INTERFACE DESIGN FOR AN AUTOMATED COMBAT IDENTIFICATION SYSTEM: DISPLAYING RELIABILITY INFORMATION

by

Heather F. Neyedli

A thesis submitted in conformity with the requirements for the degree of Master of Applied Science

Graduate Department of Mechanical and Industrial Engineering
University of Toronto

© Copyright by Heather F. Neyedli (2009)
Abstract

Title: Interface design for an automated combat identification system: Displaying reliability information

Degree: Masters of Applied Science
Year: 2009
Student: Heather F. Neyedli
Department: Mechanical and Industrial Engineering
University: University of Toronto

Users have difficulty relying on automated combat identification aids; however, verbally informing users of the automation reliability has helped them rely on the automation more appropriately. A number of interfaces that displayed automation reliability information in real time were developed and tested. In Experiment I, participants used the interfaces in the IMMERSIVE simulation, a first person shooter game. The results showed that the form of the interface affected both reliance on the automation and sensitivity in discriminating hostile and friendly targets.

The difference in sensitivity and reliance may be attributed to how participants allocated their attention among the displays. In Experiment II, still combat scenes were presented to the participants for 400 or 800 milliseconds (as opposed to 10 seconds in Experiment I) to place additional time stress on attention resources. The results replicated the results of Experiment I, but sensitivity measures showed a dependence on reliability of the automation.
Acknowledgements

I have great appreciation for my supervisors, Dr. Greg Jamieson and Dr. Justin Hollands who have mentored me in completing this work. I thank them for their insightful feedback and support that saw me through this project I care deeply about. I also thank my committee members, Dr. Mark Chignell, and Dr. Timothy Chan for taking the time to review the work. I thank the people at HUMANSYSTEMS® for their expertise on the subject matter and implementation of the prototypes in IMMERSIVE.

I must also acknowledge my CEL colleagues who were a wonderful support over the past two years. I must thank Pierre for his programming knowledge especially on the second experiment and similarly Tony for his constant ability to bail me out of technical difficulties. A special acknowledgement goes to Lu Wang whose previous work, and excellent suggestions helped me start this project. I also thank my former ActionLab colleague, Jonathan, who continued to provide an outside, alternative view on both thesis and non-thesis topics.

Finally I express gratitude to my family and friends. I thank my parents and sister who have always believed in me and encouraged me, even in difficult times, to pursue my education. I also thank my friends, who have formed a large and caring extended family that has supported me over the past two years.

Funding for this research was provided by Defence Research and Development Canada (DRDC).
Table of Contents

Abstract ............................................................................................................................ ii
Acknowledgements ........................................................................................................ iii
Table of Contents .......................................................................................................... iv
List of Figures ................................................................................................................ iv
List of Tables ................................................................................................................ vi

1. Introduction to Combat Identification........................................................................ 1
   1.1. Automation Reliability .............................................................................................. 3
   1.1. Review of CID Automation Research ......................................................................... 4
      1.1.1. General Research on CID Systems .................................................................. 4
      1.1.2. Research on Reliance and CID Systems .......................................................... 6

2. Human-Machine Interface (HMI) Design ................................................................... 9
   2.1. Scope .......................................................................................................................... 9
   2.2. Displaying Reliability Information ............................................................................ 9
      2.2.1. Human Cognition and Uncertainty ................................................................. 9
      2.2.2. Methods of Displaying Uncertainty Information .......................................... 11
   2.3. Information Requirements for CID HMI design ....................................................... 17
      2.3.1. Information Requirements Sources ................................................................. 17
      2.3.2. Information Requirements for Basic System Use ............................................ 17
      2.3.3. Information Requirements for Appropriate Reliance on CID information ....... 18
   2.4. Design Recommendations ...................................................................................... 19
      2.4.1. Review of Prior Designs .................................................................................. 20
      2.4.2. Design Concepts ............................................................................................. 20

3. Experiment I .......................................................................................................... 28
   3.1. Introduction .............................................................................................................. 28
      3.1.1. Hypotheses ....................................................................................................... 29
   3.2. Method ..................................................................................................................... 30
      3.2.1. Experimental Design ....................................................................................... 30
      3.2.2. Apparatus ........................................................................................................ 31
      3.2.3. Participants ...................................................................................................... 32
      3.2.4. Procedure ....................................................................................................... 32
      3.2.5. Data Analysis .................................................................................................. 34
   3.3. Results ..................................................................................................................... 39
      3.3.1. Order Effects ................................................................................................... 39
      3.3.2. Miss (Miss Hostile Targets) ............................................................................ 40
      3.3.3. False Alarm (Friendly Fire) ............................................................................. 41
      3.3.4. Time to Kill Target ......................................................................................... 42
List of Figures

Figure 1: Hypothetical membership functions for propositional indicators of uncertainty from Bisantz et al, 2005 .......................................................... 12
Figure 2: Example of a colour pallet to add context to the above form ......................... 14
Figure 3: Degrading graphic to convey uncertainty: graphic on the right indicates decreased certainty from the graphic on the left. (Adapted from Finger & Bisantz, 2002) ............. 15
Figure 4: Proposed CID interface ................................................................................ 21
Figure 5: Display with degraded feedback ..................................................................... 25
Figure 6: Display with continuous analogue proportion feedback. ............................... 26
Figure 7: Display with discrete analogue proportion feedback ...................................... 26
Figure 8: “unknown” and “friend” feedback views for a separated display with continuous analog proportion reliability feedback ......................................................... 27
Figure 9: Integrated pie and mesh displays as implemented in IMMERSIVE ................... 31
Figure 10: Separated pie and mesh displays as implemented in IMMERSIVE ................. 31
Figure 11: Effect of block order on sensitivity. ......................................................... 39
Figure 12: Effect of block order on average kill time .................................................... 40
Figure 13: The main effect of reliability level on miss rate ........................................ 41
Figure 14: Main Effect of Reliability Level on False Alarm Rate .................................. 41
Figure 15: Main Effect of Reliability Level on Kill Time ............................................. 42
Figure 16: Main effect of display type on sensitivity, d’ .............................................. 43
Figure 17: Interaction Effect of Display proximity and Reliability Level on C ............ 44
Figure 18: The three-way interaction effect of display format, display type and reliability level on decision criterion C ................................................. 45
Figure 19: Comparison of $\beta_{\text{actual}}$ to $\beta_{\text{optimal}}$ for each display method and proximity combination 47
Figure 19: Model of object recognition ..................................................................... 58
Figure 20: Attention Allocation - Separated display .................................................. 59
Figure 21: Attention Allocation - Integrated-Pie display ............................................ 60
Figure 22: Attention Allocation - Integrated-Mesh display .......................................... 61
Figure 23: Screen view sequence for one trial ............................................................. 66
Figure 25: Effect of block combination on $d'$ ......................................................... 68
Figure 26: Effect of block order on $d'$ .................................................................... 68
Figure 27: Interaction effect of Stimulus Duration and Display proximity on Miss Rate .......... 69
Figure 25: Interaction Effect of Reliability Level and Display proximity on Miss Rate .......... 70
Figure 26: Interaction effect of Reliability Level and Display proximity on False Alarm Rate... 71
Figure 28: Interaction Effect of Reliability Level and Display Type on \( d' \) ......................... 72
Figure 29: Interaction Effect of Reliability Level and Display proximity on \( d' \) ............... 73
Figure 30: Three-way interaction of Reliability Level, Display proximity and Stimulus Duration on \( d' \) ................................................................................................................................. 74
Figure 31: Interaction Effect of Reliability Level and Display proximity on \( C \) .................... 75
Figure 32: Comparison of \( \beta_{\text{actual}} \) to \( \beta_{\text{optimal}} \) for each Display proximity and Type combination .. 77
Figure 33. Visual Sampling of Mesh Display ........................................................................... 80
Figure 34: Sampling of Pie Display at different reliability levels ............................................. 81
List of Tables

Table 1: Outcome matrix for “unknown” feedback trials................................................................. 34
Table 2. $R^2$ values for each display combination when the $\beta_{optimal}$ equation was fitted to the
collected $\beta$. ............................................................................................................................ 46
Table 3: Simple Slopes of Stimulus Duration and Display proximity regressed on Reliability
Level............................................................................................................................................. 75
Table 4: $R^2$ values for each display combination when the $\beta_{optimal}$ equation was fitted to the
collected $\beta$. ............................................................................................................................ 76
List of Appendices

Appendix A: Information Requirements............................................................... 93
Appendix B: Informed Consent Form - Experiment I......................................... 99
Appendix C: Instruction Script- Experiment I.................................................. 101
Appendix D: Assessment of Instruction Comprehension ............................... 104
Appendix E: Trust Questionnaire ...................................................................... 105
Appendix F: Informed Consent for Experiment II ............................................ 108
1. Introduction to Combat Identification

Fratricide is an unfortunate consequence of war arising from the chaotic environment and the time pressure under which soldiers must make combat decisions. Not only is the death of a fellow soldier tragic, fratricide also undermines morale, reduces public support of an operation and can reduce mission effectiveness if soldiers hesitate to engage a threat (Bourn, 2002; Young, 2005). To reduce fratricide and its subsequent effects, effective combat identification (CID) --the ability to differentiate friend from foe-- is needed.

The characteristics of modern warfare have changed the nature of CID. Distinct sides have often characterized past conflicts (Bourn, 2002). Factors such as the type of uniform worn and equipment used distinguished between the sides’ appearances. The sides were also distinguished by physical location as troop position was divided by geographic features, or delineated through trenches and front lines. In contrast, present conflicts are more ‘asymmetric’ (Bourn, 2002), with the difference between friend, foe and non-combatant individuals becoming less evident. Enemy combatants can belong to various factions and militias whose organization is not formalized and may not have distinguishing uniforms, equipment, tactics, or locations. Friendly forces have also become diversified as coalition operations have become more common, with members from multiple nations and elements of the armed forces (army, air force, navy). In addition, there is an increase in the number of neutrals -- civilians who are not engaged in combat -- as asymmetric conflicts often occur in urban locations (Bourn, 2002). All the above factors increase the challenge of CID.

Preparation for effective CID begins before soldiers set foot on the battlefield. Members of the forces are briefed and trained in Tactics, Techniques and Procedures (TTPs), which provide a universal protocol for all units and soldiers working together (Snook, 2002). Tactics are global concepts that apply to the movement and placement of units in combat, while
Techniques and Procedures govern how individual units or soldiers perform tasks. This shared understanding attempts to establish discipline in a chaotic environment by improving knowledge of troop location and procedures. More immediate to combat, soldiers are briefed on the intelligence regarding the current situation, giving them an indication of enemy numbers, disposition, weapons, and civilians in the area (Bourn, 2002; Snook, 2002).

Once in the combat theatre, situation awareness becomes important for effective CID (Everett, 1996). Situation awareness in a military context includes knowledge of who is in the combat theatre and more specifically where the various entities (e.g., friendly and enemy troops) are located. The previously established TTPs can assist in maintaining situation awareness. Also, various technologies have been developed to assist the soldier in situation awareness and CID, one of which is the identification friend or foe system (IFF) (Boyd et al., 2005). Originally used on aircraft (Snook, 2002), personal, rifle-mounted, IFF systems have been developed for dismounted infantry soldiers (K. Sherman, 2000).

IFF systems for dismounted infantry consist of a rifle-mounted interrogator that sends out a coded laser inquiry when manually activated (K. Sherman, 2000; K. B. Sherman, 2002; “SIMLAS”, 2006). If the laser signal is intercepted by a helmet mounted transponder, the transponder emits an omni-directional radio frequency (RF) response and feedback is provided to the interrogating soldier, indicating a friendly soldier. If the laser signal is not intercepted by a transponder, no feedback is provided. The system is therefore characterized by highly reliable “friend” feedback, but provides ambiguous information when no transponder feedback is received. In the case of no RF feedback, it is possible that the CID system did not detect the RF feedback or the individual was a neutral, a friendly force without the technology, or a soldier whose unit was damaged. Therefore, the interrogating soldier cannot assume the target is an enemy, and fratricide could occur if the soldier assumes that no feedback implies ‘hostile’ status. The ambiguity can cause the soldier to rely on the automated system inappropriately.
1.1. Automation Reliability

The challenge becomes finding strategies help the users to rely on the ambiguous “unknown” feedback more appropriately; therefore, user reliance on imperfect automation must be considered. Automation can provide many benefits; however, the costs of introducing automation have been well documented (e.g., Bainbridge, 1983; Skitka, Mosier, & Burdick, 1999). The human-technology partnership must be understood to realize its benefit. Parasuraman and Riley (1997) described the problems of reliance on technology as the misuse and disuse of automation. **Misuse** occurs when individuals over-rely on the imperfect automation and **disuse** occurs when individuals under-rely or reject the automation’s capabilities. One important factor in automation use is automation reliability. A meta-analysis of studies examining automation reliance found that a reliability of at least 70% was necessary for the automation to be useful (Wickens & Dixon, 2007). A number of human and situational factors can also affect automation use. For example, if the operator’s workload is high, the operator is often more likely to engage the automation leading to potential misuse; however, if the automation has a high time or workload cost (e.g., a complicated start-up procedure), a time-pressured operator may disuse the system (Parasuraman Molloy, & Singh, 1993). Another factor is operator self-confidence, such that a highly confident operator may disuse reliable automation and a less confident operator may misuse unreliable automation.

An operator’s trust in the automation also affects automation reliance. Trust is characterized as an attitude or expectation that an agent (e.g., automation) will perform a certain behavior, or act toward a common outcome in a situation characterized by vulnerability or uncertainty (Lee & See, 2004). The relationship is then that trust is the attitude and reliance is the behaviour influenced by the attitude (Lee & See, 2004). In general, appropriate trust should lead to more appropriate reliance on automation. Trust should therefore be calibrated to the capabilities of the automation (Lee & Moray, 1994); that is, the level of trust in the automation
should correspond to the automation’s capabilities. Analogous to over- and under-reliance, users can overtrust or distrust automation if trust does not match system capabilities. Related to calibration is the resolution of trust, which refers to the precision of a user’s trust in relation to varying automation capabilities (Cohen et al, 1999). The level of trust should also be specific to each component of automation; that is, trust should have specificity (Lee & See, 2004). Specificity is important for the IFF system since high trust in and reliance on the reliable “friend” feedback is desired whereas the “unknown” feedback is less reliable. Participants in a study by Wang, Jamieson and Hollands (in press) trusted 100% reliable “friend” feedback more than less reliable “unknown” feedback (which was 67% or 80% reliable depending upon condition). In addition, trust was calibrated to automation capability: participants trusted the system more when the “unknown” feedback was 80% reliable than when it was 67%.

1.1. Review of CID Automation Research

1.1.1. General Research on CID Systems

A variety of CID systems similar to IFF systems have been examined experimentally with most showing no performance advantage and inappropriate reliance levels. Research performed with a BCIDS system (a similar IFF system mounted in tanks) found no overall performance advantage in either speed or accuracy of engagement decisions between a group provided with the BCIDS system and a control group who performed the task without the system (Karsh et al, 1995). The system had fairly high reliability (correctly identifying targets 90% of the time); however, both “friend” and “unknown” feedback could be erroneous. Even with high system reliability, participants who were provided with the BCIDS system activated it only 14.5% of the time. However, when they did activate it, accuracy was improved: activating BCIDS reduced the rate at which the participants classified an enemy as a friendly target. Karsh et al. (1995) suggested that participants might have activated the aid infrequently because they
felt they were good at the task or the task was easy, providing an example of operator self-confidence interfering with the use of reliable automation. The target slides varied in their difficulty as some targets appeared further away from the participant than others. However, the authors did not analyze whether participants activated the aid more often or showed a performance advantage on the more difficult slides. While reliance was not explicitly examined in this study, the participants erred on 11.5% of the trials where the automation erred on 10% of the trials therefore increased reliance on the automation may have been beneficial.

Kogler (2003) analyzed a CID system under varying levels of difficulty. This study used a rifle-mounted CID system and participants were placed in a realistic scenario where they fired upon targets in a shooting range. The experimenter manipulated visibility and found that an IFF system reduced the number of friendly fire incidences when visibility was poor; that is, the system was the most useful when the task was more difficult. Reliance on the automation was only measured subjectively in the study, but participants indicated they used the system more than their own vision to determine the identity of the target when visibility was decreased and when the system was more reliable. These results are consistent with the literature regarding trust and reliance on automation. There is a discrepancy in reaction time findings between this study and Karsh et al, (1995). In Karsh et al, 1995, when participants used the system, their response time increased; however, in the Kogler (2003), there was no difference in reaction time between the aid and no-aid conditions. This may have occurred as participants in Kogler (2003) were instructed through rules of engagement (ROEs) to ‘shoot on sight’ all threat targets, whereas accuracy was stressed in the BCIDS study. Kogler (2003) suggested that the additional time of system inquiry may have been offset by the decrease in decision time due to the aids’ assistance.
1.1.2. Research on Reliance and CID Systems

In the above studies, some information was provided or could be inferred about automation reliance but reliance was not directly measured. Two studies explicitly examined reliance on automated decision aids (Dzindolet et al., 2001a, Dzindolet et al., 2001b); however, the aids used in these studies assisted in target detection rather than target identification. Detection occurs much earlier in information processing than identification and cognitively is a binary response (e.g., present versus absent), whereas identification requires narrowing of almost infinite possibilities. However, although the cognitive task is quite different, some comparison can be drawn between the aids since the identification aid output is binary (friend versus unknown), similar to the detection aid output of present or not present.

