Beyond Identity : Incorporating System Reliability Information Into an Automated Combat Identification System

Heather F. Neyedli, Justin G. Hollands and Greg A. Jamieson

Human Factors: The Journal of the Human Factors and Ergonomics Society 2011 53: 338
DOI: 10.1177/0018720811413767

Beyond Identity: Incorporating System Reliability Information Into an Automated Combat Identification System

Heather F. Neyedli, University of Toronto, **Justin G. Hollands**, Defence Research and Development Canada, and **Greg A. Jamieson,** University of Toronto

Objective: The aim of this study was to evaluate display formats for an automated combat identification (CID) aid.

Background: Verbally informing users of automation reliability improves reliance on automated CID systems. A display can provide reliability information in real time.

Method: We developed and tested four visual displays that showed both target identity and system reliability information. Display type (pie, random mesh) and display proximity (integrated, separated) of identity and reliability information were manipulated. In Experiment 1, participants used the displays while engaging targets in a simulated combat environment. In Experiment 2, participants briefly viewed still scenes from the simulation.

Results: Participants relied on the automation more appropriately with the integrated display than with the separated display. Participants using the random mesh display showed greater sensitivity than those using a pie chart. However, in Experiment 2, the sensitivity effects were limited to lower reliability levels.

Conclusion: The integrated display format and the random mesh display were the most effective displays tested.

Application: We recommend the use of the integrated format and a random mesh display to indicate identity and reliability information with an automated CID system.

Keywords: automation, combat identification, decision aids, integrated displays, proximity compatibility principle, reliance, simulation, signal detection theory, system reliability, visual attention, visual displays

HUMAN FACTORS
Vol. 53, No. 4, August 2011, pp. 338–355
DOI:10.1177/0018720811413767

INTRODUCTION

A soldier must discriminate between friendly and hostile targets on the battlefield, a process known as combat identification (CID). Poor CID may result in fratricide, neutricide, and reduced mission effectiveness (Young, 2005). To perform CID, soldiers rely on tactics and procedures learned in training and information provided in premission briefings. Recently, automated CID technology has been developed to assist soldiers in the field, including rifle-mounted identification friend-or-foe (IFF) aids for dismounted infantry.

The IFF system consists of a rifle-mounted interrogator that sends out a coded laser inquiry when manually activated (K. Sherman, 2000; K. B. Sherman, 2002). Transponders are mounted on the helmet or uniform of friendly soldiers. If a transponder intercepts the laser signal, it emits an omnidirectional radio frequency response and "friend" feedback is provided to the interrogating soldier. If the transponder does not intercept the laser signal, no feedback is provided. In this case, the target could be an enemy but could also be a neutral party, a member of a friendly force without the technology, or a friendly soldier with an inoperable transponder. Therefore, the system provides highly reliable "friend" feedback but provides ambiguous information when the interrogator receives no transponder feedback. The ambiguity of this feedback (which we call "unknown" feedback) could cause the soldier to rely on the automated system inappropriately (Karsh, Walrath, Swoboda, & Pillalamarri, 1995). The challenge then becomes finding methods that allow the soldier to interpret the "unknown" feedback appropriately.

Automation and Combat Identification

Automation can provide many benefits; however, the costs of introducing automation have been well documented across a variety of domains (e.g., Bainbridge, 1983; Lee & See, 2004; Parasuraman & Riley, 1997; Skitka, Mosier, & Burdick, 1999). Human-automation performance improves as automation becomes more reliable (e.g., Maltz & Shinar, 2003). Rovira, McGarry, and Parasuraman (2007) showed that 80% reliable information automation improved performance relative to a manual condition, whereas less reliable automation (60%) did not. Wickens and Dixon (2007) performed a meta-analysis of studies examining automation reliability and human performance and found that human performance improved monotonically with system reliability and also that automation with a reliability level less than about 70% was no better than manual performance.

However, the CID automation context differs from these automation studies. As Wang, Jamieson, and Hollands (2009) noted, even with the IFF aid, the user is still actively involved in the identification task. In contrast, for most other forms of automation, the human provides a governance or oversight role, acting only when alerted by the automation (e.g., Skitka et al., 1999). Furthermore, as noted previously, CID automation reliability varies with the type of feedback the user receives, because although the automation provides highly reliable "friend" feedback, it cannot provide reliable identity information when no radio frequency feedback is received (Wang et al., 2009).

Studies that have examined imperfect automation in CID have found that availability of a decision aid either had no effect on CID performance (e.g., Dzindolet, Pierce, Beck, Dawe, & Anderson, 2001; Karsh et al., 1995) or improved performance only in the most difficult conditions (Kogler, 2003). Dzindolet et al. (2001) also reported that participants overrelied on imperfect CID automation. However, these studies did not define an appropriate level of automation reliance.

Signal detection theory (SDT; Macmillan & Creelman, 1991) has been applied in a number of domains in which an automated system and a human observer work together to detect an event of interest (e.g., Sorkin & Woods, 1985). In SDT, two parameters characterize performance. The first parameter, *sensitivity*, describes how well the observer can distinguish the event, or signal, from background noise. The second parameter, *response bias*, describes how likely the observer is to respond in a particular way (saying yes more than no, for example). Sorkin and Woods (1985) applied SDT to the use of alerts for system monitoring applications (e.g., process control), and signal detection approaches have been applied to the use of automated systems for other applications, such as radiation screening at seaports (Sanquist, Minsk, & Parasuraman, 2008). Using this approach, Sanquist et al. (2008) defined the probability of threat given an alarm, characterized the effect of automated system false alarms (FAs) on human sensitivity, and made recommendations for maximizing the effectiveness of the total system.

In the SDT approaches described by Sorkin and Woods (1985) and Sanquist et al. (2008), there is an assumed sequential process in which the automated system screens for an event and the human monitors the output. In contrast, as noted earlier, in CID, the human and automation monitor the environment in parallel. Wang, Jamieson, and Hollands (2008) reviewed measures of reliance in CID and used SDT's response bias parameter to characterize reliance on the decision aid. With their approach, an optimal level of reliance can be defined.

In the Wang et al. (2009) experiment, the aid had perfectly reliable "friend" feedback (i.e., all targets identified as "friend" were friendly soldiers), but the reliability level for "unknown" feedback (defined as the probability of a hostile target given "unknown" feedback) was varied. Wang et al. also examined whether disclosing the reliability level would help users rely on the automation more appropriately. Participants performed CID manually or with an automated aid that was 67% or 80% reliable. An optimal response bias value was defined for each reliability level. Wang et al. found that providing participants with a highly reliable aid (80% reliable) reduced identification error as compared with performance with no aid. In addition, participants who were informed of the

aid's reliability adjusted their reliance more appropriately (closer to the optimal value) than uninformed participants who were shown identity information only. However, even informed participants did not adjust their reliance as much as was warranted; in the 80% reliability condition, they underrelied on the aid.

Display Design: Reliability and Uncertainty

In the Wang et al. (2009) study, the reliability level was fixed for a block of trials and was verbally disclosed to participants in percentage form before each block. Thus, participants had to maintain a fixed reliability level value in working memory for each block. However, in operational CID situations, system reliability will likely vary over time for many reasons (e.g., distances of sensors from IFF transmitter, effects of humidity on sensors, number of friendly forces or noncombatants in the area). Presenting real-time reliability information on a visual display might help users rely on the information more appropriately. We developed four prototype displays for this purpose, shown in Figure 1.

The prospect of presenting reliability information along with identity information raises the issue of display proximity: whether it is better to integrate (*integrated display*; Figure 1, Panels A and B) or separate (*separated display*; Figure 1, Panels C and D) the two types of information. The proximity compatibility principle (PCP; Wickens & Carswell, 1995) posits that placing graphical elements close together will assist the user when the task requires information from these elements to be considered together.

