

EFFECTS OF AUTOMATION FAILURE IN A LUGGAGE SCREENING TASK: A COMPARISON BETWEEN DIRECT AND INDIRECT CUEING

Juliana Goh, Douglas A. Wiegmann, Poornima Madhavan
Aviation Human Factors Division
University of Illinois, Urbana-Champaign

The present study investigated the use of two automated aids of different reliabilities in a luggage screening task. A Direct Cue consisting of a green circle around a potential target directs attention to a specific part of the luggage image, while an Indirect Cue, consisting of a green border around an image determined to have a target, does not. Direct Cues offer an advantage in visual inspection tasks because they guide attention to specific areas of the visual image but this can also cause attentional tunneling. Furthermore, the reliance on automation may negatively impact manual performance after the aid is removed or is no longer available. Thus, two issues were investigated in the current study: (1) how do failures in Direct and Indirect Cues affect reliance and (2) how does a complete failure affect performance after operators had the use of an automated aid? Results suggest that reliance patterns were more optimal with the Direct Cue than with the Indirect Cue and performance with a more reliable Indirect Cue was not much better than a less reliable one. The results also suggest that manual performance, when the aid was removed, was better for participants who had used the automated aids compared to control participants who did not have any use of the aid previously. The advantage of previously aided performance on subsequent manual performance was greatest for those who had used the more reliable Direct Cue. Explanations and implications are discussed.

INTRODUCTION

Visual inspection tasks require individuals to search displays that contain objects of interest embedded among distractors. Examples include airport security personnel scanning X-ray images of luggage looking for banned objects, and military personnel who have to detect the presence of camouflaged weapons or enemies. More often than not, the targets of interest are not immediately obvious or occur at very low signal rates, making these tasks inherently difficult (Wickens & Hollands, 2000), and since the consequences of missing targets can often be detrimental to both personal and public safety, there is considerable impetus to develop methods to improve visual search performance in a variety of occupational settings.

One approach for improving visual search performance is to circumvent the limited information processing capabilities of human operators through the use of automation (Mosier & Skitka, 1996) which is becoming commonplace as such technologies become more readily available. However, effective integration of automation requires several issues to first be addressed, such as which aspect(s) of the visual inspection task should be automated and what are the consequences if the automation fails? Such evaluations are critical because automation is rarely perfect. Automation is likely to fail occasionally, causing changes in the way the user interacts with it, or it may fail completely, forcing the user to perform the task manually.

Automation and Luggage Screening

Since the difficulty of the luggage screening task is in detecting and recognizing banned objects, it is during the first two stages of information processing (information acquisition and diagnosis) that observers will need help. The focus of the

current study is, therefore, on the comparison between the use of an attention guiding Direct Cue – a green circle around the target and a diagnostic Indirect Cue – a green border around the luggage image and examining how (1) occasional failures affect reliance on these cues and (2) how use of these cues affect manual performance when the use of these cues is no longer available.

Occasional Failures and Reliance on Direct/Indirect Cues

Research on the use of cueing in visual inspection tasks suggests that cues that direct attention (Direct Cue) provide the advantage of reducing search space and the impact of distractors but also causes attentional tunneling (Krupinski, Nodine and Kundel, 1993; Yeh, Wickens and Seagull, 2000; Maltz and Shinar, 2003). An Indirect Cue, on the other hand, does not cause attentional tunneling but does not provide the benefits that a Direct Cue would. One study that has made a comparison between an attention-directing cue and one that does not was carried out by Wiegmann, McCarley, Kramer and Wickens (in press). They compared the use of a spatial cue which circled potential targets and a text cue which told participants what to do (“Pass/Stop Bag”) in a luggage screening task. The authors found that participants who used the spatial cue performed better and relied more on the aid than participants who used the text cue even though both cues were similarly reliable.

Cues that direct attention also provide more immediate feedback than cues that do not. When using a Direct Cue, for example, attention is drawn to a specific portion of the image to a potential target. The user is able to inspect that area immediately to verify whether a target does indeed exist. When using an Indirect Cue, on the other hand, the user is required to inspect the entire image before a determination can be made. To this end, it may be argued that

users of the Direct Cue would be better able to (1) increase their hit rates without increasing false alarm rates and (2) determine the reliability of the aid more appropriately than would users of the Indirect Cue. A more appropriate level of reliance and better performance may therefore be seen for users of the Direct Cue compared to users of the Indirect Cue. This hypothesis may be tested by varying the reliability levels of the two types of cues to examine whether reliance as measured by participant performance increases as reliability also increases and whether this differs depending on the type of cue that is used.