The first study found no performance advantage in identification for participants using the detection aid, regardless of the aid’s reliability (Dzindolet et al., 2001a). Reliance on the aid was defined by misuse and disuse rates. Disuse occurred when participants erred and the aid was correct, whereas misuse occurred when participants erred and the aid was incorrect. Using this method, the experimenters concluded participants were more likely to misuse the aid rather than disuse it regardless of trial difficulty or automation reliability even though participants were informed of the automation reliability. In a second study, the experimenters manipulated whether the participants were informed of the aid’s reliability (Dzindolet et al., 2001b) where the target-absent feedback (analogous to “friend” feedback) was perfectly reliable and the target-present feedback (analogous to “unknown” feedback) was fallible with a reliability of 67%. Participants were assigned to two groups, one which received only training where prior to the experimental sessions they were presented with the aid’s decision following their own decision to gain information about the aid implicitly. The other group received this training but was also explicitly informed of the reliability of the two types of feedback. A control group performed the task without an aid. No difference was observed between the explicitly informed and uninformed
participants and the aided participants committed more errors than the control group. Participants experienced the aid prior to measurement through either training or training and explicit instruction requested fewer second views of the slides when they received the perfectly reliable target-absent feedback. However, this indicates even when participants were explicitly informed the target-absent feedback was perfectly reliable, they still requested second views of the slides about 20% of the time. While it is understandable that participants may have had difficulty in understanding and using probability information about the target-present feedback (that was 67% reliable), it is perplexing that the participants did not use simpler target-absent feedback, which can be reduced to a if-then type statement: if the feedback is ‘target absent’, then the target is absent.

Providing the automation reliability level to the users has previously been found to help them rely on the automation more appropriately, perhaps by allowing a better calibration of trust to automation capabilities (Lee & See, 2004). However, informing participants of automation reliability taps into the analytical processes of trust when analogical and affective processes can also govern trust in automation. These processes can interact and affect one another; however this relationship is asymmetric where affective processes have a greater effect on analytic processes than analytic processes on affective (Loewenstein et al., 2001). Therefore, affective processes cannot be underestimated and their presence may provide some explanation for the contradictory results in the target detection studies.

Wang et al. (in press) investigated the relationship between trust and reliance, again manipulating whether participants were informed of the aid’s reliability. The participants performed a target identification task in a simulation consisting of a first-person shooter game at two levels of reliability. Besides being a target identification rather than a detection study, Wang et al, (in press) used a different method of assessing reliance. As discussed previously, many studies compare misuse and disuse rates to characterize reliance on the system. For example,
authors may characterize the users as either misusing or disusing the system (Dzindolet et al., 2001a, Dzindolet et al., 2001b) and while this distinction between reliance behaviour does offer some information regarding the human-automation relationship, it lacks a definition of appropriate reliance. If the goal is appropriate reliance on automation, one must be able to define what appropriate reliance is. Wang et al. (in press) adopted the decision criterion parameter from signal detection theory (Macmillan & Creeman, 1991) to define appropriate reliance. Measures of this parameter include C, D and $\beta$. Analysts can determine an optimal $\beta$ for a given situation based on the probability of a signal and noise occurring and the cost associated with errors and values associated with a correct judgment:

$$
\beta_{optimal} = \frac{P(Noise) \times V(CR) + C(FA)}{P(Signal) \times V(H) + C(M)}
$$

One can then compare the optimal value to an empirical $\beta$ to determine whether the user is relying on the automation appropriately. A more liberal $\beta$ than optimal indicates that the participant over-relied or misused the automation while a more conservative criterion indicates under-reliance or disuse. Thus this method provides a definition of appropriate reliance.

Using this new method, Wang et al. (in press) were able to determine that informing participants of the system reliability level allowed them to rely more appropriately on the aid. Reliance on the aid was also correlated with the trust in the “unknown” feedback as mentioned above and thus trust likely mediated the relationship between belief about the aid’s reliability and reliance behaviour. Unlike many of the previous studies, performance improved with the aid. Participants using the more reliable (80% correct) aid made fewer errors than those using the less reliable (67% correct) or no aid at all.

Combat identification is a high-risk, challenging task. Automated systems are intended to assist the soldier in performing the task, but studies have shown mixed results; in some cases the automation hindered CID performance. Informing the soldier of the automation reliability level
using verbal instruction has shown promise (Wang et al, in press). However, at higher reliability levels, participants’ reliance adjustment was suboptimal. Perhaps displaying the reliability information in a human-machine interface (HMI) would provide real time feedback about automation reliability, helping soldiers rely on the aid more appropriately.

2. Human-Machine Interface (HMI) Design

2.1. Scope

The interface design to provide real time feedback of automation reliability will consider a rifle-mounted, IFF automated CID system for dismounted infantry soldiers similar to prototypes as described in the literature (K. Sherman, 2000; K. B. Sherman, 2002; “SIMLAS”, 2006). The design will focus on conveying system reliability information to the user. Information requirements for the display will be described in detail; however, since the focus is on conveying reliability information, a review of the different methods for conveying this type of information is first provided. Other interface features pertaining to the system will also briefly be reviewed.

2.2. Displaying Reliability Information

2.2.1. Human Cognition and Uncertainty

Reliability information for automated decision aids implies a level of uncertainty regarding the automation’s assessment or recommendation; therefore, how human cognition deals with uncertainty information is discussed. Uncertainty information is often conveyed through proportions and probabilities. Humans can estimate proportions between 0.05 and 0.95 fairly accurately; however, Kahneman & Tversky, (1984) found subjective probability underestimates actual probability with the exception for very low probabilities where individuals overestimate the probability. The classic example of this phenomenon is the lottery where
individuals overestimate the low probability of winning, purchasing tickets, even though the chance of winning is extremely small.

Human judgment of probabilities is affected by whether the probability is framed as either a positive or negative, or as a risk or a gain. Individuals will choose a sure gain, but will take more risk regarding losses (Kahneman & Tversky, 1984). An example is the positively framed choice:

Option A: The probability of winning $60=0.8
Option B: The probability of winning $30=1.0

People will tend to choose option B because it is a sure gain even though the expected value of A is larger. However when the problem is framed negatively:

Option A: The probability of losing $60=0.8
Option B: The probability of losing $30=1.0

In this scenario, individuals will chose option A because they are risk adverse and wish to avoid a sure loss. Another example occurs when individuals are given the scenario of buying ground beef that is either 80% lean or 20% fat. Individuals will tend to prefer the positively framed choice of 80% lean beef (Keren, 2007). The caveat to the positively framed bias is the trust-choice incompatibility, where trust is greater for the negatively framed entity. If given the same statement about ground beef, and asked which butcher they would trust more, the participants would have greater trust in the butcher selling the 20% fat beef, although they prefer the 80% lean beef (Keren, 2007). While the trust-choice incompatibility may have implications for trust in automation, the subject has not been studied.

The availability heuristic affects humans’ perceptions of uncertainty regarding the frequency of events (Young, Wogalter, Brelsford, 1992) where the salience of an uncertain event affects the perceived frequency. The salience of an event in memory rather than on the probability of it occurring can affect estimation: individuals will overestimate the probability of a
more recent event. The estimation of the frequencies of an event can also be based on the severity of the consequences of an event, rather than the actual probability of occurrence. This heuristic occurs because individuals have difficulty estimating base rates of events so severe and/or recent events ‘stand out’ in the mind affecting the estimation. The representativeness heuristics is also a result of individuals having difficulty estimating base rates: the probability of an event will be estimated based on its superficial similarities to another entity rather than actual rate of occurrence.

Individuals also discount rewards that are to be delayed, where the perceived value of the reward decreases as the time to obtain the reward increases (Green, Myerson, McFadden, 1997). How much the reward decreases with time differs among individuals, with more impulsive or myopic decision making found, for example, among children (Green, Myerson & Ostaszewski, 1999) and substance abusers (Kirby, Petry, & Bickel, 1999). Finally, when working memory is taxed, individuals are more likely to be impulsive and take more risk in their decision making (Hinson, Jameson, & Whitney, 2003).

2.2.2. Methods of Displaying Uncertainty Information

2.2.2.1. Propositional Representations of Uncertainty

While probability information is numeric in nature, propositional or linguistic terms (e.g., very likely, probable, etc.) also convey probability information. Because propositional indicators are more vague than numeric indicators (e.g., 35%), a great deal of research has focused on translational issues (i.e., what numeric value is represented by the propositional indicator) and semantic issues (Teigen & Brun, 1999). Watkins (2000) proposed individuals perceive propositional indicators of uncertainty using fuzzy logic. A numeric uncertainty value can be seen as belonging more or less to a certain propositional descriptor (
Figure 1). For example, the propositional indicator *unlikely* can describe both 10% and 40%; however, 10% seems to have a higher degree of membership to the propositional term *unlikely* than 40%.

![Figure 1: Hypothetical membership functions for propositional indicators of uncertainty from Bisantz et al., 2005](image)

Semantic issues examine whether any additional meaning is conveyed beyond the numeric properties. Semantics can convey internal or external uncertainty (*I am unsure* vs. *it is unsure*) where internal uncertainty indicates uncertainty of the decision maker while external uncertainty indicates uncertainty in the situation (Kahneman & Tversky, 1982). Internally framed propositional statements are perceived as more responsible and certain than externally framed statements (Teigen & Brun, 1999). Semantics can also change whether a statement is positively or negatively framed, which has implications for the framing effect discussed above. Semantic meaning allows for additional inferences to be drawn from the probabilistic information that is not available in numeric representations.

Modifying terms can stretch, narrow or change the shape of the quantifying words’ (e.g., *unlikely, probable*) membership function (Watkins, 2000). For example, the modifier *very* shifts the membership function of the quantifying phrase *unlikely* to the left, further specifying the range of the membership function.
2.2.2.2. Analogue Representations of Uncertainty

Analogue forms or pictorial representations can also be used to represent uncertainty. Tukey boxes (i.e., box plots) and error bars are traditionally used to show variation and uncertainty in data (Watkins, 2000). These methods do not represent the uncertainty of the data, but rather the uncertainty that the data represents some ‘true’ value.

Proportional information is often displayed using a form that is rendered as partially filled, such as a meter or pie chart. When reading and comparing values from these displays, individuals use a vertical meter display more quickly and accurately than a pie display (Feldman-Stewart, Brundage, & Zotov, 2007). However, there are some important caveats. If a numeric value is included beside the display, this addition ameliorates the accuracy advantage in value comparison between the vertical meter and pie chart by increasing errors for the vertical meter and decreasing errors for the pie chart. Interestingly, in the same study, participants judged values from horizontal meters more slowly and less accurately than values from vertical meters. While it is unclear why this effect may have occurred, it may be because a vertical bar looks like a filled container or form, a situation commonly encountered due to the presence of gravity in daily life.

Pie charts are preferable to meters when the entire form of the indicator may not be clearly visible (such as when the proportion filled is lit against a dark background) as one can determine the total size of the circular form more readily than a linear, bar form. If only the proportion is lit, the unlit section in the bar form could extend indefinitely. Without knowing the size of the entire form, it is impossible to determine what proportion of the form is lit. Another consideration when using pie-charts is that providing tick marks to act as anchors around the edge of the circle may increase the accuracy of estimations. A cyclical bias pattern occurs when estimating values on a pie-chart where participants overestimate from 0 to 0.25 and 0.50 to 0.75 and underestimate from 0.25 to 0.50 and 0.75 to 1. Hollands and Dyre (2000) created a cyclical
power model, to account for the cyclical bias pattern. The addition of tick marks shortened the period of the cycle and reduced the amplitude of the participant’s bias.

2.2.2.3. Degrading Representations of Uncertainty

Degrading the quality of a graphical form is another analogue method of representing uncertainty. Some degrading forms are similar to the previous analogue forms discussed in that the form is portioned into sections to represent probability information. However, in the previous examples, the sections are congruent and organized as in a pie chart or meter, where the indicator fills in a predictable progression. With a degrading form the sections are displayed in a random pattern (e.g., random squares on a grid filled in) giving the form an additional affordance of appearing degraded as the probability decreases, thereby implying decreased reliability. If the disconnected sections indicating probability are lit against a dark background, the graphical form dims with increasing uncertainty. The disconnected sections should be evenly dispersed as participants have been shown to overestimate the proportion of clustered blocks (Goldstone, 1993) because the clusters are salient and the participant does not seek out additional cues in the rest of the form.

Another method to consider is a colour block that continuously fades or changes to another colour to indicate changing reliability or uncertainty. If a form is designed to change colour, the colour should convey meaning (e.g., red meaning danger) or a contextual colour pallet should be placed by the form (Figure 2), as only one colour will be displayed in the form.

![Colour Pallet](image)

**Figure 2:** Example of a colour pallet to add context to the above form
When patches of colour change, the user will have some contextual information regarding the meaning conveyed by the proportion; however, in a continuously fading colour block, the meaning or level of uncertainty is solely displayed through overall colour or shade.

A graphical element can also be made to appear degraded by reducing the image quality through such means as pixelation or fading. Finger and Bisantz (2002) suggested using a icon that becomes more pixelated as uncertainty increases (Figure 3).

![Figure 3: Degrading graphic to convey uncertainty: graphic on the right indicates decreased certainty from the graphic on the left. (Adapted from Finger & Bisantz, 2002)](image)

The quality of an image or photograph can also affect trust, with high quality images increasing trust (Yeh & Wickens, 2001; & MacMillan, Entin & Serfaty, 1994).

Seppelt and Lee (2006) used a fading colour graphical form to convey decreasing reliability of a sensor for an adaptive cruise control system. The form faded from bright to lighter yellow, then finally to grey as sensor reliability decreased; however, the display did not provide an advantage over a control display. The fading colour may have masked a faltering update rate, which inherently occurred with the failing sensor. The authors hypothesized that feature degradation through pixelation, as suggested by Finger and Bisantz (2002), may have been a more effective indication of the decreasing sensor reliability.

### 2.2.2.4. Comparison of Different Representations of Uncertainty

As discussed above, linguistic representations of uncertainty relate numeric representations through fuzzy membership functions. Consistent with that hypothesis, performance on tasks using each representation is similar; however, there is more variability when linguistic representations are used (Budescu & Wallsten, 1990). In a simulated stock...
buying task, participants gained the same amount of money (i.e., performed similarly) using linguistic and numeric representations, although participants made decisions more quickly (purchased stocks earlier in the simulation) using numeric representations (Bisantz et al, 2005). Analogue representations have also been compared to numeric representations of uncertainty. Feldman-Stewart et al. (2007) found that participants were more accurate and made comparison judgments more quickly with vertical and horizontal meters than with digital numeric representations on a medical decision making task.

In particular, measures of speed often differ among various analog methods of displaying proportions. The differences may be due to the cognitive processes used to judge the part-whole relationship of the proportion. If the whole is not readily available (such as in a bar graph), a summation model describes proportion judgments where the participant must sum up the parts (e.g., the bars in a graph) to determine the whole before making a subsequent ratio estimation (Hollands & Spence, 1998). As the number of components increases, the time required to accurately judge the proportion increases. If the whole is readily available (as in a pie chart), then only a ratio estimation is required, thereby decreasing the time needed for the complete estimation. As mentioned above, the ratio estimation for proportions can be described using a cyclical power model (Hollands & Dyre, 2000) where Steven’s power law is used to describe both the estimation of the part and the whole. If compression or expansion occurs for the stimuli (i.e., a power law exponent greater or less than 1), then the ratio of the part over the whole creates a cyclical pattern of increasing error between reference points.

One of the more interesting effects occurs when one compares performance using a degraded indicator alone and a degraded indicator with a digital indicator beside it. Finger and Bisantz (2002) found that, on a target identification task, participants were quicker, had higher trial scores and identified more targets using the degraded indicator alone, than using the combined (degraded and digital) indicator or the digital indicator alone. Feldman-Stewart et al,
(2007) also found similar results regarding error rates. Using degraded indicators without digital-numeric indicators, the participants had a lower error rate than when using pie charts. Adding the digital indicators ameliorated this effect by increasing errors for the degraded indicator and decreasing errors for the pie chart.

2.3. **Information Requirements for CID HMI design**

Determining information requirements is a necessary step in the successful design and evaluation of HMI that engenders appropriate trust in, and reliance on, a CID system. Information requirements identify specific information or relationships that must be presented to users through some artifact or training if the designer expects the user to have a degree of control over the acquisition, processing, or application of that information or relationship. Defining information requirements is an integral part of the systems engineering process of requirements specification (Burns & Hajdukiewicz, 2004).

2.3.1. **Information Requirements Sources**

Empirical results from Wang et al. (in press) and Jamieson, Wang and Neyedli (2008) were an essential source for information requirements for the HMI for the CID system (Appendix A, Table A-1). Further requirements were identified through review of the available functional specifications of CID systems that are already in use (e.g., K. Shurman, 2000). The latter requirements are essential for proper system operation (but perhaps not directly related to the goal of appropriate trust in, and reliance on, the CID system) as discussed below.

2.3.2. **Information Requirements for Basic System Use**

While the focus of the present interface is to engender appropriate trust in and reliance on, the CID system, it is prudent to also consider information requirements for basic system use. For example, some indication of the system power status is needed. This requirement will
become especially important if the system automatically queries targets, as it will confirm that
the system is able to provide feedback, some of which may be implicit. The HMI should also
display built in test results, battery power, and training results. As these requirements are not
essential for battlefield CID tasks, it is assumed they could be displayed in secondary or
peripheral displays.