Wickens and Carswell (1995) specified many possible proximity manipulations. One of these is a *configural display*, in which the values of system variables are mapped onto the geometric properties of a polygon (e.g., Coury, Boulette, & Smith, 1989). Wickens and Carswell describe another manipulation as *spatial integration*, in which the different information sources are assigned to different perceptual dimensions of the same object (e.g., area and color). In a time-pressured task, the elements should be within the user's foveal vision (about 2° of visual angle), thereby reducing the need

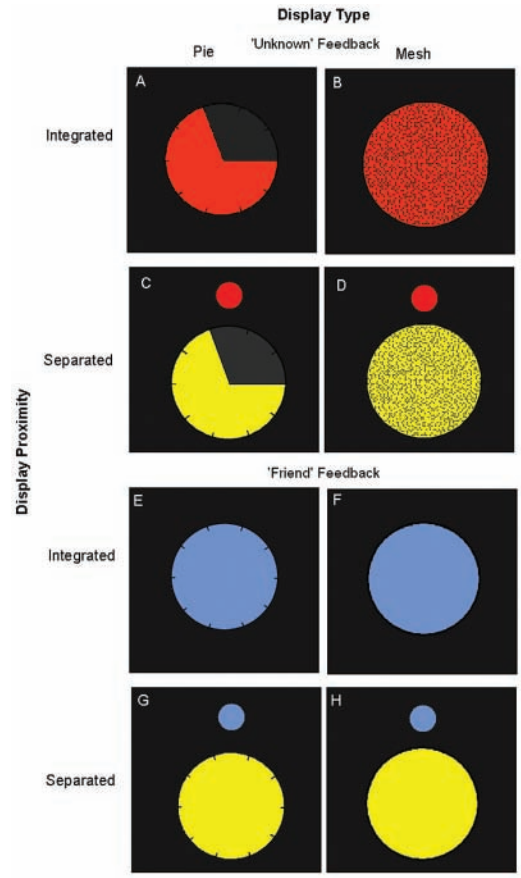

Figure 1. Examples of each display type and proximity combination. Panels A through D show "unknown" feedback in the 70% reliability condition. Panels E and F show "friend" feedback, which was always 100% reliable. See text for details.

for saccades (Brand & Orenstein, 1998) or covert shifts of visual attention. Wickens and Carswell refer to the need for movements of attention or the eye as the *information access cost* of a display.

Reliability information can be represented digitally or in an analog graphical form. Probability or reliability values are typically expressed as proportions, and pie charts have been shown to be a highly effective graphical method for depicting proportions (e.g., Hollands & Spence, 1992, 1998; Spence & Lewandowsky, 1991). Pie charts are generally read quite accurately (±2% to 3%; Hollands & Spence, 1998). Yet in a medical decision-making study, participants asked to decide which of two

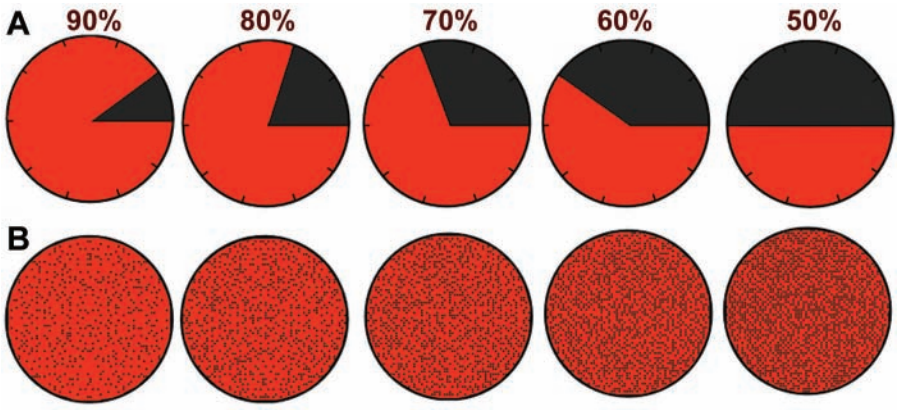

Figure 2. Integrated pie displays (A) and mesh displays (B) at each reliability level.

proportions was larger made faster and more accurate decisions with a random arrangements of black and white ovals than with pie charts (Feldman-Stewart, Brundage, & Zotov, 2007). In another study, Finger and Bisantz (2002) developed a face icon whose image resolution changed with reliability level by varying pixel size. Their participants identified targets more quickly with the degraded icons than with the seemingly more precise digital-numeric methods, and performance was no worse when using the degraded icons.

Furthermore, Chong and Treisman (2003, 2004; see also Ariely, 2001) showed that people can distribute attention across a visual array (e.g., circles of various sizes randomly arranged) and estimate statistical properties of that array (e.g., mean size) quite quickly, regardless of the number of elements within the array. Taken in combination, these results suggest that the statistical properties of a visual display can be assessed quickly in parallel and may offer an effective method for depicting system reliability information in a time-limited situation, such as CID.

For our experiments, we used a pie chart format (*pie display*; e.g., Figure 1, Panel A), in which the size of the colored slice was proportional to the reliability level. The 0% reliability level was set to the three o'clock position with reliability level increasing clockwise (Figure 2A). We also developed prototypes that used a random arrangement of tiny squares in a grid, in which the number of filled squares was proportional to the reliability (*mesh display*; e.g., Figure 1, Panel B). Since the filled squares were located randomly in the grid, the display appeared to degrade as reliability decreased (Figure 2B).

The display proximity and display color type manipulations combined to form four prototypes. We consider the displays depicting "unknown" feedback first (Figure 1, Panels A through D). For the integrated displays (Figure 1, Panels A and B), a single circle was shown. The proportion of the form that was colored in with red (as opposed to black) indicated the feedback reliability level. The integrated display was therefore constructed with the use of spatial integration (Wickens & Carswell, 1995), such that the different information sources (target identity and system reliability) were assigned to different perceptual dimensions of the object (color and filled area). For the separated displays (Figure 1, Panels C and D), the upper dot was red (indicating "unknown" feedback), and the proportion of the larger circle below that was colored in with yellow (as opposed to black) indicated the feedback reliability level. Identity information and reliability information are represented by separate objects (the small and large circle, respectively).

The bottom part of Figure 1 (Panels E through H) shows the displays used to depict "friend" feedback. For the integrated displays (Figure 1, Panels E and F), a single blue circle was shown for both pie and mesh displays (tick marks were shown on the pie chart version). For the pie display (Figure 1, Panel E), the fact that the entire circle was blue represented 100% reliable

feedback; for the mesh display, 100% reliability was indicated by the fact that the circle was solid blue. For the separated displays (Figure 1, Panels G and H), the upper dot was blue (indicating "friend" feedback), and the larger circle below was solid yellow. Thus, for the pie (Figure 1, Panel G), the fact that the entire circle was yellow meant that the "friend" feedback was 100% reliable. For the mesh, (Figure 1, Panel H), the fact that the circle was solid yellow indicated that the "friend" feedback was 100% reliable.

Consider the representation of uncertainty for a CID decision aid. By decreasing the reliability level, the user becomes less certain about the state of the world, and this uncertainty means that it is less clear how to translate the information presented in the display into an action (shoot or hold fire). In a sense, then, the uncertainty is represented by the area in the display complementary to that area representing reliability level.

This method stands in contrast to methods for displaying uncertainty in process control, such as those used by Coury et al. (1989), in which system states are defined by the combination of system variables. Coury et al. increased uncertainty in their simulated experimental system by creating an overlap between state categories, so that system variables took on values that could occur in more than one system state. They found that when uncertainty was high, configural displays were less effective than the more separable formats—bar graphs and digital-numeric indicators—used in the study. In this case, the displays represented multiple variables precisely, but the classification of variable combinations into categories was uncertain. For our application, uncertainty is inversely proportional to a single variable shown in the display. We view these approaches to the representation of uncertainty as complementary.