As mentioned, however, the cost of attentional guidance is that the user's attention tends to be drawn to the cued location so that objects located elsewhere in the visual field may go unnoticed and an unreliable cue will therefore lead to increased misses and false alarms. In addition, because attention is drawn to the image, miscues (a non-target is cued when the true target is located elsewhere in the image) could potentially be more salient and this may have implications for user reliance. Given these costs and benefits, users may have different reliance and performance patterns depending on the type of cue that is used.

Complete Failure and Manual Performance

The literature suggests that automation reliance in decision-making tasks tends to impair knowledge acquisition by novices (e.g. Glover, Prawitt and Spilker, 1997). One may question, as well, whether the use of automated cues in visual inspection tasks will impair users' abilities to perform the task manually after having relied on these cues. Based on the visual search learning literature, it appears that certain automated aids could potentially encourage user's abilities to detect and recognize targets. For example, Yund and Effron (1996) have shown that repeated exposures to specific targets lead to better performance over time and McCarley, Kramer, Wickens, Vidoni and Boot (2004) have shown this in a luggage screening task. One might argue then, that the use of Direct Cues, which direct attention to targets, will also encourage the encoding of target representations, but this may not necessarily be seen when Indirect Cues are used.

However, it is important to consider the reliability of the automated aid as well. Using a highly reliable Direct Cue would encourage more learning than using one that is less so, since the more reliable cue will bring attention more often to the target(s). Using a highly reliable Indirect Cue, however, could encourage reliance resulting in users spending less time inspecting the luggage image, and therefore having fewer opportunities to detect, inspect and recognize the targets. A less reliable cue, under these circumstances, may be more beneficial to visual skill acquisition.

Purpose of the Present Study and Experimental Design

The purpose of the present study, was the investigate how direct and indirect cues of varying reliability impact reliance and skill acquisition. Three variables were

manipulated completely between subjects (1) type of cue (Direct vs. Indirect), (2) reliability (70% vs. 90%) and (3) type of miss for direct cues (regular miss vs. miscue). To examine the effect on skill acquisition, participants performed two sessions of the same task, first with the aid then without. A control group of participants performed both sessions without automation.

METHODS

Participants

One hundred and forty students (20 per group) from the University of Illinois participated in the experiment. Approximately half received course credit for their participation while the other half were paid \$18 for their participation.

Stimuli

Four hundred unique luggage images (200 per session) were used. 80 of the contained knives (20% signal rate). 2 sets of 4 knives each were selected based on a pilot study comparing the similarity of an initial set of 20 knives. The 4 knives *within* each set were different from one another, but each had a corresponding knife with a similar rating in the other set. The difficulty of finding the knives and degree of clutter of the luggage images was controlled across sessions.

Procedure

Both sessions of the experiment were conducted within a 2.5 hour period. Participants were shown images of the knives while they were read the instructions. They were told to scan each image and to indicate whether a knife was present or absent in the images by pressing buttons on a game controller. They were informed that the knife could be placed horizontally, vertically or angled at 45° or 135°, to balance accuracy and response time when performing the task and to make a response within 20 seconds, after which the trial would time out. This continued until all 200 trials were completed. Participants who performed the task with the help of the automated aid were not told the actual reliability of the aid and were instructed that it was their decision whether or not to accept the aid's assistance or to decide on their own. Participants took a 15 minute break between the sessions after which they were shown images of the second set of knives they had to detect and reminded of how to perform the task. The second session lasted another 200 trials.

RESULTS

Session 1: Aided Performance

Figure 1 below plots d' as a function of reliability and automation condition in session 1. d' of the 70% and 90% aids are also plotted for reference. Comparing participants' performance to the aid's performance, it is clear from Figure 1 that control participants performed close to the 70% reliable

aid ($d'=1.04$). For participants who had used the Direct Cue, those in the 70% reliable condition performed better than the aid while those in the 90% reliable condition performed close to the 90% reliable aid ($d'=2.56$). For participants who used the Indirect Cue, those in the 70% reliable condition performed just as well as the aid but those in the 90% reliable condition did not appear to rely sufficiently on the aid as the d' was much lower than the aid's.

