was derived through a multiple regression of questionnaire data from six independent observers:

$$IV_{\text{unit}} = X_{\text{type}}(90 + 4 X_{\text{size}} - 5 X_{\text{dist}} - 14 X_{\text{diff}}) \times R \quad (1)$$

where  $X_{\text{size}}$ ,  $X_{\text{dist}}$ , and  $X_{\text{diff}}$  define the unit size, distance from own force, and difficulty of the intervening terrain, respectively.  $R$  is the overall reliability of the information for the unit (from 0 to 1), and  $X_{\text{type}}$  is the type (1 for enemy units, 0 for neutral or friendly). It follows from this formula that only enemy units will be perceived as a threat, and threat increases as unit size increases, separation distance decreases, and terrain difficulty eases. Reliability is used as a moderator variable. Terrain difficulty was rated on a 4-point scale by a group of four independent observers. The automation feature enhanced symbols that had information values equal to or greater than an arbitrary figure of 30, leading to approximately 25% of the items being highlighted on average.

*Memory Probe.* A memory probe was administered following two of the scenarios in order to determine the depth of processing of the raw data. The probe queried the size of a unit at a particular location in the battlefield display. One probe followed a non-automated trial (no enhancement), while another followed an automated trial (queried either an enhanced, high-relevance symbol or a non-enhanced, low-relevance symbol). Responses were scored on the basis of accuracy and degree of confidence.

*Failure.* One scenario was presented in which the automation feature failed to enhance all of the relevant units. On this trial, the enhancement appeared normal for all of the units favoring attack in one direction however did not highlight a very important (high-relevance) unit on the opposite side (one which would have a significant impact on the allocation of resources). The purpose of this trial was to determine whether participants were attending to all of the raw data on automated trials or only the enhanced units, an assessment that could be made based on the pattern of participants' allocation response.

## Results

Equation (1) was used to predict the optimal allocation based on the sum of the information values for the various units displayed on the map (comparing east versus west). Participant allocation responses were compared to the predicted values and expressed as absolute difference (error) scores in the analyses.

Overall, allocation policies were closer to the optimal level for trials with automation ( $\underline{M} = 2.66$ ) versus those without automation ( $\underline{M} = 3.05$ ;  $F(1, 745) = 6.3$ ,  $p = .01$ ). Non-ROTC (graduate) students were found to have lower error scores ( $\underline{M} = 2.63$ ) than ROTC students ( $\underline{M} = 3.0$ ) (see Figure 2). The Student  $\times$  Display interaction was not significant ( $F(1, 745) = .01$ ,  $p = .93$ ), suggesting that both groups benefited equally from automation, and further analysis suggested that



Figure 1. Sample battlefield scenario (colors inverted). Central symbol is observer's own unit. Surrounding units are comprised of enemy, friendly and

the two groups responded in a qualitatively similar way to other manipulations.

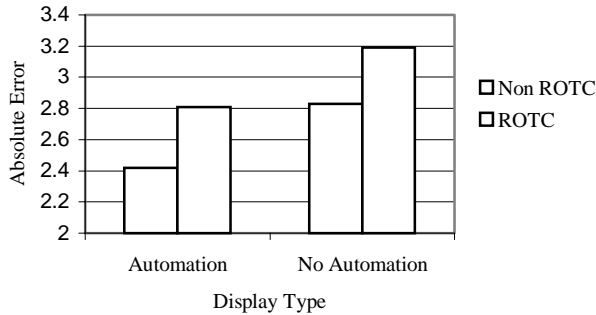


Figure 2. Absolute error by display type and student type.

Response times were found to be significantly faster in the automation condition ( $\bar{M} = 18.7$  s) than in the no-automation condition ( $\bar{M} = 20.2$  s).

*Memory Probe.* A main effect for unit relevance on the accuracy of recall for unit attributes approached significance ( $F(1, 27) = 3.46, p = .07$ ) suggesting that participants adopted the appropriate strategy of processing highly important cues more deeply ( $\bar{M} = 5.9$ ) than less important ones ( $\bar{M} = 4.2$ ).