Our study used several automation reliability levels. Following Wang et al. (2009), reliability level was defined as P(Hostile | "Unknown"). We were interested in seeing how reliance shifted with reliability level, which varied from chance level (50%) to highly reliable (90%). To test the relevant regression model, our experiments used five equally spaced reliability levels, varying in 10% increments. This range of levels crossed 70%, the approximate reliability level determining whether automation improves performance or not (Wickens & Dixon, 2007). We were also interested in the shape of the function between reliability level and reliance. In particular, we wanted to know whether the slope of observed reliance scores corresponded to optimal criterion settings across the range of reliability levels.

How should one assess the quality of human performance with a decision aid showing both identity and system reliability information? First, a more effective decision aid should improve the ability to distinguish a hostile from a friendly target—that is, improve sensitivity in the SDT sense. Second, if the decision aid discloses system reliability information to the human user more effectively, it should improve human reliance on the aid. According to Wang et al. (2009), human reliance on an automated CID aid is equivalent to the placement of the SDT decision criterion given particular feedback from the decision aid (e.g., given "unknown" feedback). The metric for the placement of this criterion in SDT is β. As noted earlier, Wang et al. defined optimal β for decision aids in the CID context. This makes it possible to compare empirical β values with optimal β values. When the reliability level of a decision aid increases, we would expect greater reliance on that aid, but we would expect more optimal shifts of reliance with those display formats that can communicate the reliability level quickly and effectively.

EXPERIMENT 1

Disclosing a decision aid's reliability level to users has been shown to improve reliance on the aid (Wang et al., 2009). Thus, the purpose of Experiment 1 was to (a) determine whether integrating target identity and system reliability information in an automated CID aid improves human performance with that aid and (b) determine which method of displaying system reliability information (pie or mesh) best improves human performance with the aid. Improved human performance is defined as improved sensitivity for detecting targets, more appropriate (closer to optimal) reliance on the aid, or both.

We provided the participants with a prototype CID aid with a in a simulation based on a first-person shooter game. The aid had perfectly reliable "friend" feedback (i.e., all targets identified

TABLE 1: Number and Identity of Targets in Each Condition for Each Participant

	Integrated (n = 420)		Separated (n = 420)	
Reliability Level	Friendly	Hostile	Friendly	Hostile
"Friend" feedback				
100	120	0	120	0
"Unknown" feedback				
50	30	30	30	30
60	24	36	24	36
70	18	42	18	42
80	12	48	12	48
90	6	54	6	54
100				
Total	210	210	210	210

as "friend" were friendly soldiers), but the probability of a hostile target given "unknown" feedback varied. We used five reliability levels, varying from 50% to 90% in 10% increments. The reliability level shown on the display accurately indicated the likelihood that the target was hostile given "unknown" feedback. Table 1 lists the number and identity of targets in each feedback condition. To ensure the participants read the reliability value on the display on each trial, the reliability level varied trial to trial.

We predicted better performance with the integrated than with the separated display formats, because the CID process must use both identity and reliability information in close temporal proximity to be performed well. The PCP predicts that for such tasks, integrated display formats should be superior, leading to improved sensitivity and more optimal placement of the decision criterion because of the reduced information access cost (Wickens & Carswell, 1995). Second, we predicted better performance with the mesh display than with the pie display. In CID, it is important to make a decision quickly, and the precise value of the probability is not of great importance. Although a pie chart can be read accurately, mesh-type formats can be read more quickly. They may provide an "at-a-glance" representation of reliability that is accurate enough for CID purposes.

Finally, our study involved several automation reliability levels that varied frequently over time. We did this to reflect moment-to-moment changes in reliability levels in combat. If participants can attend to the display and calibrate their reliance in this situation, then reliance should be affected by reliability level. To the extent that reliability level is ignored, reliance should not be affected by reliability level. Thus, the question of whether the reliability level affected the response criterion serves as a manipulation check of whether participants could interpret the displayed reliability information and respond to it.

Method

Participants. For Experiment 1, 30 University of Toronto students (20 males) with an average age of 22 years and normal or corrected-to-normal visual acuity were recruited. Complete data were collected from 28 participants, and only these data were used in the analysis. Participants were reimbursed for their participation. A small monetary bonus was provided to the top 10 participants.

Apparatus and stimuli. The IMMERSIVE (Instrumented Military Modeling Engine for Research Using Simulation and Virtual Environments) synthetic task environment served as the test bed. Developed by Defence Research and Development Canada–Valcartier, IMMERSIVE involves the modules of a commercial, first-person shooter game called *Unreal Tournament 2004*. Friendly and hostile forces can be distinguished

by differences in uniforms, weapons, actions, and feedback from the CID system.

The IMMERSIVE environment was displayed on a 50.8-cm flat-panel monitor set to high color (32-bit) resolution at 800 × 600 pixels. On each trial, a single target appeared and moved through the scene. The target's path through the scene varied from trial to trial. If the participant did not kill the target, the target would exit from view after 10 s and the trial would end. Participants controlled the direction of the weapon using a mouse and fired on targets by pressing the left mouse button. The participant obtained identification and reliability information from the display by pressing a designated key (the Insert key) on the keyboard when the gun was directed at a target.

Experimental design. A 2 (display type: pie, mesh) × 2 (display proximity: integrated, separated) × 5 (reliability level: 50%, 60%, 70%, 80%, 90%) mixed design was used. Reliability level referred to the reliability level of the "unknown" feedback. That is, it accurately represented the probability that the target was hostile given "unknown" feedback, or P(Hostile | "Unknown") for that reliability level condition. Table 1 shows how reliability level corresponded to the proportion of trials in which the type of feedback reflected the target identity. Display type was a between-subjects factor. Participants were randomly assigned to either the pie or the mesh condition. Display proximity and reliability level were within-subjects factors. Overall, half the targets were friendly and half were hostile.

The participants completed eight blocks of 105 trials each, for a total of 840 trials across two four-block sessions. Each session was 2 hr long, and the sessions were separated by at least an hour to reduce fatigue. In two of the four blocks during each session, we used separated displays; in the other two, we used integrated displays, with the order counterbalanced across participants. For each block, half the targets were friendly and the other half were hostile. Target identity and reliability level varied randomly trial to trial within a block.

Procedure. Participants were instructed to imagine themselves in a battlefield. Their task was to identify targets and, if hostile, kill the target. They were told that the value of a correct identification of a friend and of killing a hostile target was equal. Participants were informed (accurately) that a hostile target would never appear during blue-light "friend" feedback. They were also told that the red indicator, representing "unknown" feedback, was set to be less than 100% reliable to mimic system failures. Participants were informed that the reliability level was represented by a partially filled circle, representing the probability that the unknown target was hostile, and that identity information would sometimes be shown by the color of the larger circle or separately by the color of a dot placed above the circle. All participants were tested on their comprehension of instructions.

The participants were guided through a training session of 100 trials. For the first 50 trials, the experimenter pointed out relevant cues for target identity (e.g., target weapon and uniform) and provided feedback on accuracy. During the last 50 training trials, the experimenter gave tips to improve the participant's shooting skills.

Following the training session, the participants completed the first experimental session. Following a break of at least an hour, the participant completed the second session. No performance feedback was given during the experimental sessions.

Measures. All analyses described next were performed on the "unknown" feedback trials. Examination of the "friend" feedback trials (240 trials per participant) indicated that participants followed the perfectly reliable "friend" feedback more than 98% of the time.

We measured reliance on the "unknown" feedback using the response bias approach (Wang et al., 2009). We computed empirical β values using observed P(H) and P(FA) values. The probability of a hit, P(H), is equal to the probability of killing a hostile target. The probability of a false alarm, P(FA), is equal to the probability of killing a friendly soldier. Empirical β values were computed for each participant in each "unknown" feedback condition and were compared with theoretical $\beta_{optimal}$ values, which were calculated at each reliability level using the formula

$$\beta_{optimal} = \frac{V(CR) + C(FA)}{V(H) + C(M)} \times \frac{P(Friend \mid \text{“Unknown”})}{P(Hostile \mid \text{“Unknown”})},$$

where V(CR) = value of a correct rejection; C(FA) = cost of a false alarm; V(H) = value of a hit; and C(M) = cost of a miss.