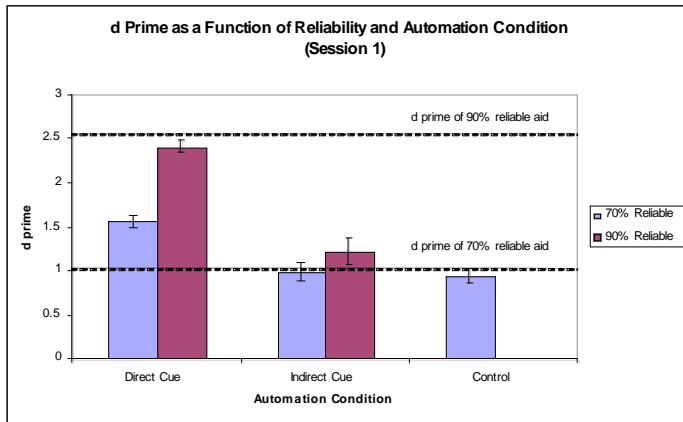


Figure 1. d' as a function of reliability and automation condition (session 1).

Figure 1 also shows d' to be highest with the 90% reliable aids, followed by the 70% reliable aids and then the Control condition and a one-way ANOVA revealed a significant effect of Reliability, $F(2, 130) = 27.423$, $p < .001$. Post-hoc comparisons revealed that all pairwise differences between the Reliability groups were significant, $p < .05$, confirming the hypothesis that sensitivity would be affected by Reliability, with the more reliable aids having higher sensitivity than less reliable aids and manual performance. It can also be seen that d' was highest for participants who used the Direct Cue followed by those who used the Indirect Cue and then Control participants. The trends are as hypothesized and this is confirmed in a one-way ANOVA which revealed a significant effect of Automation Condition, $F(2, 130) = 27.423$, $p < .001$. Post-hoc tests revealed significant differences between the Direct and Indirect Cue conditions and between the Direct and Control conditions, $p < .01$. The difference between the Indirect and Control conditions, however, did not reach statistical significance, $p > .10$.

Inspection of data for the automated groups alone shows that the 90% reliable aid provided more of a performance benefit than the 70% reliable aid in the Direct Cue condition compared to the Indirect Cue condition. Indeed, a 2 (Type of Cue) \times 2 (Reliability) ANOVA revealed significant main effects of Type of Cue, $F(1, 112) = 86.683$, $p < .001$, Reliability, $F(1, 112) = 32.645$, $p < .001$, and a significant Type of Cue \times Reliability interaction, $F(1, 112) = 10.883$, $p < .05$. t-tests indicate that d' is significantly higher with the 90% reliable cue than with the 70% reliable cue in the Direct Cue condition, $t(74) = 9.057$, $p < .001$, but the same comparison was not statistically significant for the Indirect

Cue, $p > .10$, confirming the hypothesis that performance difference would be larger between the 90% and 70% reliable conditions for the Direct Cue than with the Indirect Cue.

To gain insight into whether the differences in d' were the result of higher hits or lower false alarms or both, Figure 2 plots the hit and false alarm rates as a function of reliability and automation condition. The figure indicates that hit rate was highest for participants who had used the 90% reliable aids followed by the 70% reliable aids and then the control condition. Those who used the 90% reliable aids also made the fewest false alarms followed by those who used the 70% reliable aids and Control condition. One-way ANOVAs revealed significant differences in Reliability for hits, $F(2, 130) = 22.993$, $p < .001$, and false alarms, $F(2, 130) = 4.470$, $p < .05$ and post-hoc tests revealed significant differences for all pairwise comparisons for hits, $p > .001$, while false alarm rates differed significantly between the 90% and 70% reliable conditions, $p < .05$ and marginally significantly between the 90% and control conditions, $p = .065$. The 70% reliable and control conditions did not differ significantly, $p > .10$.