Memory recall performance for the low relevance objects was equal, regardless of automation condition ( $\bar{M} = 4.2$ ). Analyses of the raw scores indicated that performance for these units were above chance performance. Because the unit was not highlighted in both of these conditions, this suggests that the depth of processing for these cues was not hindered by the presence of automation for other items. This finding is not consistent with the findings from other research that the presence of cued targets detracts attention from non-cued objects (e.g., Yeh, et al., 1999; Yeh and Wickens, 2001, in press).

Recall for the high relevance item was slightly weaker with automation ( $\bar{M} = 5.5$ ) compared to the no automation ( $\bar{M} = 6.5$ ) condition. Performance for this high-relevant, cued memory probe was characterized by a bimodal distribution, with participants typically scoring either very high or very low in the automated condition. The high variance from this response pattern barred any significant findings, but is of considerable interest in its own right suggesting that some participants may have ignored the raw data behind the highlighted cue entirely, integrating only the fact of its highlighting, whereas others used the highlighting as a guide for deeper analysis of the threat that had been highlighted. These two strategies correspond to the effects of cueing that Yeh and Wickens (2001, in press) had associated with response bias (reduced beta) and increased sensitivity ( $d'$ ), respectively.

*Failure Trial.* On the failure trial, roughly half of the participants failed to notice the high-relevance, non-enhanced unit (as inferred from their allocation score). We examined whether there was any significant relationship between

performance on the failure trial and the bimodal pattern of responses on the memory probe for the high-relevance, cued target. A point biserial correlation between observer type (failure noticer, non-noticer) and performance on the memory probe revealed a significant relationship ( $r_{pb} = .69, p < .05$ ) between the two variables. Noticers were then those who had shown deeper processing of the data underlying the cues. It was estimated that 63% of the variance in memory probe performance was accounted for by observer type.

*Cue Weighting.* Further analyses of the impact of the separate attributes (size, type, terrain, distance) in allocation judgments suggested that different cue types were weighted differentially in the allocation responses ( $F(3, 231) = 9.3, p < .001$ ). post hoc tests revealed that unit size ( $\bar{M} = 5.8$ ) had significantly higher influence on the resource allocation response patterns than did reliability ( $\bar{M} = 3.2; p = .004$ ), terrain ( $\bar{M} = 3.1; p < .001$ ), or distance ( $\bar{M} = 3.1; p < .001$ ). This rank order of cue influence (size-reliability-terrain/distance) that was inferred from the objective performance data is not entirely consistent with subjective self-reported importance, as measured in the post-experimental questionnaire. Participants indicated that size was the most important factor ( $\bar{M} = 4.4$ ), followed by distance ( $\bar{M} = 4.0$ ), terrain ( $\bar{M} = 3.9$ ), and reliability ( $\bar{M} = 3.2$ ).

## Discussion

In the non-automated condition, performance on the allocation task was reasonable, suggesting that there was some processing of the numerous information cues in the time available. However overall, performance with the automated cueing aid was superior to unaided performance, with reduced departures from the optimal allocation scores in automated conditions. Though there were differences in performance across student type, the automation benefited both groups equally. The response times with the aid were 1.5 sec shorter than for the non-automated conditions suggesting that automation allowed the participants to make more speeded and accurate allocation decisions, presumably by allocating their attention (visual search) initially to the cued items. In general, this finding is consistent with previous research on reliable target cueing (e.g., Yeh, et al., 1999; Davison & Wickens, 2001), however it extends beyond simple detection tasks to higher-level integration tasks.

Previous research has shown that the presence of cued targets detracts attention from other uncued targets (e.g., Davison & Wickens, 2001; Yeh, et al., 1999; Yeh and Wickens, 2001, in press). This finding was not replicated in the present study. Recall scores for the low-relevant (uncued) units were equal in both the automated and non-automated conditions but still above chance performance, suggesting that the presence of automated cues did not have an adverse impact on processing for these units. The inconsistencies in the impact of automation on uncued targets may be due, in part, to the nature of the tasks employed. As mentioned previously, most research has utilized target detection tasks (level 1 SA) to demonstrate the tunneling of attention around cued target locations. The current study, however, required participants to

integrate multiple pieces of information (level 2 SA), which typically involved more than one cued target per trial and hence, more scanning behavior. Furthermore, the amount of reduction in RT allowed by the cueing, 1.5 seconds, was sufficiently small to suggest that it did not eliminate inspection of the uncued items altogether, a conclusion also supported by the above chance accuracy of memory for those uncued items.