We analyzed sensitivity on the "unknown" feedback trials using the SDT parameter d', also calculated from P(H) and P(FA) scores. We also directly examined measures of the two error types: the probability of a miss, P(M), and P(FA). On trials in which participants engaged the targets, the time from target appearance until the participant killed the target (kill time) was calculated.

Results

We calculated P(FA), P(M), kill time, and d' for each level of display type, proximity, and reliability level for each participant. A 2 (display type: pie, mesh) × 2 (display proximity: integrated, separated) × 5 (reliability level: 50%, 60%, 70%, 80%, 90%) mixed ANOVA was conducted on each of these dependent measures, with display type serving as the between-subjects variable. Effect size, r, was calculated for significant two-level effects. To increase normality and stabilize variances, an arcsine transformation was applied to all probability data submitted to the ANOVAs (Howell, 1992). A Greenhouse-Geisser correction, which results in noninteger degrees of freedom, was used when sphericity was violated.

We were interested in the quantitative effect of reliability level and not in comparisons of differences between pairs of means for each of the five levels. Therefore, a regression approach was used for the analysis of effects including reliability level. For main effects, the dependent measure was regressed on reliability level, and the individual regression slope coefficient for for a 10% change in reliability level (the difference between levels of the reliability manipulation) and its test statistic are reported. For interactions, this procedure was performed at each level of the categorical independent variable (i.e., display type or display proximity), with the slope and test statistic reported for each.

Miss rate. Display type affected miss rate, $F(1, 26) = 4.31$, $p < .05$, $r = .38$. Participants using the pie display missed more hostile targets, $M = .15$, $MS_E = .0008$, than those using the mesh display, $M = .12$, $MS_E = .0014$. Reliability level also affected miss rate, $F(1.54, 40.2) = 24.4$, $p < .001$, with miss rate decreasing as reliability level increased, $\beta_1 = -.05$, $t(278) = 10.81$, $p < .001$, $r = .54$. The untransformed means were .26, .18, .11, .09, and .06 ($MS_E = .01$) for reliability levels 50%, 60%, 70%, 80%, and 90%, respectively. No other effect was significant, $p > .05$ in each case.

FA rate. Only the main effect of reliability level on FA rate was significant, $F(2.20, 56.7) = 21.3$, $p < .001$, such that the FA rate increased with the reliability level ($\beta_1 = .06$), $t(278) = 5.87$, $p < .001$, $r = .33$. The untransformed means were .36, .50, .54, .60, and .63 ($MS_E = .01$) for reliability levels 50% through 90%, respectively.

Kill time. Only reliability level had a significant effect on kill time, $F(4, 104) = 10.9$, $p < .001$. Kill time decreased as the reliability level increased ($\beta_1 = -.08$), $t(278) = 3.03$, $p < .01$, $r = .18$. The untransformed means were 4.22, 4.01, 3.96, 3.88, and 3.88 s ($MS_E = .10$) for reliability levels 50% through 90%, respectively.

Sensitivity (d'). A total of 6 participants produced zero hits or FAs in at least one condition. SDT measures cannot be calculated in such cases, although Macmillan and Creelman (1991) have suggested corrections. However, in some cases, the corrections produced unstable SDT parameter estimates (i.e., small changes in the correction led to large changes in d' and β). Thus, data from the remaining 22 participants were used for sensitivity and reliance analysis.

Display type affected sensitivity, $F(1, 22) = 6.62$, $p < .05$, $r = .48$. As Figure 3 shows, participants using the mesh display were better able to distinguish hostile from friendly targets, consistently across reliability levels. No other main effect or interaction was significant, $p > .05$ in each case.

Reliance. To analyze the appropriateness of the participant's reliance on the aid, we computed $\beta_{optimal}$ for each reliability level, fit a linear regression model to these values, computed empirical β values for each condition, and then regressed these empirical values on the optimal model. Participants were told that their score was the sum of the number of correct identifications of friends and successful hostile target

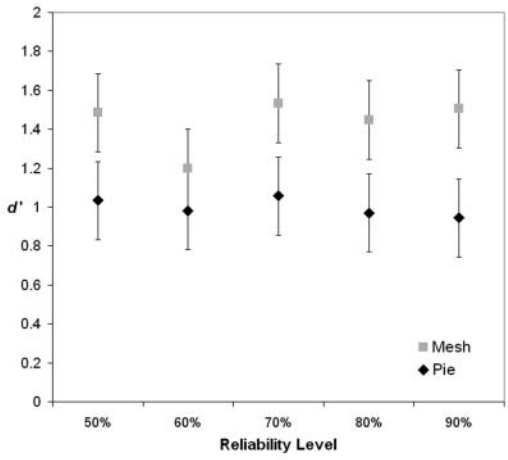

Figure 3. Experiment 1. Sensitivity (d') as a function of display type and reliability level.

kills. Therefore, the value of correct identification of friend was set equal to the value of killing a hostile target.

Given our instructions, the effect of payoffs should be

$$\frac{V(CR)+C(FA)}{V(H)+C(M)}=\frac{(1+0)}{(1+0)}=1,$$

and therefore Equation (1),

$$\beta_{optimal}=\frac{P(Friend\,|\,\text{“Unknown”})}{P(Hostile\,|\,\text{“Unknown”})}.$$

Reliability level (RL) is equal to P(Hostile | “Unknown”); therefore, P(Friend | “Unknown”) is equal to 1 – *RL*. Thus,

$$\beta_{optimal}=\frac{(1-RL)}{RL}=\frac{1}{RL}-1. \qquad (2)$$

This equation for $\beta_{optimal}$ was then fit to the β_{actual} values for each combination of display type and proximity. R^2 was calculated by computing the total sum of squares (SST) and the error sum of squares (SSE) as follows for $(i-1)$ participants in $(j-1)$ levels of the independent variables:

$$SST=\sum_i\sum_j(y_{ij}-\bar{y}_{..})^2 \qquad (3)$$

$$SSE=\sum_i\sum_j(y_{ij}-\bar{y}_{i.}-\hat{y}_{.j}+\bar{y}_{..})^2 \qquad (4)$$

The average score of each participant at each reliability level, $\bar{y}_{i.}$, was included in the calculation of SSE as reliability level was a within-subjects factor. R^2 was calculated using the formula

$$R^2=1-\frac{SSE}{SST}.$$

Fits with a larger R^2 value indicate that the $\beta_{optimal}$ equation accounted for a larger proportion of variance in β_{actual} in each condition. Given the fixed values for $\beta_{optimal}$, this model was constrained, reducing R^2 values relative to conventional linear regression. Furthermore, with a constrained model, it was possible for R^2 to be negative if a horizontal line was a better fit to the data (i.e., SSE was larger than SST). The R^2 values were .06 for integrated pie, .48 for integrated mesh, −.12 for separated pie, and −.41 for separated mesh (Figure 4).

To assess differences in reliance behavior between the conditions, we computed the linear slope of the β values across the five reliability levels for each participant included in the SDT analysis. A more negative slope indicates larger shifts in reliance with changing reliability levels. We submitted the slopes to a 2 (display type: pie, mesh) × 2 (display proximity: integrated, separated) mixed ANOVA. There was a main effect of display proximity, $F(1, 22) = 7.02$, $p < .05$, $r = .49$, such that integrated displays had a more negative slope, $M = -1.70$, than separated displays, $M = -.846$, $MS_E = .15$. Neither the main effect for display type nor the interaction was significant, $p > .05$ in each case.

To assess whether participants changed their reliance optimally, we compared the slopes for each display proximity condition to the slope of the optimal values: −2.19. Participants using the separated display had a significantly more positive slope than optimal, $t(23) = 3.78$, $p < .01$, $r = .61$, whereas the slopes for the integrated display did not differ from optimal, $p > .05$.