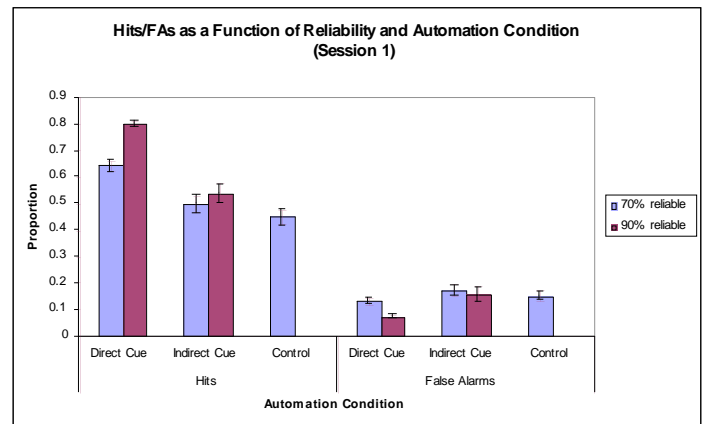


Figure 2. Hit and false alarm rates as a function of reliability and automation condition (session 1).

Comparing across automation conditions, Figure 2 shows hit rates to be highest with the Direct Cue, followed by the Indirect Cue and Control condition. In addition, false alarm rates are lowest with the Direct Cue while those for the Indirect Cue and control condition appear to be similar. One-way ANOVAs revealed significant differences in Automation Condition for hits, $F(2, 130) = 46.93$, $p < .001$, and false alarms, $F(2, 130) = 7.511$, $p < .01$. Pairwise comparisons showed that hit rate was significantly higher with the Direct Cue than the Indirect Cue ($p < .001$), which did not differ significantly from the control condition, $p > .10$. Hit rate was also higher in the Direct Cue condition compared to controls, $p < .001$. Comparing false alarm rates, participants who used the Direct Cue had lower false alarm rates compared to those who used the Indirect Cue ($p < .05$) and Control participants, $p = .056$. False alarm rates did not differ significantly between participants in the Indirect Cue and Control conditions, $p > .10$.

Comparing data from the automated groups, a 2 (Type of Cue) \times 2 (Reliability) on hit rate revealed a

significant effect of Reliability, $F(1,112)=16.496$, $p<.001$, Type of Cue, $F(1,112)=71.113$, $p<.001$, and a significant Type of Cue \times Reliability interaction, $F(1,112)=6.265$, $p<.05$. t-tests showed a significantly higher hit rate for the 90% reliable cue compared to the 70% reliable cue in the Direct Cue condition, $t(74)=6.719$, $p<.001$, but the same comparison for the Indirect Cue condition did not reveal significant differences, $p>.10$. A 2 (Type of Cue) \times 2 (Reliability) ANOVA on false alarm rate also revealed a significant effect of Reliability, $F(1, 112)=4.974$, $p<.05$, Type of Cue, $F(1, 112)=13.972$, $p<.001$ and a non-significant interaction, $p>.10$. Planned t-tests showed that participants in the Direct Cue condition who used the 90% reliable cue committed fewer false alarms than those who used the 70% reliable cue, $t(74)=4.234$, $p<.001$, but this was not true for participants in the Indirect Cue condition, $p>.10$.

To investigate the impact of automation misses and miscues we looked at the conditional probability of the human missing when the automation suggests a miss is compared across the conditions. This allows an assessment of the likelihood of participants being misled by the cue when it missed a target. Figure 3 plots the probability of miss and the conditional probability of missing given that the automation misses ($P(\text{Human}_m|\text{Automation}_m)$) across the various conditions.

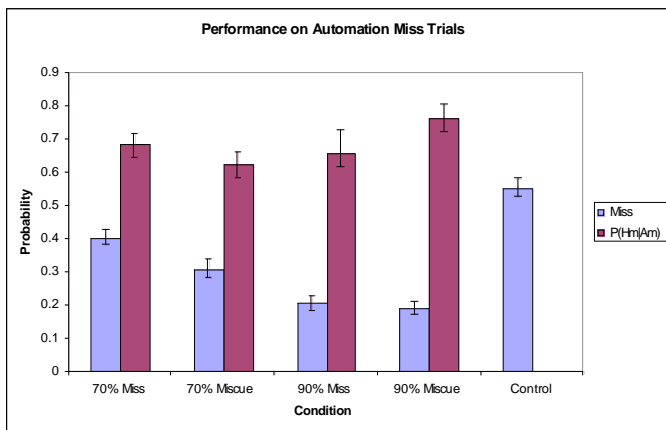


Figure 3. Performance on automation miss trials.