Recall for the attributes of the high-relevant unit exhibited a somewhat different pattern of results. The general (non-significant) trend showed inferior recall in the automated condition compared to the baseline condition, suggesting that the application of automated cueing to these high importance targets may negatively impact the depth of processing for these cues. More important was the evidence of a bimodal response pattern in the recall scores for the cued high relevance units. This suggests that different observers adopted different strategies for interacting with the early stage automation. This hypothesis was further supported by the findings from the failure trial. Observers who had poor recall for the cued target may have failed to attend to the raw data present in the display, attending only to the highlighting. For example, they may have noted the presence of 2 cued targets in the west and 4 cued targets in the east and proceeded to allocate twice as many resources to the east without processing these cues at a deeper level. Yeh and Wickens (2001, in press) found a similar response bias (beta) in observers who believed the automated system to be highly reliable. In their study, participants were found to attend more to the information suggested by the cue rather than to the raw data.

In contrast, observers who exhibited good recall may have been using the cueing to direct their attention to the relevant features for deeper analyses. This strategy would suggest an increase in sensitivity to the information in the cued target. No differences were found to suggest a demographic variable which could account for the observer type. Are there any implications of these differing beta and  $d'$  strategies in the use of automation? The former (beta) may be a more efficient strategy under time pressure however there will be costs if automation is unreliable, an issue we turn to in the next section.

*Failure Trial.* This catch trial exhibited some degree of evidence for automation induced complacency or over-reliance. Roughly half of the participants failed to notice the automation failure and hence made inappropriate allocation responses. On all trials prior to the failure trial, the automation had operated reliably, consistently highlighting the most relevant units. Over-reliance and complacency are an unfortunate negative by-product of highly reliable (yet imperfect) automated systems (Parasuraman & Riley, 1997; Mosier, et al., 1998). As such, the appropriate level of human interaction with such systems must be clarified to ensure safe and efficient use of automation (Bainbridge, 1983).

A significant finding relating to the failure trial is the strong relationship between noticing the uncued high-relevant unit in the failure trial and scoring high on the memory probe for the high-relevant, cued target. This relationship lends further support to the notion that there are different (beta and

$d'$ ) strategies for interacting with the automation. Some observers will utilize the automation to get a global sense of the situation and make their response on the basis of this high-level assessment. This strategy reduces the cognitive demands of the integration task and, given the performance findings, often leads to good allocation decisions. However it is in cases where detailed information needs to be recalled or when automation is unreliable that this advantage breaks down. Alternatively, observers may attend to the local highlighting cues, inspecting the raw data underlying each in turn.

While the strategy just described would directly predict an enhanced ability to notice that a cued item was not of high relevance (ie., an automation cueing "false alarm"), it is important to realize that the automation failures employed here (and better detected by the noticers) was of the opposite type: an automation cueing "miss". Thus the quality of deeper cue processing showed by the noticers must have applied to both cued and uncued items alike, in a way that cannot be easily revealed by the current data. However subsequent analysis revealed that this differential strategy neither slowed nor speeded the overall RT, compared to the non-noticers.

The presence of such different strategies may have important implications in real-world design and applications. The nature and conditions of the task will likely dictate which strategy is more appropriate. For instance, under time pressure adopting a beta strategy (i.e., trust the cues) may be appropriate given that overall allocation performance in the automated conditions was good. When time pressure is not significant, when a task demands recall for specific target details, or when automation is unreliable or imperfect then a  $d'$  strategy may be the best strategy. In order for automated systems to accrue their intended benefits, users must understand how to interact with the system appropriately, an end which may be attained through training or feedback implementation.