In summary, for the integrated displays, empirical criterion values dropped with reliability level, as predicted by the optimal beta model (see Figure 4). R^2 values were positive in each case. In contrast, empirical criterion values for separated displays were not well predicted by the optimal beta model. When participants used

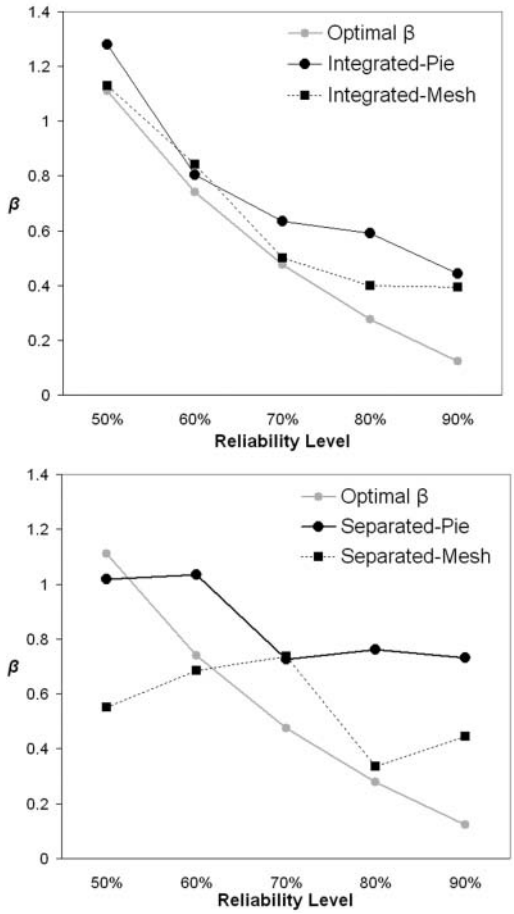

Figure 4. Experiment 1. Participants' reliance on "unknown" feedback as a function of display type and proximity.

the separated display, they relied on the aid less optimally.

Transfer effects. As a check, we examined whether there was any asymmetric transfer for display proximity, which was varied within subjects. Thus, each block was coded as to whether it was preceded by an integrated or a separated block. A one-way (preceding block: integrated, separated) between-subjects ANOVA was performed on β, d', and kill time. There was no effect of the preceding block on any of these variables, $F < 1$ in each case. Thus, no asymmetric transfer between integrated and separated displays was evident.

Discussion

The format used to display reliability and identity information affected both sensitivity and user reliance on an automated aid. Participants using the mesh display were better able to discriminate hostile from friendly targets than were participants using the pie display, in part because participants using the pie display missed more targets. Miss rate decreased and FA rate increased with increasing reliability level. Importantly, participants using the integrated display shifted their reliance level more optimally with changes in reliability level. This result implies that in the integrated condition, participants were better able to adjust their reliance with reliability level trial to trial.

Greater sensitivity was observed for the mesh display than for the pie chart. The results appear consistent with previous studies showing that mesh displays can be effective (e.g., Feldman-Stewart et al., 2007; Finger & Bisantz, 2002) and are in keeping with earlier results showing that the statistical properties of a visual array can be assessed quickly in parallel (e.g., Chong & Treisman, 2003). The mesh display may have afforded a more optimal strategy in a time-pressured situation, providing an at-a-glance indication of reliability. Although the pie display may have offered a more precise representation than did the mesh, greater precision may not be necessary to perform the task.

Assume that mesh displays provide an approximate representation of reliability level, and consider the task in terms of visual sampling: Participants judged the identity of a target by examining both the CID display (display elements showing reliability and identity information) and the visual scene (including the target's uniform and weapon). The observer has a limited amount of time to sample the scene, determine target identity, and aim and fire the virtual weapon to kill the moving hostile target. The display and target each provides an information channel that competes for the observer's attention. If it takes less time to sample the mesh than the pie display, then the observer would have more time to examine the target with the mesh display (since kill times in the mesh and pie conditions did not differ). Having more time

to examine the target should increase sensitivity (See, Howe, Warm, & Dember, 1995), thereby improving identification for the mesh display.

Consistent with this hypothesis, the group using the pie display not only had lower sensitivity than the mesh display group but also showed an increase in miss rate with no decrease in FA rate. It would appear that participants using the pie display did not have time to sample the target and were therefore more likely to indicate that a hostile soldier was not present than to shoot a friendly soldier. However, the increase in miss rate, although significant, was small (approximately 3%).

A similar visual sampling hypothesis can account for participants' reliance behavior. Separating reliability and identity information likely increased the number of visual locations from which to sample, thereby increasing the information access cost. Participants rarely fired on the "friend" feedback trials; therefore, they used the perfectly reliable identity information in that case. With the integrated display, participants could sample identity and reliability information during the same fixation and use the remaining time to examine the target to determine its identity. With the separated display, participants may have placed a higher priority on sampling the target and identity information than on determining the reliability level. However, without knowing the reliability level, participants would have had difficulty adjusting their reliance trial to trial. Again, there was no difference in kill time between the conditions, so more time spent sampling one channel meant less time to sample another. We examined these visual sampling hypotheses further in Experiment 2.

EXPERIMENT 2

In Experiment 2, we investigated the effect of reducing the viewing time on participants' ability to identify the target and their Reliance on the aid. Controlling the viewing time should enable us to further examine the role of attention in automation reliance and CID performance. Furthermore, the soldier faced with a CID decision does not always have the luxury of unlimited time to make a decision.

Consider that the observer can sample from the scene and from the CID display to obtain identity information and from the CID display to obtain the reliability level. With a shortened display time, there may be insufficient time to sample all information sources. This situation has two possible implications: (a) Reliability level is not sampled because there is insufficient time, and therefore the decision criterion is not affected by system reliability level, or (b) identity information is not sampled accurately, and therefore sensitivity is reduced. Note that to the extent that identity information is not sampled, participants are reduced to a guessing strategy, introducing a source of random error.

We argue that these effects are more likely to occur in certain conditions. For instance, if identity and reliability level are spatially separated, an eye movement (or covert attentional shift) will be required. Thus, participants using the separated display may rely on the aid less optimally or show decreased sensitivity relative to the integrated display. With the mesh display, the reliability level may be extracted more quickly (since participants were better able to identify the target with the mesh than with the pie display in Experiment 1, and they took no more time to do so). Thus, we would expect to see more optimal reliance or greater sensitivity for the mesh display than for the pie display.

We varied the amount of time participants had to sample the scene in Experiment 2 (400 ms or 800 ms). The values were based on pilot testing but also on the assumption that within 400 ms, there would not be sufficient time to fixate more than once. A fixation takes approximately 250 ms to 300 ms (Salthouse & Ellis, 1980) and a saccade approximately 20 ms to 30 ms, with approximately 150 ms necessary for the initiation of the saccade (Salvucci, 2000). Within 800 ms, one might expect two to three fixations with one to two saccades. Given that the reliability in the separated display was likely not foveated when the participant fixated on the identity indicator (greater than 3° of arc given approximately 50 cm of viewing distance), one might expect reliance on the separated display to be particularly affected by reduced display time.

Method

Participants. For Experiment 2, 26 participants (7 males) with an average age of 25 years

and normal or corrected-to-normal acuity vision were recruited. Complete data were collected from 24 participants, and only these data were used in the analysis. Experiment 2 was run at Defence Research and Development Canada–Toronto, and participants were employees of the Department of National Defence or members of the local community. All participants were reimbursed for their participation, and the top 10 participants also received a bonus.

Apparatus and stimuli. Custom software displayed still screen shots from the IMMERSIVE simulation on a 48.3-cm (diagonal) monitor with a resolution of 1,280 × 1,024 pixels. The targets in the scene either faced the participant or were in profile to the right or left. The different orientations were used to increase stimulus variability and to create a more realistic target recognition task. The target always appeared in the same location so that visual search was not necessary.

Experimental design. A 2 (display type: pie, mesh) × 2 (display proximity: integrated, separated) × 2 (stimulus duration: 400 ms, 800 ms) × 5 (reliability level: 50%, 60%, 70%, 80%, 90%) mixed design was used. Display proximity, stimulus duration, and reliability level were within-subjects factors, and display type was a between-subjects factor.