From the figure, it can be seen that miss rates for the Control condition is higher than the miss rates for the miss and miscue conditions. The miss rate for these two conditions are similar. A one-way ANOVA on miss rates across miss, miscue and control conditions which shows a significant effect of condition, $F(2,94) = 36.942$, $p<.001$. Post-hoc tests show no significant difference in miss rate between the miss and miscue conditions but each of these conditions have significantly lower miss rates than the control condition, $p<.001$. This result indicates that miscues were not especially detrimental to performance compared to regular misses. Inspection of Figure 4 also shows that participants are *less* likely to miss when the automation is 70% reliable and tends to miscue while those using the 90% reliable aid are *more* likely to miss when it is an aid that tends to miscue. These

differences, however, did not reach statistical significance, $p>.10$.

Session 2: Manual Performance

Figure 4 plots d' as a function of reliability and automation condition in session 2. As can be seen, d' appears to be relatively similar regardless of Reliability and this is confirmed with a one-way ANOVA which revealed no significant difference in d' across Reliability, $p>.10$. d' is also similar across Automation condition and a one-way ANOVA revealed no significant effect of Automation condition, $p>.10$.

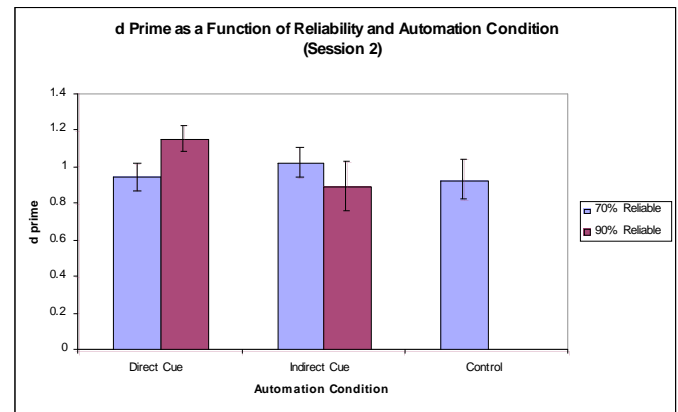


Figure 4. d' as a function of reliability and automation condition (session 2)

Within the automated groups, inspection of Figure 4 suggests that, as hypothesized, d' is higher with the 90% reliable aid compared to the 70% reliable aid in the Direct Cue condition, but the reverse is true for the Indirect Cue, d' is higher with the 70% reliable aid than with the 90% reliable aid. A 2 (Type of Cue) \times 2 (Reliability) ANOVA revealed no significant main effects of Reliability or Automation Condition, $p>.10$ but a marginally significant interaction, $F(1, 112)=3.36$, $p=.07$. t-tests show that participants in the Direct Cue condition who used the 90% reliable cue performed better than those who had used the 70% reliable cue, $t(74)=2.048$, $p<.05$. This difference was not statistically significant for participants who had used the Indirect Cue, $p>.10$.

To examine whether differences in d' observed above were a result of differences in hit and/or false alarm rates, Figure 5 plots the hit and false alarm rates across automation and reliability conditions. Hit rates for the 70% reliable condition is similar to that of the 90% reliable condition both of which are slightly higher than the Control condition. False alarm rates appear similar across both reliability conditions and the Control condition. One-way ANOVAs revealed no significant differences in hits or false alarm rates across Reliability, $p>.10$.

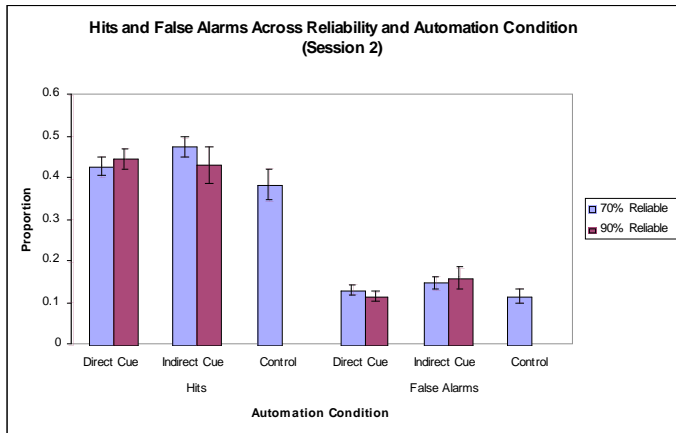


Figure 5. Hit and false alarm rates as a function of reliability and automation condition (session 2)