*Cue Weighting.* These analyses suggested that observer's judgments were influenced differentially by differences in unit size, distance, terrain, and reliability of information. Both objective and subjective measures indicated that unit size information had a more significant impact on allocation responses than the latter three cues. The military symbol (or numerical digit) for unit size was a highly concrete information cue, which may have contributed to the strong influence on response patterns. The terrain and distance cues, though concrete (physical, geographical) features themselves, were found to be less influential perhaps because the use of these cues required the observer to integrate information about the enemy unit with information regarding the position of one's own unit (hence, increasing mental workload); in the case of terrain, further integration was required with the contour information. Reliability, in contrast, is a more abstract cue than the concrete size, terrain, and distance cues. That is, reliability is a probabilistic information cue, which is often subject to biases in estimation (Tversky & Kahneman, 1981), and not always effectively used in judgments (Wickens Gordon and Liu, 1997). The current findings did not suggest any difference in cue influence between reliability (abstract probabilistic) and terrain and distance (concrete) cues perhaps

due to the graphic display of three different levels of reliability. This graphic display may have reduced the abstractness of the cue, allowing observers to treat it as if it were a concrete cue. Subjective measures, however, suggested that reliability was less influential in observer's allocation responses.

*Conclusions.* While certain benefits and costs of stage 1 automation (Parasuraman, et al., 2000) are expressed in this research, it is less clearly understood how higher stages of automation involving automatic diagnosis will impact performance in the battlefield arena, the impact of repeated failures on trust and system use, or the impact of a highly reliable system (long term) on complacency. The presence of different strategies for interacting with early-stage automation may also have a significant impact on the design and extent of automated systems as well as their task-specific training programs, which may bear a direct influence on the type of strategy a user will employ.

### ACKNOWLEDGMENTS

The authors thank Sharon Yeakel for her time and programming skills, and two anonymous reviewers for their comments and feedback. Research funding for this project was provided through a grant from the U.S. Army Research Laboratory under the Federated Laboratory Program, Cooperative Agreement DAAL0196-2-0003. The authors wish to thank LTC Keith Beurskens of the University of Illinois Army ROTC detachment, for making participants available.

### REFERENCES

Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6), 775-779.  
Barnett, B.J. & Wickens, C.D. (1988). Display proximity in multicue information integration: The benefit of boxes. *Human Factors*, 30(1), 15-24.

Craik, F.I.M. & Lockhart, R.S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671-684.  
Davison, H.J. & Wickens, C.D. (2001). Rotorcraft hazard cueing: The effects on attention and trust. *Proceedings of the 11<sup>th</sup> International Symposium on Aviation Psychology*. Columbus, OH: Ohio State University.  
Endsley, M.R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1), 32-64.  
Graham, S.E. & Matthews, M.D. (1999). *Infantry Situation Awareness: Papers from the 1998 Infantry Situation Awareness Workshop*. Alexandria, VA: US Army Research Institute.  
Metzger, U. & Parasuraman, R. (2001). Conflict detection aids for air traffic controllers in free flight: Effects of reliable and failure modes on performance and eye movements. *Proceedings of the 11<sup>th</sup> International Symposium on Aviation Psychology*. Columbus, OH: Ohio State University.  
Mosier, K.L., Skitka, L.J., Heers, S., & Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits. *International Journal of Aviation Psychology*, 8(1), 47-63.  
Parasuraman, R. & Riley, V. (1997). Humans and automation: Use, misuse, disuse, and abuse. *Human Factors*, 39(2), 230-253.  
Parasuraman, R., Sheridan, T.B., & Wickens, C.D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 30(3), 286-297.  
Tversky & Kahneman (1981). The framing of decisions and the psychology of choice. *Science*, 185, 1124-1131.  
Wickens, C.D., Gordon, S.E., & Lui, Y. (1997). *An Introduction to Human Factors Engineering*. New York: Addison Wesley Longman.  
Yeh, M. & Wickens, C.D. (2001, in press) Explicit and implicit display signaling in augmented reality: The effects of cue reliability, image realism, and interactivity on attention allocation and trust calibration. *Human Factors*, 43(3).  
Yeh, M., Wickens, C.D., & Seagull, F.J. (1999). Target cueing in visual search: The effects of conformality and display location on the allocation of visual attention. *Human Factors*, 41(4), 524-542.