Procedure. The instructions were similar to those used for Experiment 1. However, the Experiment 2 instructions stated that the participant would view a still image of a combat scene for either 400 ms or 800 ms and then indicate the target's identity. The participants performed two training sessions of 100 trials each, one in each display proximity condition. For the first 10 training trials, the display duration was 15 s to allow the participant to examine the target's appearance. The remaining training trials were 600 ms each. The experimenter provided feedback until participants could identify the target 70% of the time. The participant completed the remaining training trials unaided.

On each trial, the participant was shown the text "Press any key for next scene" in black text on a gray background. The text was centered on the display and provided a fixation point for the subsequent target. Following the key press, the screenshot was shown for either 400 ms or 800 ms. Then another gray screen with black text appeared with the question, "Do you shoot the target?" The participant used the mouse to select *yes* or *no*. The participant was then shown a *Go* button; clicking it entered the participant's response and initiated the next trial.

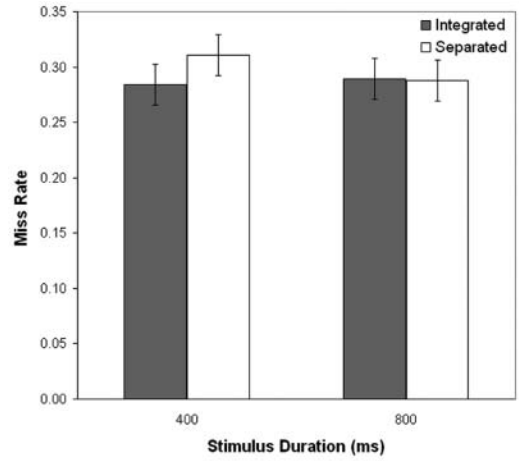

Figure 5. Experiment 2. Miss rate as a function of display type and stimulus duration.

The participants completed 1,680 trials, separated into eight blocks of 210 trials each. As in Experiment 1, participants were presented with a separated display for four blocks and an integrated display for the other four. For each display proximity condition, two of the four blocks had a stimulus duration of 400 ms and the other two a duration of 800 ms. The experiment was divided into two sessions of four blocks, each 2 hr long, separated by at least an hour to reduce fatigue. Each combination of display proximity and stimulus duration was presented once per session with the order counterbalanced across participants. For each block, the targets were friendly in half the trials and hostile in the other half, presented in random order. The reliability levels varied randomly trial to trial within a block.

Results

The analyses were similar to Experiment 1 for performance and reliance with the addition of the stimulus duration variable. Therefore, a 2 (display type: pie, mesh) × 2 (display proximity: integrated, separated) × 2 (stimulus duration: 400 ms, 800 ms) × 5 (reliability levels: 50%, 60%, 70%, 80%, 90%) mixed ANOVA

TABLE 2: Experiment 2: Miss and False Alarm Rate Means for Each Combination of Display Format and Reliability Level

	Reliability Level				
Display Format	50%	60%	70%	80%	90%
Miss rate					
Integrated	.42	.31	.26	.27	.17
Separated	.30	.35	.31	.25	.29
False alarm rate					
Integrated	.37	.47	.54	.63	.59
Separated	.54	.56	.46	.50	.52

was performed on the transformed miss rates, transformed FA rates, and *d'* values. Display type was a between-subjects factor, and all other factors were within subjects. We analyzed the appropriateness of reliance behavior using the same method as Experiment 1 for each combination of display type and proximity.

Miss rate. There was an interaction between stimulus duration and display proximity on transformed miss rate, $F(1, 22) = 5.77$, $p < .05$ (Figure 5). For the 400-ms stimulus duration, more hostile targets were missed with the separated than with the integrated display, but display proximity had no effect at 800 ms. There was a main effect of reliability level, $F(1.4, 31.0) = 5.44$, $p < .001$, but there was also an interaction between reliability level and display proximity, $F(4.1, 46.0) = 5.86$, $p < .01$. Miss rate declined with increasing reliability level for the integrated but not for the separated proximity condition (Table 2). Miss rate was regressed on reliability level for each display proximity condition, which provided simple slopes of $\beta_1 = -.06$, $t(118) = 6.04$, $p < .001$, $r = .49$, for the integrated display and $\beta_1 = -.02$, $t(118) = 1.49$, $p > .10$, $r = .14$, for the separated display. No other effect was significant, $p > .05$ in each case.

FA rate. There was a main effect for stimulus duration, $F(1, 22) = 7.76$, $p < .05$, $r = .51$, such that participants killed more friendly soldiers in the 400-ms ($M = .57$) than in the 800-ms condition ($M = .52$), $MS_E = .004$. There was a main effect of reliability level, $F(2.3, 50) = 3.06$, $p < .05$, but there was also an interaction between display proximity and reliability level, $F(3.2, 70.8) = 9.95$, $p < .01$ (Table 2). The FA rate increased with increasing reliability level for the integrated display ($\beta_1 = .08$, $t(118) = 5.91$, $p < .001$, $r = .48$), but not for the separated display ($\beta_1 = -.007$, $t(118) = .52$, $p > .10$, $r = .05$). No other effect was significant, $p > .05$ in each case.

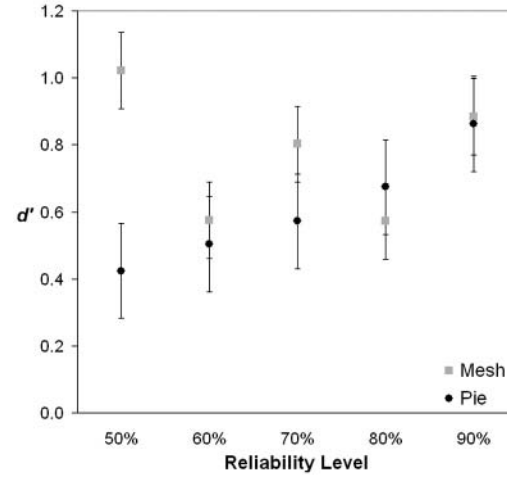

Figure 6. Experiment 2. Sensitivity (*d'*) as a function of display type and reliability level.

Sensitivity (d'). In this experiment, 6 participants were excluded from the analysis because they produced either zero hits or FAs for at least one condition; therefore, 18 sets of complete data were included for analysis. Sensitivity was higher with the longer stimulus duration, $M = .75$, than the shorter stimulus duration, $M = .55$, $F(1, 16) = 12.1$, $p < .01$, $r = .48$. Reliability level also affected sensitivity $F(4, 64) = 2.71$, $p < .05$,

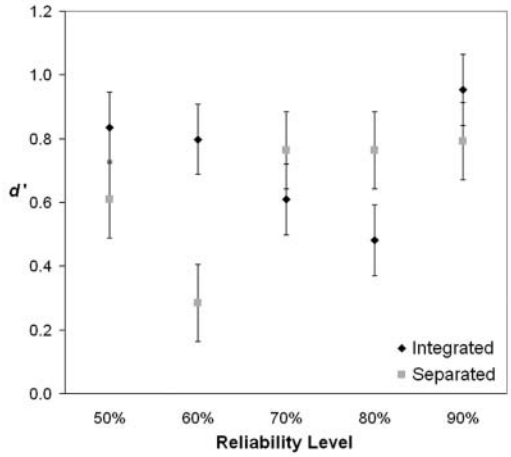

Figure 7. Experiment 2. Sensitivity (d') as a function of display proximity and reliability level.

but this effect was moderated by higher-order interactions. There was an interaction between display type and reliability level, $F(4, 64) = 3.25$, $p < .05$ (Figure 6), such that sensitivity increased with reliability level for the pie but not for the mesh display, where sensitivity appeared to vary randomly across levels of reliability. There was also an interaction between display proximity and reliability level, $F(4, 64) = 3.38$, $p < .05$ (Figure 7). There was high variability in sensitivity levels across reliability levels for both integrated and separated conditions, with sensitivity being particularly low in the separated condition at 60% reliability.