2.3.3. Information Requirements for Appropriate Reliance on CID System

To carry out effective CID, the user must obtain knowledge about the targets in the
battlefield to determine whether they are friend, foe or neutral. Soldiers can use a number of
visual cues such as the dress and actions of the target to inform their identification. In addition,
an IFF system can provide information that assists in CID. To foster appropriate reliance on the
IFF system, the display should show the reliability of the CID information, particularly the
“unknown” feedback. The emphasis on the reliability of “unknown” feedback is based on the
nature of the system, from which “friend” feedback is known to be almost perfectly reliable.

Some participants in Wang et al. (in press) received reliability information through
training (i.e., they were informed of the system reliability prior to the experimental trials). While
verbally informing participants helped them adjust their reliance, participants were not relying on
the system optimally at some reliability levels. This may have been because the display provided
no information to reinforce the uncertainty of the “unknown” feedback; instead, participants
were required to remember this information. In contrast, participants were largely successful in
relying on the “friend” feedback (i.e., they adjusted their decision criterion appropriately), the
reliability of which had also been provided through training. Perhaps the certainty of the
information about the “friend” feedback decreased the load on working memory demand as
compared to the “unknown” feedback. One could hypothesize that displaying reliability
information for the “unknown” feedback may reduce working memory load and allow for more appropriate reliance on the system.

Providing system activation information on a manually activated system may be important especially when the “unknown” feedback is implicit. Jamieson et al. (2008) examined trust and reliance in a similar rifle-mounted CID aid when the “unknown” feedback was either implicit or explicit. The participants were more likely to activate the aid multiple times per trial when the “unknown” feedback was implicit and the participants trusted the implicit feedback less than the explicit feedback. Second, based on participant reports, multiple activations occurred because participants were unsure if they were successful in activating the aid initially. Therefore, any future CID system with a manually activated interrogator should provide an indication that the system has been activated. This would be intended to foster appropriate trust in the system. When the aid provided explicit feedback through a red light, the inquiry feedback also provided a form of activation information. Another possibility for displaying activation information would be to have a separate graphical form to indicate that the system has been activated. Explicitly displaying the activation information is especially important considering that existing systems take up to one second to provide feedback following system activation (K. Sherman, 2000 & Zari et al, 1997). This one-second delay would be perceptible to the user; therefore there is a need for separate activation feedback, even if the inquiry feedback is explicit.

2.4. Design Recommendations

Defining the information requirements allowed for redesign of the CID display so to include all of the information requirements. The information requirements also allowed for comparison between various designs.
2.4.1. Review of Prior Designs

An analysis of the display used in the IMMERSIVE simulation against the requirements is summarized in Table A-2 of Appendix A. The display in IMMERSIVE was similar to those available in existing systems that were reviewed (see K. Sherman, 2000 for an example) and an analysis of existing CID systems against the information requirements is contained in Table A-3, Appendix A. The existing displays provided information regarding the identity of individuals in the environment, but not probability information. The display used in Wang et al. (in press) & Jamieson et al. (2008) provided identity but not probability information and also did not give feedback about whether the participant had activated the system correctly.

2.4.2. Design Concepts

2.4.2.1. Assumptions

Two assumptions were made in conceptualizing graphical forms for the CID system HMI. The most important was that the HMI would have to be viewed in low lighting conditions. This assumption was imperative when considering displays of proportional information (which will be used when displaying reliability information) because the user must be able to determine the ‘whole’ even though only a portion (representing a proportion) of the display is lit. In the dark it may be difficult to determine how large the unlit part of a linear indicator display is, thus making it difficult to determine what proportion of the entire display is lit.

A second assumption was made regarding the precision of the reliability information, as the exact level of precision of the IFF aid was unknown. It was assumed that, in the unpredictable environment of asymmetric warfare, the level of precision might not require a continuous display. A continuous display may in fact convey to the user a higher level of precision than actually exists. Thus, the display concepts discussed below assumed that reliability indications would increment at discrete and equal intervals of 10%. With knowledge
of actual precision levels, the increments used for the following concepts could easily be modified.

2.4.2.2. Graphical Elements for Basic System Use

The display must contain graphical elements that aid in basic system use, such as a power status indicator and a battery power indicator, but are not directly relevant to battlefield CID. These elements, which are discussed below, are demonstrated in Figure 4.

![Proposed CID interface](image)

**Figure 4: Proposed CID interface**

Note that the figure also contains a circular graphical element that displays reliability information that will be discussed further below. The first two elements for basic system use are indications of whether the system is operational or not (i.e., on/off) (shown by the green word ‘ON’ in Figure 4), and of the system mode (training vs. combat), (shown by the orange word ‘TRAIN’ in Figure 4). Colour, position, and semantic meaning provide redundant cues to the status and meaning of the two indicators. The third graphical element is a screen containing a text field to display battery power, test results, etc. Since the tasks of checking battery power and built in tests results are not immediately important for battlefield CID, they do not need to be continuously displayed. Reducing them to one graphical element will reduce clutter. The text box, shown in gray in Figure 4, should be flexible enough that it is able to display a range of system information discussed above. As the specifics of this information were unavailable, and
the information displayed is not essential for battlefield use of the system, the specifications of this text box, (e.g., number of lines of text, font, etc.) are deemed to be beyond the scope of this project.

### 2.4.2.3. Design Concept for Acknowledgement of System Activation

Displaying acknowledgement of system activation was shown in Jamieson et al, 2008 to affect participants’ trust in, and reliance on, the CID feedback with participants activating the system multiple times when the feedback was implicit. Given that queries can take up to one second, even with explicit feedback, there is a noticeable transition state during which the CID system has been properly activated, but feedback is not yet available. One way to obtain appropriate trust in, and reliance on automation is to indicate the system state (Lee & See, 2004). Thus, the following design features aim to make explicit the state transition from activation, through delay, to feedback.

It is proposed that activation state information be displayed through two sensory modalities to increase channel security through redundancy. An auditory tone should alert the soldier that the system has been activated, taking advantage of the fast processing of auditory stimuli (Woodworth & Schlossberg, 1965). The auditory tone should be between 250-550 Hz at least 60db for reliable detection of the stimulus (Bridger, 2003). In addition, the activation should be echoed visually in the display as a redundant cue in case the auditory signal is masked by louder noise, which can be expected in the operational environment. It is suggested that the friend or foe indicator light (described below) turn yellow to indicate system activation and stay yellow until the feedback has been received (See Figures 5-7 for visual display designs containing an indication of system activation).
2.4.2.4. Design Concept for CID information

For soldiers to rely properly on the system and carry out CID effectively, both identification information and reliability of the identification information should be displayed. The initial design proposes the integration of these two requirements into one graphical form to facilitate the rapid parallel processing of this information. This is expected to allow for better integration of the information, leading to quicker decisions in the time-pressured environment. The elements in an integrated display can be within the user’s foveal or central vision (two degrees of visual angle), thereby reducing the use of saccades (Brand & Orenstein, 1998) or even covert shifts of attention, which may be an important consideration for a time-pressured task. One potential disadvantage of integrated displays is that it can be more difficult to extract component information (Bennett, Nagy & Flach, 1997). However, it is anticipated that soldiers should be able to extract feedback reliability from the integrated element because the colour and shape of the display are separable dimensions. A second possible disadvantage of the integrated display approach is that the “unknown” feedback must be explicit. The participants in Jamieson et al, 2008 that preferred the implicit feedback stated that this was because, even though the information was unreliable, the explicit feedback was salient and heavily influenced their decision towards shooting the target. However, the display of reliability information may ameliorate this effect. The “friend” feedback should be displayed explicitly. It is important that this information is salient due to the high reliability of this information and the role it plays in preventing fratricide.

While both the identity information and reliability information are used in performing the CID task, their relationship is dissimilar to other examples of high processing proximity in the proximity compatibility principle (PCP) literature. One may even consider the reading of the reliability level as a task that requires the user to determine a precise value; a task that is best displayed in a low proximity format (Wickens & Carswell, 1995). While the time-pressured
environment seems to call for an integrated display, because of the aforementioned disadvantages, both a separated and integrated display should be compared empirically. If no clear advantage for the integrated display were found, a separated display would allow for either implicit or explicit “unknown” feedback.

2.4.2.5. Comparison of Design Concepts to Display System Reliability

Three prototype graphical elements for displaying reliability information were developed. One prototype concept makes use of a degraded stimulus to display reliability information, while the other two make use of analogue proportion displays. A circular arrangement was used in all of the designs for the indicator of the combined inquiry results/reliability information, as it is easy to determine the whole size of a circle even if one part is not visible. This satisfies the requirement for a graphical element that is amenable to low light conditions.

Figure 5 presents a prototype that displays reliability information through degradation. When a “friend” signal is received, a blue circle is shown, conveying the near-perfect reliability of the “friend” feedback. If no signal is received, the unknown feedback appears as a red circle. However, random blocks in the grid are rendered in black, their number proportional to the $P(\text{unknown}|\text{“unknown” feedback})$, with the red blocks conveying $P(\text{enemy}|\text{“unknown” feedback})$. As the $P(\text{enemy}|\text{“unknown” feedback})$ decreases, the indicator would appear to dim and degrade with a decreased proportion of the grid rendered in red.
As an alternative to stimulus degradation, the reliability information could be displayed using congruent, organized proportions, two examples of which are shown in Figure 6 and Figure 7. The colour and shape (as the “unknown” feedback will be an incomplete circle and the “friend” feedback a complete circle) of the graphical element provide identification information, and the shape of the display (i.e., the proportion lit) provides reliability information.
Figure 6: Display with continuous analogue proportion feedback.

Figure 7: Display with discrete analogue proportion feedback
The concept in Figure 6 presents the reliability information in a continuous manner while the concept in Figure 7 presents it using segmented levels of probability. The segments were rendered as triangles, rather than having the outer edge as a rounded arc to provide better definition between the segments. The angle of the outer edge last segment provides a low-level visual processing cue as to the segments order in the form. Although the continuous display affords greater sensitivity in information display, it may indicate to the user a higher level of precision than actually exists. This may invite the user to form an incorrect mental model of the precision of the reliability information.

A proposed design for a separated display is contained in Figure 8. The graphical forms for basic system requirements remain the same as in the integrated display; however, separate circular graphical forms are now present for the identification and reliability information. The upper, smaller circle renders blue for “friend” feedback and in red for “unknown” feedback. The lower, larger circular form displays the reliability information, rendered in yellow, using any one of the three methods described above. When “friend” feedback is received, the lower form is completely filled indicating the highly reliable nature of the “friend” feedback. When no signal is received, the form fills in yellow proportional to $P(\text{enemy}|\text{“unknown” feedback})$ using one of the three methods described above.

Figure 8: “unknown” and “friend” feedback views for a separated display with continuous analog proportion reliability feedback.
2.4.2.6. Automatically versus Manually Activated Systems

All three of the displays discussed above use explicit feedback, which necessitates a manual query. If an automated system were to use explicit feedback, the ‘red light’ would continuously be present, except when the interrogator was pointed at a transponder, draining battery power. Jamieson et al. 2008 showed that participants in the manually-activated condition had a longer reaction time than those using an automated system. However, there are two benefits to a manually activated system. First, if the system is manually activated, the “unknown” feedback can be explicit and discrete (i.e., only occurs when a suspected target is queried, rather than a continuously displaying a ‘red light’ or other explicit feedback). Explicit feedback has the combined benefits of increased trust in the feedback as well as allowing for integration of reliability information into the feedback. Second, an automatic system requires continuous laser inquiries, which presumably drains battery power and could conceivably increase the soldier’s signal exposure in the field. This may be why currently implemented CID systems are manually activated. Because of the practicality of manual activation, both for the information that can be displayed as well as technical requirements, it seems prudent to focus display recommendations on manually-activated systems even with the tradeoff of decreased response time. However, any system that is automatic should have the ability to continuously display the reliability of the “unknown” feedback.

3. Experiment I

3.1. Introduction

The display concepts described in Section 2.4.2 varied in the means used to display the reliability of the inquiry feedback and whether the reliability information was integrated with or separated from the identity information. Integrating the information may be beneficial as both
pieces of information would often be used simultaneously in the time-pressured battlefield environment but separating the information allows for more flexibility in interface design. For the reliability information one display conveyed decreasing reliability using a degrading and dimming graphical element (Figure 5). The other two displays used an analog graphical element to indicate the reliability of the feedback in either discretely or a continuously (Figures 6 and 7). The continuous analog proportion display (Figure 6) and the degrading stimulus display (Figure 5) were used in the experimental design. They will be referred to as the pie and mesh display respectively. The discrete proportion display was not included in the present experiment to keep the protocol of reasonable size with the rationale that this display was not sufficiently distinguished from the continuous proportional display to warrant its inclusion.

The objectives of Experiment I were to determine 1) which method of displaying feedback reliability information afforded the best performance, and most appropriate trust in and reliance on an automated CID system and 2) whether integrating or separating feedback reliability and feedback identification information affords a performance advantage, and/or more appropriate reliance on and trust in the CID system.

### 3.1.1. Hypotheses

As the probability of a signal occurring affects response bias, knowledge of signal probabilities should affect the response criterion but not participants’ sensitivity in detecting targets (Macmillan & Creelman, 1991; Wickens & Hollands, 2000). Therefore, the best method or display of conveying feedback reliability would be that which induces more optimal shifts in the decision criterion.

Integrating reliability and with identification information into one graphical element should allow for more rapid processing in a time-pressured environment. If so, faster kill times should occur for the integrated HMIs; however, a performance advantage such as higher
sensitivity or more appropriate reliance may also occur. However, if participants could more accurately determine the reliability level in the separated display, the advantage could dissipate. Thus participants may rely on the separated HMI more optimally than the integrated HMI as the reliability information will have been separated from the feedback, allowing for more precise reading.

3.2. Method

3.2.1. Experimental Design

A 2 (display type: pie, mesh) x 2 (display proximity: integrated, separated) x 5 (reliability level: 50%, 60%, 70%, 80%, and 90%) mixed design was used. Display proximity and reliability levels were within subjects factors, while the display type was a between subjects factor. In other words, each participant used one of the display types in both an integrated and separated display at five reliability levels.

The participants completed 840 trials, separated into eight blocks of 105 trials. Participants were presented with a separated HMI for four blocks of trials and in the other four blocks an integrated HMI (the order was counterbalanced between participants). The participants completed the eight blocks of trials over two sessions, each two hours long, separated by at least an hour to reduce participant fatigue. Each session consisted of four blocks of trials. Within a session the blocks could be ordered IISS, SSII, SISI, ISIS, SIIS or ISSI (I=integrated, S=separated). Each order was used the same amount of times in the pie and mesh groups of participants. For each block, the targets were “friendly” for half the trials and “hostile” for the other half, in random order. The reliability levels varied randomly trial-to-trial within a block. The reliability level referred to the probability that the target was a hostile soldier on a given "unknown" feedback trial.
3.2.2. Apparatus

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 uses the modules of a commercial, first-person shooter game called Unreal Tournament 2004. Friendly and hostile targets were distinguished by differences in uniforms, weapons, actions, and feedback from the CID system. The HMI manipulations were implemented onto the CID system interface (Figure 9 & Figure 10). The graphical elements that corresponded to the information requirements for basic system use (Section 2.4.2.2) were excluded from the implementation as this information was not pertinent to the minute-to-minute battlefield CID.

Figure 9: Integrated pie and mesh displays as implemented in IMMERSIVE

Figure 10: Separated pie and mesh displays as implemented in IMMERSIVE
The integrated display (Figure 9) consisted of one graphical element containing the feedback information (i.e., the display rendered in either red for “unknown” feedback or blue for “friend” feedback) using either the pie method or the mesh method to display the reliability of the inquiry feedback. The separated display (Figure 10) contained similar pie and mesh graphical elements; however, these elements were rendered in yellow to separate this information from the inquiry feedback. The friend or foe indicator light appeared above the graphical element. This light was rendered in red or blue to indicate “unknown” and “friend” feedback, respectively. The participants were shown the feedback identification and feedback reliability information by manually activating the system using a key-press when the gun was directed at a target.

3.2.3. Participants

30 University of Toronto students (males: n=20) with an average age of 21.6 +/- 2.63 years with normal or corrected-to-normal visual acuity were recruited for Experiment I. Complete data was collected from 28 participants and only these data were used in the analysis. Participants were paid $40 CAD for their participation with a bonus of $10 CAD for ‘good’ performance. Good performance was defined as being above the 66th percentile for all participants for sensitivity (d’ values). Because of the length of the experiment as well as the large number of participants, the decision was made to base the bonus on relative performance rather than the ‘best’ performing participant. The assumption was made that this would set a realistic goal for the participants to achieve (as compared to the goal of being the best performer), and therefore be more likely to motivate the participants for the duration of the experiment.

3.2.4. Procedure

Prior to the first session each participant first passed a Snellen vision test, signed an informed consent form (Appendix B) and filled out a demographic information survey. The
participants were then given a sheet of instructions to explain the experimental procedure (Appendix C), and were tested on their comprehension of the instructions using the questionnaire in Appendix D.

Participants were instructed to imagine themselves in a battlefield. Their task was to identify targets in the scene and shoot the hostile targets. The participants were informed that the blue indicator, signifying “friend” feedback, would never appear when a target was hostile. However, the participants were told that the red indicator, representing “unknown” feedback, was set to be less than 100 percent reliable to mimic system failures. The instructions informed the participants that the reliability was displayed through a form that became partially filled, representing the probability that the unknown target was hostile. They were also told that the reliability and identity information would sometimes be shown as separate forms or in a single form. To mimic the time pressure of an engagement scenario, the participants had to kill hostile targets before they ran off the screen (approximately 10 s).