Comparing hits and false alarm rates across Automation Condition, Figure 5 indicates that hit rates are similar between the Direct and Indirect Cue conditions, each of which appear to have higher hit rates compared to the Control condition. The figure also shows false alarm rates to be similar across Automation Condition. Indeed, one-way ANOVAs revealed no significant differences in hit and false alarm rates across Automation Condition, $p > .10$. Planned one-tailed t-tests show that hit rates were not significantly different between the Direct and Indirect Cue conditions, $p > .10$, but each of these groups had marginally significantly higher hit rates than the control condition, [Direct Cue vs. control: $t(93) = 1.462$, $p = .074$; Indirect Cue vs. control: $t(53) = 1.545$, $p = .064$]. A 2 (Type of Cue) \times 2 (Reliability) ANOVA on hit rate and false alarm rate showed all effects to be non-significant, $p > .10$. Post-hoc t-tests reveal non-significant differences in hit and false alarm rates between the 90% reliable and 70% reliable cues for both the Direct and Indirect Cue conditions, $p > .10$.

DISCUSSION

The purpose of the study was to examine how different types of automation of varying reliabilities affected performance and visual skill acquisition in a luggage inspection task. Specifically, the study was concerned with how a Direct Cue and Indirect Cue affected the extent to which users relied on the automation and how the use of these cues affected users' abilities to perform the task manually after having performed the task with the help of the automated aids.

The results show that using a direct cue leads to better performance than when an indirect cue is used and this advantage is greater in more reliable systems. The advantage of getting immediate feedback from direct cues appears to be the reason for this; direct cueing led to higher hit rates without increasing false alarm rates. In addition, the results suggest that direct cueing is beneficial to visual skill acquisition but this may be true only for a highly reliable direct cueing system. Indeed, the results suggest that learning from a less reliable direct cue is worse than the learning that could be

achieved with an indirect cue of similar reliability. The advantage of using the highly reliable direct cue for visual skill acquisition appears to be the result of a combination of a slightly higher hit rate and lower false alarm rate compared to control participants and one might argue that participants who had used this cue could possibly have developed stronger target representations than those who did not have the benefit of attentional guidance previously.

The practical implications of this study are as follows. Firstly, implementing a direct cueing system that is not perfectly reliable may not be too detrimental to performance even when it fails. Secondly, a simulated automated aid may be used to train security screeners to improve their ability to pick out banned objects. Clearly, a direct cueing system would be more beneficial as a training tool than an indirect cueing system. An unresolved issue, however, is what objects security screeners should be trained with. The present study limited itself to knives as targets but the list of banned objects in the real world is more extensive, thus acquiring the ability to detect one group of objects does not necessarily translate into ability to pick out other objects. Further research in this area is necessary.

REFERENCES

- Dzindolet, M.T., Pierce, L.G., Beck, H.P., Dawe, L.A., & Anderson, B.W. (2001). Predicting misuse and disuse of combat identification systems. *Military Psychology*, 13(3), 147-164.
- Glover, S.M., Prawiit, D.F., & Spilker, B.C. (1997). The influence of decision aids on user behavior: Implications for knowledge acquisition and inappropriate reliance. *Organizational Behavior and Human Decision Processes*, 72(2), 232-255.
- Krupinski, E.A., Nodine, C.F., & Kundel, H.L. (1993). Perceptual enhancement of tumor targets in chest X-ray images. *Perception and Psychophysics*, 53(5), 519-526.
- McCarley, J.S., Kramer, A.F., Wickens, C.D., Vidoni, E.D., & Boot, W.R. (2004). Visual skills in airport security screening. *Psychological Science*, 15(5), 302-306.
- Wickens, C.D., & Hollands, J. (2000). *Engineering Psychology and human performance* (3rd ed.). Upper Saddle River, N.J.: Prentice Hall.
- Wiegmann, D.A., McCarley, J.S., Kramer, A.F., Wickens, C.D. (under review). Effects of age on utilization and perceived reliability of an automated decision-making aid in a luggage screening task. *Human Factors*.
- Yeh, M., Wickens, C.D., & Seagull, F.J. (2000). Target cueing in visual search: The effects of conformality and display location on the allocation of visual attention. *Human Factors*, 43(3), 355-365.