Finally, there was a three-way interaction between display proximity, stimulus duration, and reliability level, $F(4, 64) = 2.56$, $p < .05$. To explore the interaction, an analysis of simple interactions was carried out whereby the d' values for the integrated and separated displays were submitted to separate 2 (stimulus duration: 400 ms, 800 ms) × 5 (reliability level: 50%, 60%, 70%, 80%, 90%) repeated-measures ANOVAs. There was a Stimulus Duration × Reliability Level interaction for the integrated, $F(4, 68) = 2.68$, $p < .05$, but not the separated display, $p > .1$. As shown in Figure 8, although the difference between the two display duration conditions varied across reliability levels for the integrated displays, with the differences smaller at the higher reliability levels, sensitivity was generally higher

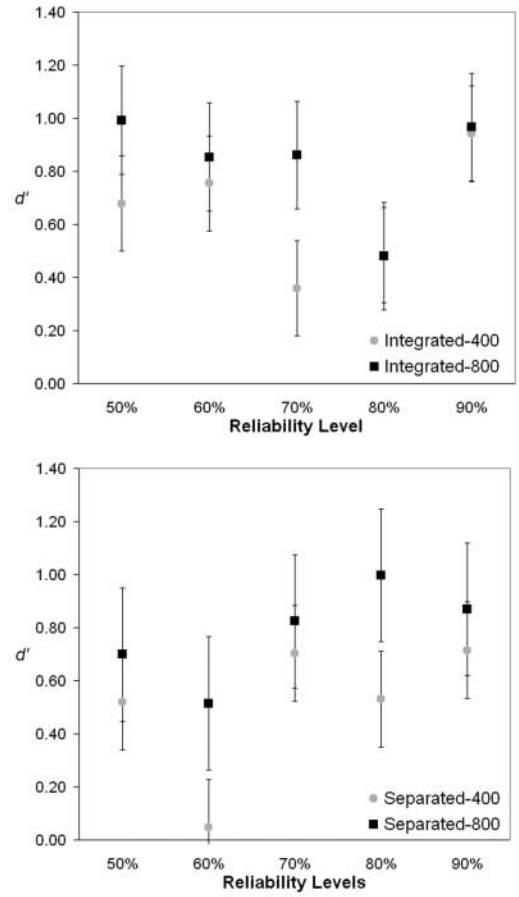

Figure 8. Experiment 2. Sensitivity (d') as a function of display proximity, stimulus duration, and reliability level.

for the longer stimulus duration for the separated displays.

Reliance. Regression slopes were calculated as described in Experiment 1 for each combination of display type, proximity, and stimulus duration. These slope values were submitted to a 2 (display type: pie, mesh) × 2 (display proximity: integrated, separated) × 2 (stimulus duration: 400 ms, 800 ms) mixed ANOVA. The main effect of display proximity approached conventional significance levels, $F(1, 10) = 4.81$, $.05 < p < .06$, $r = .57$. The integrated displays had a more negative slope ($M = -1.21$), than the separated displays ($M = -0.146$), $MS_E = 0.21$. No other main effect or interaction was significant, $p > .1$ in each case. The slopes of display

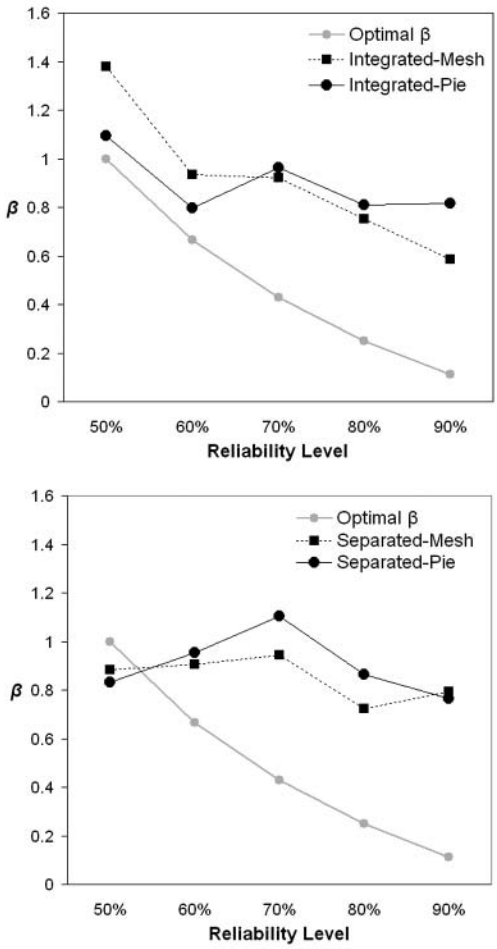

Figure 9. Experiment 2. Participants' reliance on the "unknown" feedback as a function of display type and proximity.

proximity were compared with the optimal value of −2.19. The slopes for the separated display were significantly more positive than optimal, $t(23) = 8.73$, $p < .01$, $r = .81$, which indicates that participants did not reduce their criterion sufficiently at higher reliability levels in the separated condition. In contrast, the slopes for the integrated display were not significantly different from optimal, $t(23) = 2.03$, $p > .05$, $r = .39$.

The equation for $\beta_{optimal}$ (Equation 2) was fit to the β_{actual} values for each combination of display type and proximity. (We did not include stimulus duration as a factor in this analysis because the slope analysis showed that it had no effect on response bias.) R^2 was calculated as in Experiment 1.

As Figure 9 shows, The R^2 values reveal that participants adjusted their response criterion more optimally when using the integrated mesh display than when using any other display combination. The R^2 values were as follows: integrated pie, .00; integrated mesh, .51; separated pie, −.41; separated mesh, −.22. In the conditions other than the integrated mesh, participants maintained a constant criterion just below 1.0 regardless of reliability level. Participants in all conditions were generally more conservative than optimal, especially at high reliability levels.

Discussion

Relative to Experiment 1, participants in Experiment 2 had limited time to view the CID scene. Similar to Experiment 1, when reliability level was integrated with identity information into a single display and represented using a random mesh format, increasing reliability level monotonically lowered the decision criterion such that it was closer to optimal. It would appear that the integrated mesh display best helped participants to take reliability level into account. With this display arrangement, participants appeared to be able to make use of the reliability level even given the reduced display times in Experiment 2. The other display formats led to reliance values that were more conservative than optimal for most reliability levels. In these conditions, the reliance was close to a neutral value, which would be optimal only in a 50% reliability condition. It would appear that participants did not use the reliability information in time-limited conditions.

There was no effect of stimulus duration on response bias, which is consistent with previous research showing that viewing time has a greater effect on sensitivity than response bias (See et al., 1995). The difference between the two stimulus durations was likely not large enough to affect bias.

Display type and proximity interacted with reliability level to affect sensitivity. Sensitivity was lower for the pie than for the mesh display at the lowest reliability level. Sensitivity improved with reliability level for the pie chart and was equivalent to sensitivity for the mesh display at

the highest reliability level. Given the limited time to sample, perhaps participants used the reliability information portrayed in the pie chart to the exclusion of information available in the scene, whereas with the mesh display, they could also sample the scene itself, improving performance at the low reliability levels.

Interaction comparisons showed that the three-way interaction with stimulus duration was attributable to an interaction between stimulus duration and reliability level for the integrated but not the separated display. For the separated display, sensitivity was generally higher in the 800-ms condition, which supports the suggestion that participants used the extra time to process additional information sources. This result also occurred for the integrated displays at low reliability levels, but there was little difference in sensitivity between the two display durations at the highest reliability levels (80% and 90%). At lower reliability levels, participants may have been able to use the additional time to process information in the scene, which may have been seen as unnecessary with high-reliability feedback.

Especially revealing is the high variability in the d' results. We hypothesized that with the brief presentations, participants would not be able to examine both the target and the display. It may be that on one trial, the participant could sample enough sources to make an accurate decision, whereas on the next, he or she could not and was essentially performing at chance, introducing a source of random error. This combination may have led to high variability in the responses between reliability levels.