The participants were guided through a training session of 100 trials. In the first half of the training trials, the experimenter helped the participants with their identification skills by giving feedback on their accuracy and identifying relevant cues to target identity (e.g., weapons and uniform). In the second half of the session the experimenter gave tips to improve their shooting skills. Following the training session, the participants completed four mission blocks. After each block the participants completed a trust questionnaire (Appendix E). Following a break of at least one hour the experimenter briefly reviewed the instructions with the participant, who then completed the last four mission blocks and corresponding trust questionnaires.
3.2.5. Data Analysis

3.2.5.1. Main Analysis

The response bias method (Wang et al, in press) was used to analyze CID performance with signal detection theory parameters for sensitivity: $d'$, and response bias: $C$ and $\beta$ calculated on the “unknown” feedback trials using the outcome matrix in Table 1 and the following formulas:

$$d' = Z_{Hit} - Z_{FA} \quad (1)$$

$$C = -\frac{1}{2}[Z_{Hit} + Z_{FA}] \quad (2)$$

$$\beta = \exp\{d' \times C\} \quad (3)$$

Table 1: Outcome matrix for “unknown” feedback trials.

<table>
<thead>
<tr>
<th>Participant Response</th>
<th>States of the World</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hit (H)</td>
<td>False Alarm (FA)</td>
</tr>
<tr>
<td>P(H</td>
<td>Unknown)</td>
</tr>
<tr>
<td>Value=V(H)</td>
<td>Cost=C(FA)</td>
</tr>
<tr>
<td>Miss (M)</td>
<td>Correct Rejection (CR)</td>
</tr>
<tr>
<td>P(M</td>
<td>Unknown)</td>
</tr>
<tr>
<td>Cost=C(M)</td>
<td>Value=V(CR)</td>
</tr>
</tbody>
</table>

Both $C$ and $\beta$ were used: $C$ has the simplest statistical properties for parametric tests while $\beta$ can be easily compared with the optimal $\beta$ calculated based on the target probability and payoffs to assess reliance on the automated system. False alarm rates (the percentage of trials participants shot at friendly targets) and miss rates (the percentage of trials where participants did not kill hostile targets) were also analyzed to give an indication of what type of errors occurred in the CID task. To increase the normality of the probability data, an arcsine transformation was applied:

$$\text{Transformed Probability Data} = 2 \times \text{arcsine}[\text{Probability Data}]^{1/2} \quad \text{(Dzindolet et al., 2001a; Howell, 1992; Winer, 1991).}$$
Kill time was calculated as the time from when the trial started until the time the target was killed for the trials in which the participant killed the target. Trust in the whole aid, the “unknown” feedback, and the “friend” feedback was measured using 7-point scales. Questions were extracted from an empirically validated trust questionnaire (Jian, Bisantz, & Drury, 2000) with two questions added to assess trust in the “friend” and “unknown” feedback (Wang et al. in press).

A 2 (display type: pie, mesh) x 2 (display proximity: integrated, separated) x 5 (reliability level: 50%, 60%, 70%, 80%, 90%) mixed ANOVA was conducted on the transformed false alarm (FA) rate, transformed miss rate, time to kill a target, C and d’. A 2 (display proximity: integrated, separated) x 2 (display type: pie, mesh) mixed ANOVA was conducted separately on the subjective ratings of trust for the whole system, the “unknown” feedback and the “friend” feedback. Since reliability varied trial-to-trial within block, trust could not be assessed at each reliability level. Effect size was calculated for two level factors and contrasts using:

\[ r = \sqrt{\frac{F(1, df_R)}{F(1, df_R) + df_R}} \]  

(4)

Where df_R=residual degrees of freedom. If sphereicity was violated the Greenhouse-Geisser correction on degrees of freedom was used. All error bars displayed in results (section 3.3) represent the standard error of the mean.

### 3.2.5.2. Exploring Interaction Effects

The five-level reliability factor presented a challenge for exploring differences between levels of the factor as well as interaction effects involving the factor. Previous studies (e.g., Wang et al., in press; & Dzindolet et al., 2001) used no more than three levels of reliability level, which allowed for contrasts to be performed between factor levels. In the present study with five reliability levels, performing multiple comparisons would inflate the experiment-wide Type I
error rate and would not inform about the trends across the multiple reliability levels. To the author’s knowledge, no other automation study has used as many reliability levels; therefore, a novel approach had to be considered. Reliability level can be considered a ratio variable with quantifiable difference between levels and a defined zero point, which made it amenable to use regression to explore changes between factor levels. For main effects involving reliability level, reliability level was entered into a regression model and the individual regression coefficient (i.e., the slope) for the change in the dependent variable for a 10% change in reliability level is given. A test statistic to assess the significance of the coefficient was computed by dividing the coefficient by its standard error. A 10% change in reliability level data was chosen for reporting the slope value as data was collected at 10% intervals of automation reliability.

For two-way interaction effects, the above procedure was performed at each level of the other categorical independent variable (i.e., display type or display proximity), with the slope and test statistic reported for each.

Finally for three-way interactions, the difference in slopes was assessed to explore the interaction effects. The decision was made was compare the difference between slopes since the significance of each combination of independent variable (e.g., display type and display proximity) would not necessarily be indicative of the interaction effect. For example, in the case of the two-way interaction, for the interaction to be significant at least one of the simple slopes computed has to be significant. This is not the case for the three-way interaction as the effect of one of the independent variables may depend on the other variable being present in the model. In an effort to interpret the nature of the interaction, simple slopes of reliability level were calculated at each combination of display type and (i.e., integrated-pie, integrated-mesh, separated-pie and separated mesh). However, the categorical nature of display proximity and display type presented difficulties in testing for significant differences between the slopes. Therefore, the two categorical variables were given values of -1 (pie display and separated
format) and 1 (mesh display and integrated format) in order to standardize each variable with a mean value of 0. This produced a model with simple slopes that was amenable to using the method described by Aiken and West (1991) and Dawson and Richter (2006) for testing differences in simple slopes using values in the variance-covariance matrix to produce a test statistic. The difference tested the effect of one categorical variable while the other was held constant, effectively giving the same information as the tests performed for simpler two-way interactions.

### 3.2.5.3. Assessing Appropriateness of Reliance

To analyze the appropriateness of the participant’s reliance on the aid, a function was fit relating the reliability level to $\beta_{\text{optimal}}$ and then compared this equation with the participants’ $\beta_{\text{actual}}$ for each combination of display proximity and display type. As mentioned above, participants were told that their score was the sum of the number of correct identifications of friends and successful hostile target kills. Therefore, the value of correct identification of a friend was not the same as the value of correct identification of a hostile target, but rather the same as the value of killing a hostile target. In this experiment, when participants decided to shoot, they successfully killed targets 90% of the time. In other words, even if participants correctly identified a hostile target (i.e., a hit), only 90% of the time they received credit for a hit. To receive a better score, the participant needed to be more liberal (i.e., attempt to shoot more targets). Therefore, the value of correct identification of a hostile target should be the value of killing of hostile target multiplied by the successful kill rate of 90%. Therefore:

$$V = \frac{V(CR) + C(M)}{V(H) + C(FA)} = \frac{1 + 0}{1 \times 0.9 + 0} = 1.11$$  \hspace{1cm} (5)

Reliability level (RL) is equal to $P(\text{hostile} | \text{“unknown” feedback})$; therefore $P(\text{friend} | \text{“unknown” feedback})$ is equal to $1-\text{RL}$. If we consider in an “unknown” feedback trial that a hostile target is a signal, the equation for $\beta_{\text{optimal}}$ can be arranged to:
\[ \beta_{\text{optimal}} = V \left( 1 - \frac{RL}{RL} \right) = V \left( \frac{1}{RL} - 1 \right) \]  

This equation for \( \beta_{\text{optimal}} \) was then fit to the \( \beta_{\text{actual}} \) values for each combination of display type and proximity which allowed \( R^2 \) to be calculated. \( R^2 \) was calculated by calculating the total sum of squares (SST), the error sum of squares (SSE) as follows:

\[ \text{SST} = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 \]  

\[ \text{SSE} = \sum_i \sum_j (y_{ij} - \bar{y}_i - \hat{y}_{ij} + \bar{y}_j)^2 \]

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:

\[ R^2 = 1 - \frac{\text{SSE}}{\text{SST}} \]

Fits with a larger \( R^2 \) value indicate the \( \beta_{\text{optimal}} \) equation accounted for a larger proportion of variance in \( \beta_{\text{actual}} \) in each condition. Because this was 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).

### 3.2.5.4. Order Effects

C, d’ and average kill time were calculated for each block of trials for each participant. To test for order effects, a 2 (display method: pie, mesh) x 8 (blocks number) mixed ANOVA was performed on C, d’ and average kill time with display method as the between subjects factor and block number as the within subjects factor. While counterbalancing attempted to control for order effects related to display proximity, analysis was carried out to ensure there was no asymmetric transfer of skills between the integrated and separated interface. Each block was coded as to whether it was preceded by an integrated or separated block. A 2 (preceding block: integrated, separated) between subjects ANOVA was performed on C, d’ and average kill time.
The first block of each session was omitted from the analysis since no other block preceded it. Because of the way the blocks were counterbalanced, each block could be preceded by a block of like or different display proximity; therefore, each block was coded by its relationship to the previous block. This allowed for four different two-block combinations: I-S, S-I, I-I and S-S. Again the first block of each session was omitted from the analysis and a 4 (block combination: I-S, S-I, I-I and S-S) ANOVA was performed on C, d’ and average kill time.

3.3. Results

3.3.1. Order Effects

There was no effect of either the preceding block or block combination on d’, C or average kill time, F>1, indicating there was likely no asymmetric transfer of skill between the integrated and separated displays. There was a main effect of block order on d’, F(7,182)=2.25, p<0.05, (Figure 11).

Figure 11: Effect of block order on sensitivity.

Figure 11 demonstrates that d’ was lower in the first block of trials, indicating the participants were still learning the identification task. More training should be provided in subsequent experiments to ensure learning is completed before the experimental session begins.
There was also significant effect of block order on average kill time, $F(7,182)=4.28, p<0.01$, Figure 12: Effect of block order on average kill time. Participants took longer to kill the target on the first block of each session (Block 1 and 5), with kill time decreasing as the session progressed. Kill time includes the time to identify the target and the time to manually aim and kill the target. Identification skills remained fairly constant after the first block; therefore the decrease was likely mostly driven by experience with the manual-aiming task. The increased average kill time in block 5 is attributable to participants familiarizing themselves with the aiming task at the start of the second session as there was no corresponding decrease in $d'$ for that block.

3.3.2. Miss (Miss Hostile Targets)

There was a main effect of display type on miss rate, $F(1, 26) = 4.31, p < 0.05, r = 0.38$. Participants missed more hostile targets $M = 0.15$, $MS_E=0.0008$ with the pie display than those using the mesh display, $M = 0.12$, $MS_E = 0.0014$. The main effect of reliability level was also significant, $F(1.54, 40.2) = 24.4, p < 0.001$, with miss rate decreasing as reliability level
increased ($\beta_1=-0.05$, $t_{278}=10.81$, $p<0.001$, $r=0.54$) (Figure 13). The main effect of display proximity was not significant, nor was any interaction, $p > .05$ in each case.

3.3.3. False Alarm (Friendly Fire)

The main effect of reliability level on false alarm rate was significant, $F(2.20, 56.7) = 21.3$, $p < 0.001$, such that the FA rate increased with the reliability level ($\beta_1=0.06$, $t(278)=5.87$, $p<0.001$, $r=0.33$) (Figure 14). That is, when the display indicated an increased probability of a terrorist when there was “unknown” feedback, participants killed more friendly soldiers.
3.3.4. Time to Kill Target

Reliability level had a significant effect on kill time, $F(4, 104) = 10.9, p < 0.001$, with kill time generally decreasing as the reliability level increased ($\beta_1 = -0.08, t(278) = 3.03, p < 0.01, r = 0.18$) (Figure 15). The main effects of display proximity and display type were not significant, nor were the interactions, $p > .05$ in each case.

![Figure 15: Main Effect of Reliability Level on Kill Time](image)

3.3.5. Sensitivity: $d'$

Six participants produced zero hits or FAs for some reliability levels. SDT measures cannot be calculated in such cases, although Macmillan and Creelman (1991) have suggested corrections. However, at high reliability levels, the corrections produced unstable SDT parameter estimates (i.e., small changes in the correction lead to large changes in $d'$ and $\beta$) and therefore were not used. Thus, data from the 22 other participants were used for sensitivity analysis.

The effect of display type on sensitivity was significant, $F(1,22) = 6.62, p < 0.05, r = 0.48$. Participants using the mesh display were better able to distinguish hostile from friendly
targets (see Figure 16, which also shows the consistency of this effect across reliability levels).

No other main effect or interaction was significant, $p > .05$ in each case.

![Figure 16: Main effect of display type on sensitivity, d'.](image)

3.3.6. Decision Criterion: C

There was a significant effect of reliability level on the decision criterion, $F(4, 88)=21.6$, $p<0.001$, such that participants were more liberal at higher levels of reliability. This is to be expected with the increased probability of a signal. There was a significant interaction between display proximity and reliability level, $F(4, 88)=8.27$, $p<0.001$ (Figure 17). When the participants used the integrated format, they were more conservative at the low reliability levels and more liberal at the high reliability levels than with the separated display with slopes of $\beta_1=-0.23$, $t(110)=6.96$, $p<0.001$, $r=0.55$ and $\beta_1=-0.15$, $t(110)=4.17$, $p<0.001$, $r=0.37$ respectively. It appears that the separated display led to a more ‘sluggish’ change in decision criterion.
The three-way interaction between display proximity, display type and reliability level was also significant, $F(1, 88)=4.00, p<0.05, r=0.21$ (Figure 18) so this interaction will be explored in more detail using the method described by Aiken and West (1991) to compare slope differences. The integrated-pie display had a significantly steeper negative slope than the separated-mesh display $t(264)=-2.69, p<0.01, r=0.16$. All other comparisons were not significant $p>0.10$. 

**Figure 17: Interaction Effect of Display proximity and Reliability Level on C.**
Figure 18: The three-way interaction effect of display format, display type and reliability level on decision criterion C.
3.3.7. Appropriateness of Reliance

When participants’ decision criterion was compared to an optimal decision maker the $\beta_{optimal}$ equation was a better representation of the participants’ actual $\beta$ values for the displays in a integrated display as they had positive $R^2$ values (Table 2). When participants used the display in a separated display they relied on the aid less optimally, especially the participants using the mesh display (Figure 19). How display type affected participants’ is less clear. Participants using the pie display seemed to become more conservative than optimal as reliability level increased in both the integrated and separated display. Participants using the mesh display in the integrated condition had the most appropriate reliance; however in the separated format, participants’ reliance did not vary with changing reliability level.

<table>
<thead>
<tr>
<th>Display</th>
<th>$SS_{Error}$</th>
<th>$SS_{Total}$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integrated-Mesh</td>
<td>8.08</td>
<td>15.5</td>
<td><strong>0.48</strong></td>
</tr>
<tr>
<td>Integrated-Pie</td>
<td>11.4</td>
<td>13.7</td>
<td><strong>0.17</strong></td>
</tr>
<tr>
<td>Separated-Pie</td>
<td>17.8</td>
<td>16.9</td>
<td>-0.05</td>
</tr>
<tr>
<td>Separated-Mesh</td>
<td>10.6</td>
<td>6.36</td>
<td>-0.66</td>
</tr>
</tbody>
</table>

Table 2. $R^2$ values for each display combination when the $\beta_{optimal}$ equation was fitted to the collected $\beta$. 
Figure 19: Comparison of $\beta_{\text{actual}}$ to $\beta_{\text{optimal}}$ for each display method and proximity combination.
3.3.8. Subjective Measures of Trust

The trust ratings in the “unknown” and “friend” feedback were not normally distributed with a right skew and left skew, respectively. Trust in the fallible “unknown” feedback was generally low and trust in the reliable “friend” feedback was high. A logarithmic transformation and a power transformation were performed correspondingly to the “unknown” and “friend” feedback distributions.

Separate analysis conducted on trust in the whole aid, “unknown” feedback and “friend” feedback showed display type did not have a significant effect on participants’ trust in the whole aid, “friend” feedback or “unknown” feedback, $p > 0.05$ in each case. Display proximity had a significant effect on trust in the “unknown” feedback $F(1, 26) = 12.03, p < 0.05, r = 0.56$. The logarithmic means were: Integrated: $M = 0.58$; Separated: $M = 0.55$, $(\text{MS}_E=0.007)$, which translate into means of 3.8 on a scale of 7 for the integrated display and 3.6 on a scale of 7 for the separated display. Therefore, participants trusted the “unknown” feedback slightly more when shown the integrated display. Display proximity did not have a significant effect on trust in the whole aid or “friend” feedback and no interactions were significant, $p > 0.05$ in each case.

3.4. Discussion

The form of a human-machine interface can affect CID performance and users’ reliance on an automated aid. Participants using the mesh display were more sensitive in discriminating hostile targets from friendly soldiers than participants using the pie display. The effect on sensitivity was unexpected as it was hypothesized that displaying probability information would affect response bias. Display proximity did affect response bias: participants using the integrated display relied on the aid more appropriately than when using a separated display, shifting their criterion with changes in reliability level more optimally.
3.4.1. Effects on Sensitivity

A decrease in sensitivity may be attributable to a salient or distracting display. Suppose that the task is considered 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 target (characteristics of the target’s uniform and weapon). The observer has a limited amount of time to sample the scene, determine target identity, manually aim and fire the virtual weapon, and kill a moving hostile target. The display and target thus create two information channels that compete for the participant’s attention. Because kill times were not different for the method or proximity factors in the present study, the difference in sensitivity might be explained by the shooter using the mesh display having more time to examine the target, thereby improving identification. The pie group may have spent more time sampling the display information, leaving them less time to sample the target. Consistent with this hypothesis, the group using the pie display not only had lower sensitivity than the mesh display group but also had a higher miss rate with no corresponding decrease in FA rate.

The results would appear to be consistent with previously observed performance advantages of mesh-like displays compared to pie charts. As noted in Section 2.2.2.4, more precise quantitative representations are not always beneficial. Finger and Bisantz (2002) found no difference in target identification performance between a degraded icon and numeric or linguistic indicators. Indeed, there appear to be advantages for mesh-like displays when greater precision is not required (e.g., deciding which of two probabilities is larger, Feldman-Stewart et al, 2007). Similarly, an advantage was shown for a “less precise” mesh display over a pie chart, a display that is generally considered an effective display for proportional data.

A possible reason for sub-optimal sampling or attention allocation for the pie display may be that the participants attempted to capitalize on the extra precision afforded by the pie display, when more time should have been spent sampling the target scene. Individuals can be quite
precise at reading pie charts, especially when landmarks (such as the tick marks in the pie display) are provided (Hollands & Dyre, 2000). Based on previous empirical studies it is reasonable to assume participants should have be able to estimate the percentages from the pie chart in the present study with absolute error of at least +/- 2-3 percent (Hollands & Spence, 1998). It is unlikely that participants counted the number of filled-in boxes in the mesh display to obtain a precise number; however, they could still get a sufficiently accurate reading to perform the task. Instead, the case could be made that the less ‘precise’ mesh display afforded a more optimal strategy in a time-pressured situation providing an ‘at a glance’ indication of uncertainty.

Finger and Bisantz (2002), showed that participants were more accurate with a degraded indicator alone than with combined degraded and digital indicators; that is, participants were less accurate when a more precise indicator was added to the display. A similar pattern of results was shown with a gambling task where individuals performed as well or better with fuzzy linguistic descriptors (e.g., likely, very unlikely, improbable etc.) than with precise numeric indicators (e.g., 5 percent) (Budescu & Wallsten, 1990). A possible way to test this theory is to experiment with the segmented, analog display (Figure 7), which may have reduced this effect on sensitivity, as it did not afford as high of level of precision as the pie display.

3.4.2. Effects on Response Bias

Participants were particularly insensitive to changes in reliability level when they were shown the separated display, failing to shift their decision criterion appropriately. They instead maintained a decision criterion that was slightly more liberal than optimal for a reliability level of 70 percent, regardless of the displayed reliability level (which, it is worth noting, was the mean reliability level across trials). Separating reliability and identity information likely increased the number of visual locations from which to sample. In a time-pressured situation, participants may have chosen to sample the target and the identity information more often, while
mostly ignoring the reliability information. Participants used the identity information because they did not fire on the “friend” feedback trials. The identity information was of particular value to the participant because given “friend” feedback the identity of the target was certain. 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. Again, there was no difference in kill time between the conditions, so more time spent sampling one channel meant less time to sample another. The participants also trusted the separated “unknown” feedback less. This is consistent with the findings of Wang et al. (in press) who found that when reliability levels were not disclosed to participants, the participants had less trust in the aid.

3.4.3. Other Effects

The reliability level also affected the time it took to kill the target where the participants killed the target more quickly at higher reliability levels. The increased certainty may have allowed the participants to make the decision to kill the target more quickly. The display provided no information when it displayed a reliability level of 50%, therefore the participant would have to rely on their identification alone. At high reliability levels, the participants could still perform quite well if they adopted a simple, time-efficient strategy of killing all targets at the 90% reliability level.

3.4.4. Experiment I Conclusion

The results of the study provide evidence that the form of a display can affect reliance on an automated aid as well the ability to distinguish friends from unknown targets. Displaying probability information was expected to affect reliance behavior but not sensitivity. The observed sensitivity differences may have been caused by different attention allocation strategies used between the difference displays; however, empirical evaluation of this hypothesis is required.
4. Attention and CID

Attention plays an important role in CID. For effective CID, the soldier must detect and locate a potential target, and then identify the target using a number of possible cues. CID uses cognitive processes such as visual search, orientation and object recognition, which all relate to the use of attention. When automation is used to assist in CID, attention must also be allocated to the HMI, which as shown in Experiment I, the form of the HMI can affect reliance on the automated system and sensitivity in for the CID task.

Attention in the context of CID can be understood by considering the processes of divided and selective attention as well as how individuals shift their attention. Divided attention involves the human allocating attention between two more stimuli in order to process information from the multiple sources (Matlin, 2005). In the chaotic battlefield environment, the soldier must divide attention among numerous stimuli in order to complete the mission. In the aided CID task, the user divides attention between the target to be identified and the rifle mounted display in order to gain information. On the other hand, selective attention describes the ability to focus attention on relevant cues while filtering out irrelevant cues (Matlin, 2005; Wickens & Hollands, 2000). Selective attention is an important component of object recognition, which is important for CID (Graboi & Lisman, 2003). Important to both divided and selective attention is how individuals shift attention to sample multiple channels for divided attention or to change the location of focused attention for selective attention. An overview of attention shifting will be provided, followed by a more in-depth examination of issues of attention specific to interface design and CID.

4.1. Orienting and Shifting Attention

There are two major dichotomies to consider when discussing the orienting of attention. First, individuals can trigger an attention shift endogenously (i.e., the trigger comes from
motivations within the individual) or exogenously (i.e., the trigger comes from the environment outside of the individual) (Klein & Shore 2000). The two types of triggers correspond to the cognitive constructs of top-down versus bottom-up processing respectively. Exogenous orienting is often thought of as reflexive in nature, with the individuals responding to a salient or novel stimulus in the environment even when the stimulus is irrelevant to the present task (Funes, Lupianez, & Milliken, 2005); whereas endogenous orienting is thought to be the strategic direction of attention to relevant areas of space to perform the desired task. The second dichotomy is between overt and covert shifts of attention. Eye movements accompany overt shifts of attention whereas covert shifts of attention occur without the eyes moving.

The modes of orienting attention relate to each other and other aspects of attention in numerous ways (for more in-depth reviews refer to Klein, 2004; Klein & Shore, 2000; Yantis, 1998). In order to narrow the scope, the aspects of orienting attention that are relevant to the present work should be considered. Much of the work regarding exogenous versus endogenous triggers examines the effect of exogenous stimuli appearing and capturing attention over competing endogenous cues. While exogenous cues are important for CID given the unpredictable battlefield environment (e.g., the detection of a potential threat), the scope of the present work assumes the target has been detected and attention has been captured but the target has not yet been identified. The tasks the soldier must perform are object recognition and identification. These are generally driven by an interplay of endogenous and exogenous cues (Graboi & Lisman, 2003), with the exogenous cues informing the soldier where next to endogenously shift attention to gain more information, a concept that will be discussed in much greater detail in the following sections.

The degree to which overt and covert shifts of attention differ is a point of debate. As all saccades are preceded by a shift of attention to the intended location of the saccade (Shepard, Findlay & Hockey, 1986) it has been hypothesized that all shifts of attention may be
accomplished by motor preparation to fixate on the attended location; however, the motor response is suppressed in the case of covert orienting (Klein, 1980 & Rizzolatti et al, 1987). This hypothesis is contradicted by evidence showing that sometimes when a saccade is prepared, but not executed, there was no corresponding attention shift and visa versa (Klein & Pontefract, 1994). However it is certain that a motor program that is then suppressed by an inhibitory mechanism accompanies most covert shifts of attention (Rafal et al, 2000).

Shifts of attention can occur every 20-30 milliseconds (Graboi & Lisman, 2003). For overt shifts of attention, i.e., saccades, a number of parameters are used to describe the fixations. Spatially the fixation can be defined by the center point well as the useful field of view -- the diameter around the center point from which information can be extracted (Wickens & Hollands, 2000). Temporally the fixation can be defined by the dwell time, which is how long the eye remains at a particular location. The saccadic shifts can happen rapidly; however, for information to be extracted the dwell time needs to be at least 200-300 milliseconds, which increases with the complexity and amount of the information to be acquired during the fixation (Moray, 1986). These concepts are important when considering how attention is divided among multiple stimuli.

### 4.2. Object Recognition and Scene Comprehension

CID requires that the soldier integrate a number of cues to identify individuals in the battlefield. Some cues can be considered bottom-up (e.g., what is the individual wearing, their location in the battlefield, etc.), while others are top-down (e.g., prior intelligence briefings). The soldier must integrate numerous top-down and bottom-up cues. These cues can influence how he or she perceives and then identifies a target as well as his or her situational awareness. Attention can influence this integration and facilitate recognition if relevant cues are attended to; however, a complex visual environment or a high mental workload can lead to distraction and errors on the recognition task. As effective CID requires the soldier to use information from both the
environment (including the display) and the target, the following section will examine scene
comprehension and object recognition together. While often considered separate topics, the two
are clearly related as scene comprehension requires one to know the identity of objects in the
scene, and context information from the scene can act as cue for object identification (e.g., if one
is viewing a computer workspace, one can reasonably expect a rectangular object on the desk is a
keyboard.)

When an observer is first presented with a scene, attention guides where the observer is to
focus to interpret the scene or locate and recognize relevant objects in it. The observer likely
creates a salience map of the scene, and then attends to locations in order of decreasing salience
to gain information (Navalpakkam & Itti, 2005). Salient locations provide exogenous, bottom-up
cues to where important information can be obtained; however, even this initial ‘gist’ of the
scene is influenced by top-down information. Depending on the task, different features such as
orientation, size, closure, hue, intensity or direction of motion may be considered salient features
to form the map (Navalpakkam & Itti, 2005). This concept echoes the classic work of Yarbis
(1967), which showed that task affected eye movements across a scene. These overt shifts of
attention may depend on different feature salience maps to guide how information can be
garnered effectively for the different tasks.

Object recognition consists of a series of bottom-up and top-down iterations where an
individual systematically narrows down the possible identity of an object by analyzing a number
of features or components (Matlin, 2005). For example, the letter H may be made up of three
features; two vertical lines and one shorter horizontal line that are in a particular relation to one
another. Extending a similar theory to three-dimensional objects, geons are generic three
dimensional shape components that comprise objects (Matlin, 2005).

A major issue with many feature recognition theories is that they only consider shape, in
either two or three-dimensional form, for providing information for object recognition. While
shape is a distinctive property of all objects, cues like hue and motion can provide additional information, especially in complex situations. For example, if an individual observed a small red orb, both the colour and the shape of the object might give an indication that the object is an apple. In the complex task of combat identification, a friend or enemy may appear to have relatively similar shapes, and identification may rely on colour and other more subtle cues. Ullman, (2006), suggests that features do not have to be generic, instead they can be class specific and pictorial which allows Ullman’s theory to consider object recognition cues beyond shape. The features, while less generic, do not have to belong to only one class or subclass of objects; instead the utility of a feature for object recognition is judged based on how much the feature reduces uncertainty. A feature with low redundancy between classes has higher entropy. This information concept can be expanded to combinations of features (e.g., the likelihood of two particular features being found together).

Feature recognition theories suggest attention plays a role in selecting what cues, components or areas of space to retrieve information for identification. As mentioned above, when presented with an object or scene, a salience map (a bottom-up cue) will direct an observer where to attend in order of importance. Once information is taken from that location, the observer combines that information with prior knowledge to decide where attention should be directed next by recalculating, in a top-down manner, a new feature probability landscape (Graboi & Lisman, 2003). To identify an object efficiently, attention should be directed to a feature whose presence or absence would most reduce uncertainty regarding the object’s identity. For the class specific features proposed by Ullman (2006), an observer uses top-down resources to bind features together and determine which features have the least redundancy. The features that have the least redundancy can provide the most information about object identity. Attention is then directed to the informative feature’s expected location in space.
The iterations between top-down and bottom-up cues occur rapidly, allowing the observer to effectively perceive and interact with his or her environment. While using class specific features allows observers to consider cues other than shape for recognition, the number of features may be considered unwieldy to process thereby making the use of class specific features an inefficient means for object identification. The two views of generic versus class-specific features can be reconciled by considering that people may use both types of features in object recognition. Low-level information processing may use generic, shape-like, features to determine object class where class can be as generic or specific as need be for the level of processing required (e.g., inanimate object, plant, food, fruit). If higher order processing is required within a class, then object class can act as a cue to truncate class specific features to use for higher-order identifications. These theories of object recognition can be applied more specifically to the task of CID in the presence of an HMI to further inform experimentation.

4.3. Attention Model for CID

A second experiment was necessary to investigate the attention allocation hypotheses proposed based on the results of Experiment I. To clarify the purpose and hypotheses for Experiment II a descriptive attention model for the object and scene identification task was created based on information processing model (e.g., Wickens & Hollands, 2000).

Wicken’s information processing model was chosen as it provides discrete steps of processing. The model fit well with other, more specific models (discussed in Section 4.2) of how attention plays a role in object recognition and scene comprehension. While having discrete steps is a useful explanatory tool, much of the recent work in cognitive psychology has focused on neural networks where cognitive processes occur in parallel (Matlin, 2005). While a detailed explanation is beyond the scope of the current project, most models state that a certain level of activation must be achieved before a particular outcome (e.g., an action or a decision) is
executed. Different areas and neurons in the brain work in an interconnected manner to achieve the needed level of activation using both inhibitory and excitatory influences. The inhibitory influences are necessary to suppress competing outcomes. These concepts have been merged with the information processing model, to provide a complete discussion of attention and CID.

4.3.1. **Attention Model of Object Recognition/Identification**

![Figure 20: Model of object recognition](image)

In Figure 20 an information processing model is used to summarize the object recognition models discussed in Section 4.2, The model demonstrates the iterations that occur during scene comprehension as the individual scans for cues in a strategic way. The search for bottom-up cues is informed by top-down influences in the form of salience maps and feature probability landscapes. In the case of object recognition, each cue may excite or inhibit possible object identities, which leads to the recalculation the feature probability landscape. Iterations occur until sufficient excitatory neural evidence has been accumulated for one possible object identity.
4.3.2. ‘Bottom-up Cues’ for CID

To further examine the role of interface design in CID, the role of the bottom-up cues are
examined more in depth in light of the results from Experiment I. In Error! Reference source
not found., and Figure 20 limited attention resources are shown as being equally allocated among
the components; however, this is likely not the case. For optimal performance attention should be
allocated to the cues that will provide the most information thereby reducing uncertainty about
the situation. Section 3.4 proposed that the difference in sensitivity between the pie and mesh
displays observed in Experiment I may have been due to a difference in the attention allocation
strategies of the participants. Attention allocation strategies may also explain the difference
between the integrated and separated displays. By again narrowing the scope of the model,
attention allocation to the bottom-up cues of the target and interface can be examined and then
tested further.

Figure 21: Attention Allocation- Separated display

Figure 21 shows a proposed attention allocation model for the separated display. To
comprehend the scene for the purpose of target identification the participant has three channels
of information to sample: the target, the identification indicator light and the reliability graphical
form. The identification indicator light can eliminate uncertainty if “friend” feedback is received therefore making it an important source to sample. The participants had almost perfect reliance on the blue light trials indicating that they likely prioritized an amount of attention resources for this channel of information. If “unknown” feedback is received the participant must strategize where to sample the scene to perform an efficient CID. The target and the reliability graphical form likely compete for attention resources. As mentioned above the target contains the veridical identity of the target, while the graphical form conveys a level of uncertainty. The participants using the separated display could have prioritized attention resources on the target rather than on the reliability graphical form as they relied on the aid less optimally than with the integrated interface.

As can be seen in Figure 22, the integrated interface allows the participants to sample the reliability information while they were obtaining the identification information.

![Diagram](image)

**Figure 22: Attention Allocation - Integrated-Pie display**

If “unknown” feedback was received the participant could focus attention resources on the target having already gained reliability information from the same form as the identification information. There was no sensitivity difference between the integrated and separated display,
which may indicate the participants expended similar amounts of attention resources on the targets; however, for the participants using the separated display this was at the cost of the gaining the reliability information.

Attention allocation can also be used to explain the sensitivity difference between the pie and mesh displays. If more attention resources are allocated to the graphical form instead of the target, sensitivity may decrease (Figure 22 and Figure 23). However, it is more difficult to explain why attention resources may be allocated to one display type over the other. One suggestion made in Section 3.4 was that a general level of uncertainty could be more quickly gained from the mesh display. This may have allowed the participant to use the reliability information more efficiently and expend more resources attending to information contained in the target. Conversely, the pie display may have afforded a more precise reading, leading to the seemingly suboptimal strategy of expending more resources on gaining a more precise reading of the reliability when such precision was not required for effective CID.