GENERAL DISCUSSION AND CONCLUSION

The goal of this study was to design and evaluate visual displays that could help users rely on an automated CID system appropriately. Given the results of the two experiments in a simulated CID environment, there are advantages to integrating system reliability and identity information into a single graphical form. The decision criterion was closer to optimal for the integrated display (Experiment 1) or the integrated mesh (Experiment 2) than for separated display equivalents. In addition, mesh displays generally appeared to be the more effective format for depicting reliability information. Sensitivity was generally higher (and never lower) for the random mesh display as compared with the pie chart.

Some limitations should be noted. First, one could argue that it may have been difficult for participants in our study to adjust their criterion quickly enough trial to trial. If true, participants would have not adjusted reliance behavior as reliability level increased. This result did occur in some conditions but stood in contrast to other conditions in which reliance changed monotonically with system reliability levels. This finding implies that some display formats (most notably, integrated mesh) are in fact more effective in helping a user dynamically respond to changes in reliability level. Second, none of our displays produced perfectly optimal reliance, and further variations in display format might lead to even more optimal criterion setting. Our participants did not respond as much to changes in reliability level (i.e., probability) as they should have; this result is not a new one, being yet another example of the sluggish-beta phenomenon (Wang et al., 2009; Wickens & Hollands, 2000). Participants may also have had difficulty treating the cost of killing a friendly soldier and missing a hostile target as equal.

One design solution—unexplored in CID, to our knowledge—is to present higher-than-veridical reliability levels. However, as users work closely with such a system over time, we suspect it might lead to the mistrust and disuse issues so often observed with automation studies (e.g., Parasuraman & Riley, 1997). Regardless, we believe we have clearly identified the need to integrate reliability information with identity information for soldier decision aiding in CID. Finally, we acknowledge that the manipulation of display duration in Experiment 2 did not produce the clear anticipated effects of our sampling model. However, it did provide a controlled method for teasing apart some of the display effects, and the interactions with reliability level suggest that in time-limited conditions, one will see guessing strategies, especially with separated formats.

Providing users with system reliability information in a well-designed visual display improves both reliance on an automated CID aid and sensitivity in detecting targets. As CID is conducted in a high-risk environment under time pressure, the display must allow the user to quickly and easily obtain information from the display. The allocation of visual attention in CID with decision aiding deserves further study. Training can help soldiers attend to important cues, and attention allocation can help explain why some methods of displaying information are better than others. Direct measures, such as eye tracking, should be used to explore attention allocation in CID and validate the visual sampling hypothesis outlined in this study. In addition, eye tracking results could provide insight into interface design for other similarly demanding environments.

ACKNOWLEDGMENTS

This research was conducted under Contract No. W7711-06800/001/TOR awarded to the University of Toronto by Defence Research and Development Canada–Toronto (DRDC Toronto), as part of the DRDC Combat Identification project (15au). The first author received a scholarship from the National Science and Engineering Research Council for the duration of the project and a DRDC Post-Graduate Scholarship Supplement. We thank HUMANSYSTEMS (Guelph, Ontario), which provided technical assistance with the IMMERSIVE (Instrumented Military Modeling Engine for Research Using Simulation and Virtual Environments) synthetic task environment.

KEY POINTS

- Informing users of the reliability level of an automated combat identification system may help the users rely on the information more appropriately.
- We created displays that indicated reliability level in a pie or mesh graphical form, integrated with or separated from the identity information.
- Both in an interactive simulation and when shown still images, participants using the mesh display were generally better able to discriminate between friendly and hostile targets.
- Participants relied more appropriately on displays that integrated reliability and identity information.
- The mesh and integrated formats allowed participants to obtain reliability information more efficiently.

REFERENCES

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*, 157–162.

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*, 775–779.

Brand, J. L., & Orenstein, H. B. (1998). Does display configuration affect information sampling performance? *Ergonomics*, *41*, 286–301.

Chong, S., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*, 393–404.

Chong, S., & Treisman, A. (2004). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics*, *66*, 1282–1294.

Coury, B. G., Boulette, M. D., & Smith, R. A. (1989). Effect of uncertainty and diagnosticity on classification of multidimensional data with integral and separable displays of system status. *Human Factors*, *31*, 551–569.

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*, 147–164.

Feldman-Stewart, D., Brundage, M. D., & Zotov, V. (2007). Further insight into the perception of quantitative information: Judgments of gist in treatment decisions. *Medical Decision Making*, *27*, 34–43.

Finger, R., & Bisantz, A. M. (2002). Utilizing graphical formats to convey uncertainty in a decision making task. *Theoretical Issues in Ergonomics Science*, *3*, 1–25.

Hollands, J. G., & Spence, I. (1992). Judgments of change and proportion in graphical perception. *Human Factors*, *35*, 313–334.

Hollands, J. G., & Spence, I. (1998). Judging proportion with graphs: The summation model. *Applied Cognitive Psychology*, *12*, 173–190.

Howell, D. C. (1992). *Statistical methods for psychology* (3rd ed.). Belmont, CA: Duxbury Press.

Karsh, R., Walrath, J. D., Swoboda, J. C., & Pillalamarri, K. (1995). *The effect of battlefield combat identification system information on target identification time and errors in a simulated tank engagement task.* (Tech. Rep. ARL-TR-854). Aberdeen Proving Ground, MD: Army Research Laboratory.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*, 50–80.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide.* New York, NY: Cambridge University Press.

Maltz, M., & Shinar, D. (2003). New alternative methods in analyzing human behavior in cued target acquisition. *Human Factors*, *45*, 281–295.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*, 230–253.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors*, *49*, 76–87.

Salthouse, T. A., & Ellis, C. L. (1980). Determinants of eye-fixation duration. *American Journal of Psychology*, *93*, 207–234.

Salvucci, D. D. (2000). A model of eye movements and visual attention. In *Proceedings of the International Conference on Cognitive Modeling* (pp. 252–259). Veenendaal, the Netherlands: Universal Press.

Sanquist, T. F., Minsk, B., & Parasuraman, R. (2008). Cognitive engineering in radiation screening for homeland security. *Journal of Cognitive Engineering and Decision Making*, *2*, 204–219.

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.

Sherman, K. (2000). Combat identification system for the dismounted soldier. In *Proceedings of SPIE 2000: Digitization of the Battlespace V and Battlefield Biomedical Technologies II* (pp. 135–146). Orlando, FL: International Society for Optical Engineering.

Sherman, K. B. (2002). Combat ID coming for individual soldiers. *Journal of Electronic Defense*, *25*, 34–35.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, *51*, 991–1006.

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

Spence, I., & Lewandowsky, S. (1991). Displaying proportions and percentages. *Applied Cognitive Psychology*, *5*, 61–77.

Wang, L., Jamieson, G. A., & Hollands, J. G. (2008). Improving reliability awareness to support appropriate trust and reliance on individual combat identification systems. In *Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting* (pp. 292–296). Santa Monica, CA: Human Factors and Ergonomics Society.

Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and reliance on an automated combat identification system: The role of aid reliability and reliability disclosure. *Human Factors*, *51*, 281–291.

Wickens, C. D., & Carswell, C. M. (1995), The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors*, *37*, 473–494.

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

Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology and human performance* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Young, C. J. (2005). Fratricide. *Dispatches: Lessons Learned for Soldiers*, *11*(1).

Heather F. Neyedli is a PhD candidate in the Department of Exercise Science at the University of Toronto. She received her MASc in mechanical and industrial engineering from the University of Toronto in 2009.

Justin G. Hollands is a senior advisor for the Human Systems Integration Section at Defence Research and Development Canada–Toronto. He received his PhD in psychology from the University of Toronto in 1993.

Greg A. Jamieson is an associate professor of mechanical and industrial engineering at the University of Toronto. He received his PhD in mechanical and industrial engineering from the University of Toronto in 2002.

Date received: July 28, 2010
Date accepted: May 12, 2011