Figure 23: Attention Allocation - Integrated-Mesh display
5. Experiment II

5.1. Introduction

The purpose of Experiment II was to test the attention allocation explanation given for the results in Experiment I by reducing the time available to identify a target in a simulated combat scene. If the model is plausible, reducing stimulus exposure time should impact performance on the CID task and reliance on the CID aid. Based on the models provided in the previous section, reducing the time available to view the CID scene will further stress attention resources since attention is serial in nature, exacerbating the effects shown in the previous study.

5.1.1. Hypotheses

If participants are able to more efficiently acquire information (i.e., use less attention resources) with the mesh display than the pie display, there should be less of an effect on sensitivity (d’) for the mesh display as viewing time decreases as the participant will have more time and/or attention resources available to examine the target. Similarly, with a shorter viewing time, participants using the separated display may rely on aid less optimally or show a decrease in sensitivity as the participants will have less time to attend to the three channels of information provided for identification rather than the two channels for the integrated display.

5.2. Methods

5.2.1. Experimental Design

A 2 (display type: pie & mesh) x 2 (display proximity: integrated & separated) x 2 (stimulus duration: 400ms & 800ms) x 5 (reliability level: 50%, 60%, 70%, 80% & 90%) mixed design was used. Display proximity, stimulus duration and reliability level were within subjects factors, while display type was a between subjects factor. In other words, each participant
experienced one of the display types in both an integrated and separated display at five different
dependency levels for 400ms and 800ms.

The participants completed 1680 trials, separated into eight blocks of 210 trials each. Similar to Experiment I, the participants were presented with a separated HMI in four blocks and an integrated HMI in the other four blocks. For four blocks the combat scene remained on the screen for a duration of 400ms and for the other four blocks, 800ms. The stimulus duration variable was blocked, so participants would know how long they had to identify the target and strategize appropriately. If stimulus duration varied randomly, participants could not determine how long they had to sample the scene and plan accordingly; therefore the participants may have planned for the worse case scenario and always sampled as if they only had 400ms to view the scene. The stimulus durations were equally distributed between the integrated and separated HMI so participants saw each combination of display proximity and stimulus duration twice. The participants completed the eight blocks of trials over two sessions, each two hours long, separated by at least an hour to reduce fatigue. Each session consisted of four blocks of trials. Similar to Experiment 1 within a session the blocks could be ordered IISS, SSII, SISI, ISIS, SIIS or ISSI (I=integrated, S=separated), however the addition of stimulus duration doubled the ways the blocks could be ordered (e.g., for the first sequence IISS stimulus duration could be distributed as ‘I-400ms, I-800ms, S-400ms, S-800ms’ and ‘I-800ms, I-400ms, S-800ms, S-400ms’). The 12 block orders were counterbalanced across participants. For each block, the targets were “friendly” for half the trials and “hostile” for the other half, presented in random order. The reliability levels varied randomly trial-to-trial within a block. The reliability level referred to the probability that the target was a hostile soldier on a given “unknown” feedback trial.


5.2.2. Participants

Twenty-six participants with normal or corrected to normal vision volunteered for the study (Males: n=7, 25.1 +/- 6.0 years of age) with data collected at Defence Research and Development Canada-Toronto. Complete data was collected from 24 participants and used for analysis. Participants were paid 22.20 CAD if they were employees of Defence Research and Development Canada or the Department of National Defence. Otherwise participants were paid 45.20 CAD. Similar to Experiment I, performers above the 66th percentile received a 10 CAD bonus.

5.2.3. Apparatus

Custom software was used to display the combat identification scenes. Modifications to the lighting of the IMMERSVE map were made to create a still stimulus set that was sufficiently difficult to prevent a ceiling effect on the identification task. The targets in the scene were positioned either facing the participant or facing to the right and left. The different orientations were used to increase the stimulus variability and to create a more realistic object recognition task. To ensure the task was solely object identification the target always appeared in the same location so the task was not confounded by the effect of visual search times. Informal pilot testing was carried out to ensure the participants could discriminate between the friendly and hostile targets at a rate greater than chance but less than 100% to avoid a ceiling effect.

5.2.4. Procedure

Similar to Experiment I participants provided informed consent (Appendix F) and had their visual acuity measured with a Snellen eye chart. The participants were given a sheet of instructions about the experiment procedure to read which was followed by a questionnaire to demonstrate their comprehension of the instructions. The instructions were identical to those
used in Experiment I except for stating the participant would view a still combat scene for either 400ms or 800ms and information regarding the actual shooting of the target was removed.

The participants performed two training sessions of 100 trials, one with the integrated display and one with the separated display. The goal of the training sessions was to train participants to identify the targets at a rate better than chance (~70%). For the first ten trials of the first training session the scene remained on the screen for 15 seconds. This was so the participant had sufficient time to examine the target’s appearance and relate the appearance to the target’s identity. After the first ten trials, all other trials were 600 milliseconds during the training session. For trials 10-20, the experimenter informed the participant of the target identity before presentation of the scene so the participant could become accustomed to the rapid nature of the task. Following trial 20, the participant was asked to identify the target himself or herself and the experimenter provided feedback. Once participants were able to identify targets correctly approximately 70% of the time, the experimenter let them run through the rest of the training session without feedback. If the participants had difficulty identifying the targets on the faster trials, the session was restarted so the participant could view the slower trials again.

For the second training session all trials were 600ms in duration. The experimenter again provided the identities of the targets before the scene was displayed for the first ten trials so the participant could become familiar with the new display proximity (i.e., the display proximity not used in the first training session). To ensure participants were still able to identify the targets at a rate better than chance, the experimenter gave feedback for the next 10-30 trials as needed. Early in the session (presumably with less fatigue), and with feedback, most participants were performing with an accuracy of close to 80-90%. Only one participant was excluded on the basis that she could not identify the targets at a level greater than chance.

Each trial started with a gray screen with black text in the center (Figure 23) instructing the participants to ‘Press any key for next scene’.
Figure 24: Screen view sequence for one trial.

The text was situated to provide a fixation point on the screen of where the target was to appear. The key press to start the trial was added so the participants could self-pace the trials. Self-pacing allowed for a break mid-block if the participant felt they were fatiguing or their vigilance was decreasing. Following the key press, the IMMERSIVE scene was displayed for either 400ms or 800ms depending the stimulus duration condition. Following scene presentation, another gray screen with black text appeared asking the participant ‘Do you shoot the target?’ The participant used the mouse to select ‘Yes’ or ‘No’. Once they made a selection, another button appeared that said ‘Go’. When the participants used the mouse to press the ‘Go’ button, their answer was entered, the trial ended and the first screen appeared.
5.2.5. Data Analysis

The analyses performed were similar to those performed for Experiment I (Section 3.2.5). The only differences were to account for the addition of stimulus duration into the analysis. Therefore, a 2 (display type: pie, mesh) x 2 (display proximity: integrated, separated) x 2 (stimulus duration: 400ms, 800ms) x 5 (reliability levels: 50%, 60%, 70%, 80% & 90%) mixed ANOVA was performed on the transformed miss rates, transformed FA rates, d’ and C. Display proximity, stimulus duration and reliability level were within subjects factors and display type was a between subjects factor. When a regression model was used to explore interaction effects, the 400ms condition was assigned a value of -1, and the 800ms condition a value of 1. Display proximity did not seem to produce asymmetric transfer of skills in Experiment I; thus, analysis of order effects focused on stimulus duration. Each block was coded as to whether the block preceding it used a 400 or 800 millisecond stimulus duration. The four block combinations (400-400, 800-800, 400-800 and 800-400) were also coded. Analysis then proceeded as described in Section 3.2.5.4 of Experiment I.

5.3. Results

5.3.1. Order Effects

There were no significant effects of the preceding block on either d’ or C. There was a significant effect of block combination on d’, F(3,140)=5.21, \( p<0.05 \) (Figure 25). Simple contrasts indicated that d’ was significant higher when the combination was 400-400, when compared to the combinations of 800-400 and 800-800 \( p<0.05 \). The combinations, 400-400 and 800-800 only occurred eight times (i.e., twice for four different participants) whereas the combinations 400-800 and 800-400 occurred 64 times. The four participants that received the combination 800-800 were among the participants that had the lowest overall sensitivity. The decrease in sensitivity for the 800-400 combination may have been due to the participants having
to adjust their identification strategies from the easier to the more difficult stimulus duration condition.

Figure 25: Effect of block combination on $d'$. There was a significant effect of block order on $d'$, $F(7,154)=2.79$, $p<0.05$, (Figure 26).

Sensitivity decreased throughout each session (session 1: blocks 1-4, session 2: blocks 5-8). This vigilance decrement was expected. In future studies with blocks of a similar duration, a longer
break should be given after two blocks of trials instead of four. Block order did not significantly affect C.

5.3.2. Miss Rate (Miss Hostile Targets)

There was a significant interaction between stimulus duration and display proximity on the transformed miss rate $F(1, 22) = 5.77, p < 0.05$ (Figure 27). Participants using the separated display missed more hostile targets when the stimulus was displayed for the shorter period of time but not for the longer period.

There was also a significant interaction between display proximity and reliability level on the transformed miss rate $F(4.1, 46.0) = 5.86, p < 0.01$ (Figure 28) where there was a more pronounced decline in miss rate over increasing reliability levels for the integrated interface than for the separated interface. Display proximity was regressed over reliability level to provide simple slopes of $\beta_i = -0.06$, $t(118) = 6.04, p < 0.001$, $r = 0.49$, for the integrated display and $\beta_i = -0.02$, $t(118) = 1.49, p > 0.10$, $r = 0.14$ for the separated display.

Figure 27: Interaction effect of Stimulus Duration and Display proximity on Miss Rate

There was a significant interaction between stimulus duration and display proximity on the transformed miss rate $F(1, 22) = 5.77, p < 0.05$ (Figure 27). Participants using the separated display missed more hostile targets when the stimulus was displayed for the shorter period of time but not for the longer period.

There was also a significant interaction between display proximity and reliability level on the transformed miss rate $F(4.1, 46.0) = 5.86, p < 0.01$ (Figure 28) where there was a more pronounced decline in miss rate over increasing reliability levels for the integrated interface than for the separated interface. Display proximity was regressed over reliability level to provide simple slopes of $\beta_i = -0.06$, $t(118) = 6.04, p < 0.001$, $r = 0.49$, for the integrated display and $\beta_i = -0.02$, $t(118) = 1.49, p > 0.10$, $r = 0.14$ for the separated display.
Figure 28: Interaction Effect of Reliability Level and Display proximity on Miss Rate

The decrease in miss rate for increasing reliability level for both the separated and integrated display also manifested in a main effect of reliability level $F(1.4, 31.0)=5.44$, $p<0.001$. No other main effects were significant.

### 5.3.3. False Alarm Rate (Friendly Fire):

The main effects of display type and display proximity were not significant, $F>1$. The main effect of stimulus duration was significant, $F(1,22)=7.76$, $p<0.05$, $r=0.51$, where participants killed more friendly soldiers in the shorter (400ms) condition $M_{\text{untransformed}}=0.57$, than in the longer (800ms) condition, $M_{\text{untransformed}}=0.52$ ($MS_E=0.004$).
The interaction of display proximity and reliability level was also significant, F (3.2,70.8)=9.95, \( p < 0.01 \) (Figure 29). The interaction was characterized by FArate increasing with increasing reliability level for the integrated display (\( \beta_i = 0.08, t(118)=5.91, p<0.001, r=0.48 \)) whereas, FArate remained fairly constant over reliability levels for the separated display (\( \beta_i = -0.007, t(118)=0.52, p>0.10, r=0.05 \)). The increase in FArate for the integrated display over reliability level also manifested in a main effect of reliability level F(2.3,50)=3.06, \( p<0.05 \).

5.3.4. Sensitivity – d’

As with the first experiment, six participants had to be excluded from the analysis of the SDT parameters d’ and C as they produced either zero hits or FAs for one of the combination of factors; therefore, 18 sets of complete data were included for these analysis. There were two main effects for the measure d’. First, stimulus duration had a significant effect on sensitivity
$F(1,16)=12.1, p<0.01, r=0.48$, where participants were more sensitive for the longer stimulus duration, $M_{\text{transformed}}=0.75$, than the shorter stimulus duration, $M_{\text{transformed}}=0.55$, ($MSE=0.003$), indicating the magnitude of stimulus duration manipulation was sufficient to affect participants performance. Second, reliability level had a significant effect on sensitivity $F(4,64)=2.71$, $p<0.05$.

The main effect of reliability level on sensitivity was driven by higher order interactions. There was a significant interaction between display type and reliability level $F(4,64)=3.25, p<0.5$ (Figure 30) where sensitivity increased with increasing reliability levels for the pie display but not the mesh display. The simple slope for the mesh display was negligible at $\beta_1=-0.003$, $t(88)=0.58, p>0.10, r=0.09$ and the simple slope for the pie display was positive with a value of $\beta_1=0.10$, $t(88)=2.27, p<0.05, r=0.24$. Participants were less sensitive using the pie display than the mesh display at lower reliability levels; however, as reliability levels increased participants sensitivity increased when using the pie display until $d'$ values reached mesh display $d'$ values.

![Figure 30: Interaction Effect of Reliability Level and Display Type on $d'$.](image-url)
A similar interaction was observed with display proximity and reliability level $F(4,64)=3.38, p<0.05$ (Figure 31). Participants were less sensitive using the separated display than the integrated display at lower reliability levels; however, as reliability levels increased participants’ sensitivity increased when using the separated display until $d'$ values reached or exceeded the integrated display’s $d'$ values. The simple slopes for both displays were positive, but the separated slope was steeper, $\beta_1=0.09$, $t(88)=2.37$, $p<0.05$, $r=0.14$, than the integrated slope, $\beta_1=0.004$, $t(88)=0.11$, $p>0.1$, $r=0.01$.

![Graph showing interaction effect](image)

**Figure 31: Interaction Effect of Reliability Level and Display proximity on $d'$.**

Finally there was a display proximity x stimulus duration x reliability level interaction effect on sensitivity $F(4,64)=2.56$, $p<0.05$ (Figure 32). The simple slopes (Table 3) reveal that sensitivity increased as reliability level increased for the separated display in both time conditions. A similar but less drastic increase was shown for the integrated display in the 400 milliseconds stimulus duration. However, the pattern was reversed in the 800 milliseconds
stimulus duration, with sensitivity decreasing with increasing reliability level for the integrated display (Figure 32). There were no other significant interaction effects on sensitivity.

Figure 32: Three-way interaction of Reliability Level, Display proximity and Stimulus Duration on $d'$
<table>
<thead>
<tr>
<th>Condition</th>
<th>Simple Slope Across RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>400ms-Integrated</td>
<td>0.05</td>
</tr>
<tr>
<td>800ms-Integrated</td>
<td>-0.04</td>
</tr>
<tr>
<td>400ms-Separated</td>
<td>0.08</td>
</tr>
<tr>
<td>800ms-Separated</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Table 3: Simple Slopes of Stimulus Duration and Display proximity regressed on Reliability Level

5.3.5. Decision Criterion-C

Stimulus duration had a significant effect on C, $F(1,16)=6.88$, $p<0.05$, $r=0.54$, where participants were more conservative in the 800 millisecond condition $M_{transformed}=-0.37$, than the shorter stimulus duration, $M_{transformed}=-0.42$, ($MS_E=0.001$).

There was a main effect of reliability level on C, $F(1.5, 25.0)=3.85$, $p<0.05$; however, this effect was driven by the integrated display as demonstrated by the display proximity x reliability level interaction, $F(4,64)=9.60$, $p<0.01$ (Figure 33).

![Figure 33: Interaction Effect of Reliability Level and Display proximity on C.](image-url)
The simple slope for the integrated display was $\beta_i = -0.18$, $t(88)=6.80$, $p<0.001$, $r=0.59$, indicating that participants became more liberal with the integrated display as reliability level increased. The simple slope for the separated display was $\beta_i = 0.001$, $t(88)=0.22$, $p>0.10$, $r=0.02$, indicating that when using the separated display participants did not adjust their decision criterion with reliability level.

The interaction effect of display proximity and reliability level on decision criterion was driven by similar interactions in the error (miss and FA) rates, where the miss rate decreased as reliability level increased and the FA rate increased as reliability level increased for the integrated display.

**5.3.6. Appropriateness of Reliance**

The appropriateness of the participants’ reliance was assessed using a similar method to the method used in Experiment I (Section 3.2.5). The value of a hit was 1.0 as the experimental apparatus allowed the participant to receive a hit each time they identified the target as hostile; therefore, the $\beta_{optimal}$ is reduced to:

$$\beta_{optimal} = \frac{1}{RL} - 1 \quad (10)$$

This equation for $\beta_{optimal}$ was then fit to the $\beta_{actual}$ values for each combination of display type and display which allowed $R^2$ to be calculated as in Experiment I (see Section 3.2.5). (Figure 34 & Table 4).

<table>
<thead>
<tr>
<th>Display</th>
<th>SSE</th>
<th>SST</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integrated-Mesh</td>
<td>16.8</td>
<td>34.6</td>
<td>0.51</td>
</tr>
<tr>
<td>Integrated-Pie</td>
<td>27.7</td>
<td>27.7</td>
<td>0.00</td>
</tr>
<tr>
<td>Separated-Pie</td>
<td>21.7</td>
<td>15.4</td>
<td>-0.41</td>
</tr>
<tr>
<td>Separated-Mesh</td>
<td>18.5</td>
<td>15.10</td>
<td>-0.22</td>
</tr>
</tbody>
</table>

*Table 4: R² values for each display combination when the $\beta_{optimal}$ equation was fitted to the collected $\beta$. 
Figure 34: Comparison of $\beta_{\text{actual}}$ to $\beta_{\text{optimal}}$ for each Display proximity and Type combination.

The $R^2$ values reveal that participants changed their reliance more optimally when using the integrated display than when using the separated display. However, as Figure 34 demonstrates participants in all conditions were more conservative than optimal. When
participants used the separated display, they did not appear to shift their decision criterion as reliability level increased, instead maintaining a constant criterion just below 1.0. Therefore, for reliability levels above 50% the participants were more conservative than optimal. For the displays in the integrated display participants were consistently more conservative throughout all reliability levels as they shifted their decision criterion with changing reliability levels. This indicates that while their strategy was more conservative than optimal, the participants were attuned to the changing reliability of the “unknown” feedback.

5.4. Discussion

Experiment II aimed to provide insight into the results of Experiment I; therefore, it is difficult to separately discuss the results of Experiment II without comparing the protocol to Experiment I. The following is a combined discussion of the two experiments, with a focus on the new information Experiment II provides. First, a summary of the results of Experiment II will be provided followed by a comparison of the protocols for Experiment I and Experiment II. The sensitivity and response bias results will be discussed in detail followed by a limitation section that takes into account limitations of both protocols.

5.4.1. Summary of Experiment II Results

The form of the human-machine interface was again shown to affect CID performance and users’ reliance on the automated aid. The display proximity effects were replicated from Experiment I where participants relied on the aid more appropriately when using the integrated format. Display type was again shown to affect participants’ sensitivity; however, unlike Experiment I, the relationship was also dependent on reliability level. Participants using the mesh display had higher sensitivity than those using the pie display, but only for lower reliability levels. The participants using the mesh display maintained a stable level of sensitivity for all five reliability levels. For the participants using the pie display, sensitivity was lower at low
reliability levels but monotonically increased with increasing reliability levels. Display proximity also interacted with reliability level to have an effect on participants’ sensitivity. The pattern was similar to that of display type, where the separated interface had lower sensitivity than the integrated interface at lower reliability levels but the sensitivity increased with increasing reliability level for the separated display. A comparison between the Experiment I and Experiment II protocols can help explain the results and refine the explanations put forth in the Discussion of Experiment I (Section 3.4).

### 5.4.2. Differences between Experiment I and Experiment II

A stimulus duration variable was added in Experiment II to add a time stress. Because attention is serial in nature, if more attentional resources were required to obtain reliability and identity information from an interface, the participants would have less time to gain information from the target’s appearance as the time available to sample the display decreased. With less time to sample the target there should be a greater decrease in sensitivity. Conversely, as stimulus duration decreased, participants may choose to attend less to the interface, hampering his or her ability to gain reliability information which may lead to less appropriate reliance. This manipulation was only somewhat successful for its intended purpose; the limitation of implementing it will be discussed later in this section. The addition of the stimulus duration variable reduced the amount of time available for the task and necessitated the use of still screen shots for a well-controlled stimulus duration variable. Still screen shots eliminated the interactive nature of the simulator.

One of the less obvious, but more important differences between the two protocols was the nature of the decision required of the participants. In Experiment II the participants’ decision was forced after the scene had been presented for its set, short, duration. Even if the participants could not determine the target’s identity during the stimulus presentation, a decision was still
required. In Experiment I participants had more time to sample the scene and could continue to gain information until they either shot the target, or the target exited the screen (~10 seconds, much longer than the 400 or 800 milliseconds durations used in this experiment). If the participants did not initially perceive the target as hostile and shoot it on sight, the participant could continue to sample the scene for cues until enough evidence was present to suggest the target was hostile allowing for meta-decision, where the participants could decide if they were certain about the identity of the target within the 10-second window.

5.4.3. Effects on Sensitivity

The decrease in stimulus duration was expected to affect sensitivity for the pie display more so than the mesh display. This display type x duration interaction did not occur. However, there was a significant interaction between display type and reliability level on sensitivity. Scanning behaviour may explain the sensitivity difference between reliability levels for the pie display. Participants using the mesh display can sample anywhere on the graphical form (Figure 35).

![Visual Sampling of Mesh Display](image)

**Figure 35. Visual Sampling of Mesh Display**

Individuals using the pie display must scan a larger area of the form to find the two radii that portion the form into levels of uncertainty (Figure 36) adding a visual search component to the task.
For the pie display it would be strategic for the participants to sample the grey area of the form rather than the red area (shown in black in Figure 36). According to Weber’s theory of just noticeable differences (JNDs) (Weber, 1834), the detectable change in a stimulus is proportional to the magnitude of the stimulus; therefore, participants would more readily detect changes on the smaller, grey portion. As shown in Figure 36 at a reliability level of 90% the radii of the pie display are much closer together and therefore may be more readily sampled leaving more time to examine the target as compared with lower reliability levels. The effect is likely small since, due to the size of the display and the visual angle, it is probable that only one or two additional fixations would be required or that only a covert shift of attention would be required to find the radii. However, since visual fixations need to be at least 200ms in length to gain information in the useful field of view, even one additional fixation to find the two radii would be detrimental when the combat scene was only available for 400 or 800ms.

This subtle effect between reliability levels may not have been present in Experiment I since participants had much more time to examine the entire scene thereby negating the effect of a small number of additional fixations. This visual search pattern may have drawn their attention away from the target since the participant had to sample particular area on the pie display and therefore aim for a smaller area on the pie form rather than aim for any area on the mesh form.
This effect can be compared to Fitt’s Law (Fitts, 1954), where movement time decreases with increasing target size. Since the participants could sample anywhere within the mesh display this created a larger target. Participants could make less precise eye movements, reducing movement planning time.

A similar interaction between display proximity and reliability level also affected $d'$ where participants were more sensitive when using the integrated interface at lower reliability levels but had similar sensitivity as when using the separated interface at higher reliability levels. The pattern was not as well defined as the display type x reliability level interaction but may suggest another explanation beyond a visual sampling theory for the dependence of $d'$ on reliability level. Increasing the probability of a signal is expected to affect the participants’ decision criterion, but, in some cases, it can also increase sensitivity. The increased reliability level may have increased the participants’ expectancy of a signal through a reduction in uncertainty, which has been show to increase sensitivity (e.g., Milosevic, 1975). The increased expectancy may lead to more appropriate observing behaviour, especially in loosely coupled displays (i.e., where the participant can orient away from the area of interest) (Warm, Dember, Murphy & Dittmar, 1992). In Experiment I, kill times decreased with increasing reliability levels (Section 3.3.3) indicating in the free choice scenario (vs. set time, forced choice scenario in Experiment II), evidence for a hostile target was more quickly obtained or a decision more quickly made when the reliability level increased.

Expectancy does not explain why reliability level affected sensitivity levels for only the pie and separated displays. Each was at a disadvantage in Experiment I with a decrease in sensitivity for the pie display and less appropriate reliance for the separated display. An expectancy advantage at high reliability levels may have ameliorated some of the detrimental aspects of the pie display and separated display when they were subjected to an increased time stress. The three-way interaction between stimulus duration, display proximity and reliability
level may support this result. Only the assumingly most advantageous combination of display
proximity and stimulus duration (integrated display shown for 800ms) showed a steady d’ over
all reliability levels. For the other combinations (separated display shown for 400 and 800ms and
integrated display shown for 400ms) d’ started off lower and increased as reliability level
increased.

The attention allocation and expectancy hypotheses are not mutually exclusive. As
alluded to above, expectancy of a signal may compensate for interface features that are
disadvantaged due to suboptimal attention allocation strategies. The expectancy advantage may
have manifested itself in a time measure of performance in Experiment I whereas, in the time
controlled Experiment II, it manifested itself in the sensitivity measure.

5.4.4. Effects on Response Bias

Display proximity also had an effect on decision criterion replicating the results of
Experiment I. Participants using the separated display had a more sluggish change in decision
criterion than those using the integrated display. Both in the parametric analysis of the parameter
C and the comparison of participants $\beta$ to $\beta_{optimal}$ demonstrated this effect. This effect also
manifested itself as a display proximity and reliability level interaction for both miss rate and
FArate. Miss rate decreased and FArate increased over increasing reliability levels for the
integrated but not the separated display. Finally the stimulus duration and display proximity
interaction effect on miss rate also demonstrates the advantage for the integrated display under
time pressure. When the participants viewed the scene for a shorter duration, their miss rates
increased when they were given the separated display. This supports the hypothesis that the
additional channels on the separated display may draw attentional resources away from the
target, thereby increasing the number of targets missed. There is a clear advantage for integrating
identity and reliability information on a rifle-mounted CID identification system.
5.4.5. Limitations

The attention allocation hypothesis provides an explanation for the difference in performance and reliance behaviour between the display types and display proximities; however, it has not been tested directly. Ideally, given time and resources, an eye-tracking protocol could examine overt attention shifts to give evidence support or refute the attention allocation hypothesis.

The stimulus duration manipulation attempted to examine attention allocation indirectly but it had many limitations. First, it was difficult to adjust the stimulus duration variable to levels that would show an effect. The short and long interval had to be designed to be within a span where the task would be difficult but not impossible. The stimulus duration levels were set using information from protocols of other studies (e.g., Dzindolet et al., 2001a) and pilot testing. Even with this diligence, the range over which performance may decline (i.e., show an interaction effect of stimulus duration and display type or display proximity) could be different for each of the independent and dependant variables in the study (e.g., the levels may have straddled a range which was appropriate for the effect of display proximity on d', but not, the effect of display type on C). The short and long interval also had to straddle an unknown threshold where performance may rapidly decline for the multiple variables of interest. The one advantage this indirect measure has over the direct measure of eye-tracking is that it could include effects of covert shirts of attention which eye-tracking would not measure.

The use of SDT to assess CID performance and reliance on an automated aid is superior to other methods of analysis as discussed in Section 1.2; however, there are challenges that experimenters must be aware of when implementing this analysis in a protocol. Many of the challenges stem from SDT analysis requiring a large number of trials. In both Experiment I and II, the use of multiple reliability levels led to a rapid increase in the number of trials necessary. Compounding the issue was the use of extreme reliability levels (e.g., 90%) where there is a
possibility of a false alarm on only 10% of the trials. If the participant does not commit a FA at a
given reliability level, the SDT parameters cannot be calculated and the participant has to be
dropped from the analysis. The experimenter must balance resources and the priorities of the
study to attempt to maximize the amount of data collected. In Experiments I and II, the display
type manipulation was a between-subjects variable to control the length of the experiment which
reduced the power of this variable.

5.4.6. Future Work

The two experiments were able to demonstrate which displays helped the users rely on
the automation more appropriately and allowed for greater sensitivity in target detection.
However, the explanations offered for the results are hypothetical. Direct measures of attention
allocation, such as eye-tracking, will be able to explore why performance measures differed. An
empirically evaluated explanation can further inform interface design in other domains.

Exploring the differences between different methods of displaying reliability or
uncertainty information should be a priority. In Experiment I, it was suggested differences in
performance were due to the different levels of precision between the methods tested, with
participants allocating more attention to the pie display to gain the additional precision it
afforded. A number of previous studies (e.g., Feldman-Stewart et al., 2007; Finger & Bisantz,
2002) have also shown both speed and accuracy advantage to seeming less precise displays. An
empirical study that systematically varies the precision of a graphical form could support that the
variation in results is due to inherent level of precision available between display types.

In Experiment II, it was also suggested that visual scanning of the pie and mesh forms
may differ, where the participants can look anywhere on the mesh display but have to look at
particular locations on the pie display to read reliability level. Direct measures of attention
allocation could examine sampling behaviour between different displays. Between Experiment I
and II there are multiple explanations to why there are performance differences between display methods. This contrasts with display proximity where the results are more consistent across experiments and are also consistent with the PCP. An experiment that uses both SDT and eye-tracking will be resource intensive, and therefore display method should be a priority for subsequent experiment designs.

The trust-choice incompatibility principle (Keren, 2007) has interesting implications for framing effects when display the reliability level of automation. The framing of the reliability level could affect users trust in automation. The relationship between trust and automation use and framing and automation use becomes less clear. More trust in an automated system should lead to more automation use. But negative framing can positively affect trust, but negatively affect use (i.e., the choice portion of the trust-choice incompatibility). Therefore the relationship between reliability framing and automation use is unclear and open to empirical evaluation.

6. Conclusion

Providing users with system reliability information in a well-designed HMI can affect reliance behaviour on an automated CID aid; however, the form of the interface can affect reliance behaviour and participant sensitivity in identifying the targets. SDT was a useful tool in uncovering these effects. As CID has to be completed in a high-risk environment while under time pressure, the interface must allow for the user to quickly and easily obtain the information from the interface.

The original goal of the project was to design an interface that could help users rely on an automated CID system appropriately. Clearly the form of the interface can affect reliance behaviour as the participants with the integrated interface relied on the system more appropriately likely because integrating the information into one form assisted the user in obtaining the information more quickly. However, the display method effects on sensitivity show
that an interface designer cannot ignore other performance measures when designing for a particular goal (in the present cause, more appropriate reliance).

The mesh display likely provided a speed advantage as the participant could sample anywhere on the graphical form to garner the automation reliability; whereas the pie display required specific, perhaps time consuming, sampling of the form. The pie display also provides a high level of precision that was not necessary to successfully complete the task. The designer of an interface to be used in a time-pressured environment will want to favour speed over precision, especially if a high level of precision is not required to complete the task successfully.

Attention, due to its serial nature, is an important consideration in a time-pressured task; therefore, attention allocation in both CID and interface design 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. In addition such results may provide insight into interface design in other demanding environments.

7. References


## Appendix A: Information Requirements

### Table A-1: Information Requirements

<table>
<thead>
<tr>
<th>Task</th>
<th>Requirement</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>System use</td>
<td>power status (on/off)</td>
<td>The system must be able to be turned off when not in exercise in order to preserve battery power. As the system can only identify friendly targets and provide other information when it is on, it is important the soldier has some knowledge and reassurance that the system is operational. The reassurance may be especially important if some aspects of the task (e.g., interrogating a target) are potentially automatic.</td>
</tr>
<tr>
<td>mode (training, combat)</td>
<td></td>
<td>It has been operational requirement for prior systems (e.g., Shurman 2000) to have training mode. Operating in training mode would be detrimental to CID performance during actual operation.</td>
</tr>
<tr>
<td>Battery Power (percentage, proportion?)</td>
<td></td>
<td>System requirement</td>
</tr>
<tr>
<td>Training results</td>
<td></td>
<td>System requirement</td>
</tr>
<tr>
<td>CID</td>
<td>Identity of entities in field</td>
<td>Friend, foe or neutral, obtained through system or other sensory cues (e.g., visual identification)</td>
</tr>
<tr>
<td>Appropriate system reliance</td>
<td>Acknowledgement of interrogator activation</td>
<td>If the system is designed to manually interrogate a target, there is a delay of up until a second between button press and feedback (see Sherman, 2000 &amp; Zari et al, 1997) as the system cycles through the inquiry. Although a delay of a second is within the requirements in a high pressure situation, it is imperative the individual have feedback they have correctly activated the system. It was show in previous work on the project that during manual activation, if unknown feedback was not explicit (the feedback was ‘nothing happening’) participants would activate the system numerous times, because they were not able to tell the difference between improperly activating the system and the implicit feedback for unknown targets.</td>
</tr>
<tr>
<td></td>
<td>P(enemy</td>
<td>“unknown” feedback from system)</td>
</tr>
<tr>
<td></td>
<td>P(friend</td>
<td>“friend” feedback from system)</td>
</tr>
<tr>
<td>Task</td>
<td>Requirement</td>
<td>Achieved?</td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------------------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>System use</td>
<td>power status (on/off)</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>mode (training, combat)</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>Battery Power (percentage, proportion?)</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>Training results</td>
<td>×</td>
</tr>
<tr>
<td>CID</td>
<td>Identity of entities in field</td>
<td>✓</td>
</tr>
<tr>
<td>Appropriate system</td>
<td>Acknowledgement of interrogator activation</td>
<td>×</td>
</tr>
<tr>
<td>reliance</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$P(\text{enemy}</td>
<td>\text{“unknown” feedback from system}) $</td>
</tr>
<tr>
<td></td>
<td>$P(\text{friend}</td>
<td>\text{“friend” feedback from system}) $</td>
</tr>
</tbody>
</table>
Table A-3: Informational Requirements Displayed with Previous System (Shurman, 2000)

<table>
<thead>
<tr>
<th>Task</th>
<th>Requirement</th>
<th>Achieved?</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>System use</td>
<td>power status (on/off)</td>
<td>?</td>
<td>Difficult to determine. Information on screen may indicate the system is on</td>
</tr>
<tr>
<td></td>
<td>mode (training, combat)</td>
<td>?</td>
<td>impossible to determine</td>
</tr>
<tr>
<td></td>
<td>Battery Power (percentage, proportion?)</td>
<td>✓</td>
<td>shown in screen (Sherman, 2000)</td>
</tr>
<tr>
<td></td>
<td>Training results</td>
<td>✓</td>
<td>shown in screen upon command, detailed information compiled centrally (Sherman, 2000)</td>
</tr>
<tr>
<td>CID</td>
<td>Identity of entities in field</td>
<td>✓</td>
<td>1) Sensory cues of target 2) Through colour change in indicator, blue=friend signal received, red=no signal received (unknown)</td>
</tr>
<tr>
<td>Appropriate system</td>
<td>Acknowledgement of interrogator</td>
<td>✓</td>
<td>indicator turning yellow upon activation, and remaining yellow until feedback is received</td>
</tr>
<tr>
<td>reliance</td>
<td>activation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>P(enemy</td>
<td>“unknown” feedback from system)</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>P(friend</td>
<td>“friend” feedback from system)</td>
<td>✓</td>
</tr>
<tr>
<td>Task</td>
<td>Requirement</td>
<td>Achieved?</td>
<td>Comment</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>--------------------------------------</td>
<td>-----------</td>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td>System use</td>
<td>power status (on/off)</td>
<td>✓</td>
<td>When the power is on, the screen is activated, as well as the green indicator is lit.</td>
</tr>
<tr>
<td></td>
<td>mode (training, combat)</td>
<td>✓</td>
<td>if system is in training mode, orange indicator is lit</td>
</tr>
<tr>
<td></td>
<td>Battery Power (percentage, proportion?)</td>
<td>✓</td>
<td>shown in screen</td>
</tr>
<tr>
<td></td>
<td>Training results</td>
<td>✓</td>
<td>shown in screen upon command</td>
</tr>
<tr>
<td>CID</td>
<td>Identity of entities in field</td>
<td>✓</td>
<td>Through colour change in indicator, blue=friend signal received, red=no signal received (unknown)</td>
</tr>
<tr>
<td>Appropriate system reliance</td>
<td>Acknowledgement of interrogator</td>
<td>✓</td>
<td>indicator turning yellow upon activation, and remaining yellow until feedback is received</td>
</tr>
<tr>
<td></td>
<td>activation</td>
<td></td>
<td>Auditory tone</td>
</tr>
<tr>
<td></td>
<td>P(enemy</td>
<td>“unknown” feedback from system)</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>P(friend</td>
<td>“friend” feedback from system)</td>
<td>✓</td>
</tr>
</tbody>
</table>
Table A-5: Information Requirements for Display with Continuous Analogue Proportion Unknown Feedback

<table>
<thead>
<tr>
<th>Task</th>
<th>Requirement</th>
<th>Achieved?</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>System use</td>
<td>power status (on/off)</td>
<td>✓</td>
<td>When the power is on, the screen is activated, as well as indicator X is lit.</td>
</tr>
<tr>
<td></td>
<td>mode (training,…)</td>
<td>✓</td>
<td>if system is in training mode, indicator X is lit</td>
</tr>
<tr>
<td></td>
<td>Battery Power (percentage, proportion?)</td>
<td>✓</td>
<td>shown in screen</td>
</tr>
<tr>
<td></td>
<td>Training results</td>
<td>✓</td>
<td>shown in screen upon command</td>
</tr>
<tr>
<td>CID</td>
<td>Identity of entities in field</td>
<td>✓</td>
<td>1) Sensory cues of target 2) Through colour change in indicator, blue=friend signal received, red=no signal received (unknown)</td>
</tr>
<tr>
<td>Appropriate system reliance</td>
<td>Acknowledgement of interrogator activation</td>
<td>✓</td>
<td>indicator turning yellow upon activation, and remaining yellow until feedback is received</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Auditory tone</td>
</tr>
<tr>
<td></td>
<td>P(enemy</td>
<td>“unknown” feedback from system)</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>P(friend</td>
<td>“friend” feedback from system)</td>
<td>✓</td>
</tr>
<tr>
<td>Task</td>
<td>Requirement</td>
<td>Achieved?</td>
<td>Comment</td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>--------------------------------------------------</td>
<td>-----------</td>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td>System use</td>
<td>power status (on/off)</td>
<td>✓</td>
<td>When the power is on, the screen is activated, as well as indicator X is lit.</td>
</tr>
<tr>
<td></td>
<td>mode (training,…)</td>
<td>✓</td>
<td>if system is in training mode, indicator X is lit</td>
</tr>
<tr>
<td></td>
<td>Battery Power (percentage, proportion?)</td>
<td>✓</td>
<td>shown in screen</td>
</tr>
<tr>
<td></td>
<td>Training results</td>
<td>✓</td>
<td>shown in screen upon command</td>
</tr>
<tr>
<td>CID</td>
<td>Identity of entities in field</td>
<td>✓</td>
<td>1) Sensory cues of target 2) Through colour change in indicator, blue=friend signal received, red=no signal received (unknown)</td>
</tr>
<tr>
<td>Appropriate system reliance</td>
<td>Acknowledgement of interrogator activation</td>
<td>✓</td>
<td>indicator turning yellow upon activation, and remaining yellow until feedback is received</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Auditory tone</td>
</tr>
<tr>
<td></td>
<td>P(enemy</td>
<td>“unknown” feedback from system)</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>P(friend</td>
<td>“friend” feedback from system)</td>
<td>✓</td>
</tr>
</tbody>
</table>
Appendix B: Informed Consent Form - Experiment I

Developing Human-Machine Interfaces to Support Appropriate Trust and Reliance on Automated Combat Identification Systems

Principal Investigator: M.A.Sc. Candidate Heather Neyedli  
Faculty Supervisor: Professor Greg Jamieson  
Department of Mechanical and Industrial Engineering  
University of Toronto

This study is sponsored by Defense Research and Development Canada. The purpose of this study is to discover the necessary information that can facilitate soldiers’ use of an automated Combat Identification (CID) system. The results of this study will guide the interface design for the CID system. It is expected that with this interface soldiers can better utilize the CID system in the battlefield and consequently reduce friendly fire incidents. You are invited to participate in this study because you are a student with normal or corrected-to-normal vision at the University of Toronto (UofT). The experiment will be conducted in the Cognitive Engineering Laboratory at UofT and there will be altogether 30 participants involved.

During the experiment, you will be seated in front of a computer workstation to interact with a CID virtual simulation and you will be asked to identify whether a target appearing in a simulated combat scene is a friend or enemy. The whole experiment will last approximately four hours, spread over two session (approx. 120 minutes each session with at least a two hour break inbetween) which includes the following sections:

1. Instruction (5 min): the investigator will give you instruction on how to complete tasks in the CID simulation. 
2. Training (3 min): you will practice in 2 training sessions. 
3. Formal Test (120 min): you will complete tasks in 8 mission blocks and answer a questionnaire at the end of each block. 

The Formal Test session (120 min) will also take place on the second session with a chance to review instructions and the informed consent for before proceeding.

The risk is minimal in this study and is comparable to playing a video computer simulated shooter game. You will receive a cash compensation of 40 CAD for your time and effort in this study. In addition, you will have the potential to earn a bonus 10 CAD if you achieve a level of a good performer. The cash compensation and bonus will be paid to you by a financial officer in the Mechanical and Industrial Engineering (MIE) Department at U of T.

Your privacy and identity will be carefully protected in this study. Once the experiment has been completed, the unidentifiable raw data of each participant will be assigned a “non-descriptive alias” and the Master List will be destroyed. In any publication, information will be provided in such a way that you cannot be identified.
Your participation in this study is completely voluntary. You may refuse to participate without any negative consequences. In addition, you may withdraw from the study at any time without any penalty, and request your data be destroyed. In that case your remuneration will be calculated based on the actual time you would have spent in the study, at a rate of 10 CAD per hour.

M.A.Sc. Candidate Heather Neyedli is undertaking this study in partial fulfillment of Master’s Degree requirements. If you have any additional questions later about this study, Ms. Neyedli (neyedli@mie.utoronto.ca, 416-978-0881) will be happy to answer them. For information about participants’ rights in scientific study, you can contact the Ethics Review Office at ethics.review@utoronto.ca or 416-946-3273.

You will be given a copy of this form to keep.

PARTICIPANT CERTIFICATION:

I have read this Informed Consent Form. I have had the opportunity to ask any questions that I had regarding the study, and I have received answers to those questions. By my signature I affirm that I agree to take part in this study as a research participant and that I have received a copy of this Informed Consent Form.

..................................................................................  ..................................................................................
Signature of Research Participant  Signature of Investigator

..................................................................................
(Please PRINT name)  (Please PRINT name)

..................................................................................
Date
Appendix C: Instruction Script- Experiment I

Instruction 1: (experiment procedure)

You will complete 4 mission blocks in this session, each consisting of 105 trials. In each trial, an unknown soldier, which we call the “target”, will appear in the simulated combat scene. These targets can be either hostile terrorists or friendly Canadian soldiers. Your task is to kill terrorists as soon as possible, while holding fire on Canadian soldiers.

There are two types of errors that can be made. One is made when you shoot at a friendly Canadian soldier; the other is made when you don’t kill a terrorist. Both errors are equally serious, and you should try to avoid them.

Your final score, which will determine whether or not you receive the bonus, will be the total number of the trials that you hold fire on a Canadian soldier or successfully kill a terrorist. For each block, the targets will be half terrorists and half Canadian soldiers. The order of the trials has been randomized.

After each block, you will be asked to complete a short questionnaire.
Instruction 2: (combat identification system)
The combat ID aid in this experiment simulates a real-world combat ID system which comprises two parties, an interrogator and a transponder. As shown in the graph below, a soldier with an interrogator can send out an electronic message to another soldier, and if the second soldier is fitted with a compliant transponder he will send a message back to identify himself as a friend.

Figure 1. Interrogation process of the combat ID system

To activate the interrogator you will press the ‘insert’ key which will activate the system and you will receive feedback after your weapon is pointed at a target and activated: a blue light indicates a Canadian soldier and a red light indicates a terrorist. Although this aid is usually reliable, it is not 100% reliable all the time. This is because of the occasional failures in communications between an interrogator and a transponder in a chaotic battlefield. It is possible that a red light is shown when the target is actually friendly. Either the red circle will appeared partially filled indicating the reliability of the feedback (the probability the target is in fact a terrorist) or a similar indicator below the red light will convey the same information. In contrast, blue lights will always correctly identify Canadian soldiers: the blue light will never appear when the target is a terrorist.

Fig.2 reliability of redlight feedback ***note participants who were in the pie display type group had images containing the pie display
**Instruction 3: (appearance of Canadian soldiers and terrorists)**

The different appearance of terrorists and Canadian soldiers are illustrated in the graphs below. Note that they have different helmets, masks, weapons, etc. Please take your time to observe it and when you are ready we can move on to the training session.

![Canadian](image1.png) ![Terrorist](image2.png)

**Instruction 4: (training)**

The purpose of this training session is to develop your skills in identifying the targets, practice shooting and to familiarize you with the simulation.

In the first portion of training, there will be practice trials that are similar to the trials in the mission blocks. Your task is to kill terrorists as soon as possible, while holding fire on Canadian soldiers. When the experimenter thinks that you’ve gained a certain level of accuracy identifying the targets you will move on to the second portion of training.

Second you will practice accurately shooting the target. You need to attempt to kill every target that appears on your screen.

The tasks are hard, so please don’t feel frustrated even if you make a lot of errors. The experimenter will help you to improve your performance.
Appendix D: Assessment of Instruction Comprehension

1. Please fill out the blank:

In each block, ___% of all targets will be Canadian soldiers.

When a target is a terrorist, the light on the combat ID aid should be _____.

When a target is a Canadian soldier, the light on the combat ID aid should be _____.

2. Please circle the right answer:

When a target is a terrorist, I should ______.
A. hold fire   B. shoot it as soon as possible

When a target is a Canadian soldier, I should ______.
A. hold fire   B. shoot it as soon as possible

The mistake of shooting a Canadian soldier and the mistake of not shooting a terrorist are ______.
A. equally serious   B. not equally serious

When the light on the combat ID aid is blue, it is ____ that the target is a terrorist.
A. possible   B. not possible

When the light on the combat ID aid is red, it is ____ that the target is a Canadian.
A. possible   B. not possible
Appendix E: Trust Questionnaire

Please circle the number which best describes your feeling or your impression in the mission block you just completed. Remember, there are no right answers.

1. The aid is deceptive

1 2 3 4 5 6 7
not at all extremely

2. The aid behaves in an underhanded (concealed) manner

1 2 3 4 5 6 7
not at all extremely

3. I am suspicious of the aid’s outputs

1 2 3 4 5 6 7
not at all extremely

4. I am wary of the aid

1 2 3 4 5 6 7
not at all extremely

5. The aid’s action will have a harmful or injurious outcome
6. I am confident in the aid

7. The aid provides security

8. The aid is dependable

9. The aid is reliable

10. I can trust the aid

11. I am familiar with the aid
12. I can trust that **blue** lights indicate Canadian soldiers

13. I can trust that **red** lights indicate terrorists
Appendix F: Informed Consent for Experiment II

Visual Sampling of Human-Machine Interfaces for Automated Combat Identification Systems
Principal Investigator: Dr. Justin Hollands- DRDC Toronto
Co-Investigator: M.A.Sc. Candidate Heather Neyedli & Professor Greg Jamieson
Department of Mechanical and Industrial Engineering
University of Toronto

I, ______________________ (name) of ___________________________ (address and phone number) hereby volunteer to participate as a participant in the study, “Visual Sampling of Human-Machine Interfaces for Automated Combat Identification Systems” (Protocol #L-678). I have read the information package on the research protocol, and have had the opportunity to ask questions of the Investigator(s). All of my questions concerning this study have been fully answered to my satisfaction. However, I may obtain additional information about the research project and have any questions about this study answered by contacting Heather Neyedli (416) 978-0881.

I have been told that I will be asked to participate in a two sessions of approximately 1.5 hours duration each. I understand that I will receive a copy of this consent form that provides thorough and truthful explanations of the purpose and procedures of the experiments before beginning.

I have been told that the principal risks of the research protocol are minor eyestrain and fatigue associated with working on a computer. Also, I acknowledge that my participation in this study, or indeed any research, may involve risks that are currently unforeseen by DRDC Toronto. I have been advised that the experimental data concerning me will be treated as confidential (‘Protected B’ IAW CF Security Requirements) and not revealed to anyone other than the DRDC Toronto Investigator(s) or external investigators from the sponsoring agency without my consent except as data unidentified as to source. Also, I understand that my name will not be identified or attached in any manner to any publication arising from this study. Moreover, should it be required, I agree to allow the experimental data to be reviewed by an internal or external audit committee with the understanding that any summary information resulting from such a review will not identify me personally. I understand that my name and/or service number will be part of the data stored for this experiment. Moreover, only authorized personnel will have access to the data and only group results will be presented.

For Canadian Forces (CF) members only: I understand that I am considered to be on duty for disciplinary, administrative and Pension Act purposes during my participation in this experiment. This duty status has no effect on my right to withdraw from the experiment at any time I wish and I understand that no action will be taken against me for exercising this right.

I understand that I am free to refuse to participate and may withdraw my consent without prejudice or hard feelings at any time. Should I withdraw my consent, my participation as a participant will cease immediately. I also understand that the Investigator(s) or their
designate responsible for the research project may terminate my participation at any time, regardless of my wishes. I have been informed that the research findings resulting from my participation in this research project may be used for commercialization purposes.

I understand that for my participation in this research project, I am entitled to remuneration in the form of a stress allowance in the amount of $45.20 ($22.36 for CF members and public servants on duty) for the completed experiment. Stress remuneration is taxable. T4A slips are issued only for amounts in excess of $500.00 paid during a year.

I have informed the Principal Investigator that I am currently a participant in the following other DRDC Toronto research project(s):

______________________________________________ (cite Protocol Number(s) and associated Principal Investigator(s)), and that I am participating as a participant in the following research project(s) at institutions other than DRDC Toronto:

___________________________________________ (cite name(s) of institution(s))

I understand that by signing this consent form I have not waived any legal rights I may have as a result of any harm to me occasioned by my participation in this research project beyond all risks I have assumed.

Volunteer’s Name ______________________ Signature: Date: ______________

Secondary Use of Data: I consent/do not consent (delete as appropriate) to the use of this study’s experimental data involving me in unidentified form in future related studies provided that review and approval have been given by DRDC HREC.

Volunteer’s Signature __________________________ Date ___________________

Name of Witness to Signature: Signature: __________________________ Date: ______________

Family Member or Contact Person (name, address, daytime phone number & relationship)

Section Head/Commanding Officer’s Signature (see Notes below) ______________________

CO’s Unit: __________________________

Principal Investigator: Heather Neyedli. Signature: Date: ______________

Notes:

For Military personnel on permanent strength of CFEME: Approved in principle by Commanding Officer; however, members must still obtain their Section Head’s signature designating approval to participate in this particular research project. For other military personnel: All other military personnel must obtain their Commanding Officer’s signature designating approval to participate in this research project.

For civilian personnel at DRDC Toronto: Signature of Section Head is required designating that volunteer participant is considered to be at work and that approval has been given to participate in this research project.

FOR PARTICIPANT ENQUIRY IF REQUIRED: Should I have any questions or concern regarding this project before, during, or after
participation, I understand that I am encouraged to contact Defence R&D Canada – Toronto (DRDC Toronto), P.O. Box 2000, 1133 Sheppard Avenue West, Toronto, Ontario M3M 3B9. This contact can be made by surface mail at this address or in person, by phone or e-mail, to any of the DRDC Toronto numbers and addresses listed below:

Principal Investigator or Principal DRDC Toronto Investigator: Principal Investigator: Dr. Justin Hollands, Phone: 416-635-2073, Email: Justin.hollands@drdcrddc.gc.ca
Co-Investigator: Heather Neyedli, Phone: 416-978-0881, Email: heather.neyedli@drdc-rddc.gc.ca
Chair, DRDC Human Research Ethics Committee (HREC):
Dr. Jack P. Landolt, Phone: 416-635-2120, Email: Jack.Landolt@drdc-rddc.gc.ca