# RESPONSE CRITERION PLACEMENT MODULATES THE EFFECTS OF GRADED ALERTING SYSTEMS ON HUMAN PERFORMANCE AND LEARNING IN A TARGET DETECTION TASK

BY

LEAH ANN SWANSON

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Human Factors
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Advisor:

Professor Jason McCarley

# ABSTRACT

Human operators can perform better with the use of an automated diagnostic aid than without the use of an aid in a signal detection task. This experiment aimed to determine whether any differences existed among graded aids—automated diagnostic aids that use a scale of confidence levels reflecting a spectrum of probabilistic information or uncertainty when making a judgment—that enabled better human detection performance, and either binary or graded aid produced better learning. Participants performed a visual search framed as a medical decision making task. Stimuli were arrays of random polygons ("cells") generated by distorting a prototype shape. The target was a shape more strongly distorted than the accompanying distracters. A target was present on half of the trials. Each participant performed the task with the assistance of either a binary aid, one of three graded aids, or no aid. The aids' sensitivities were the same ($d' = 2$); the difference between the aids lay in the placement of their decision criteria, which determines a tradeoff between the aid's predictive value and the frequency with which it makes a diagnosis. The graded aid with 90% reliability provided a judgment on the greatest number of trials, the graded aid with 94% reliability gave a judgment on fewer trials, and the third graded aid with 96% reliability gave a judgment on the least number of trials. The binary aid with 84% reliability gave a judgment on each trial. All aids improved human detection performance, though the graded aids trended towards improving performance more than the binary aid. The binary and graded aids did not produce significantly better or worse learning than did unaided performance. The binary and graded aids did not significantly help learning, but they certainly did not worsen human detection performance when compared to the unaided condition. These results imply that the decision boundaries of a graded alert might be fixed to encourage appropriate reliance on the aid and improve human detection performance, and indicate employing either a graded or binary automated aid may be beneficial to learning in a detection task.

## TABLE OF CONTENTS

# CHAPTER 1: INTRODUCTION

In complex task domains such as aviation, military, and nuclear power operation, operators must monitor system behavior and make frequent safety-critical decisions. These decisions depend on a diagnosis of the state of the environment, often while the operator is overwhelmed with a multiple tasks or external stressors such as fatigue or time pressure. Under such circumstances, automated diagnostic aids can reduce operator workload and improve system reliability; the aid's processing capabilities integrated with a human's cognitive capabilities can produce a human-automation system that performs better than either the human or automation alone (Mosier & Skitka, 1996; Parasuraman, 1987; Sorkin, Hays, & West, 2001).

However, aids may not always improve human performance. In some cases, while meant to decrease the human's workload, the automation may add to it (Kirlik, 1993; Parasuraman & Riley, 1997; Woods, 1995). Additionally, the reliability of the aid and the ways in which users perceive automation can also affect its efficacy (Lee & Moray, 1992; Lee & See, 2004). Contingent on the context, the automated aid should be designed to engender appropriate reliance in order to optimize human-automation performance. The current study examines an aspect of automated diagnostic aid's design, the placement of response criteria, likely to affect the behavior of human users.

# CHAPTER 2: LITERATURE REVIEW

Automated aids can enhance human performance in a variety of ways, facilitating information acquisition, information and system analysis, decision making, action execution, learning a task, and detection performance in a visual search task (Glover, Prawitt, & Spilker, 1997; Parasuraman, Sheridan, & Wickens, 2000; Riley, 1997; Sheridan, 2002). However, the benefits of automated aids to human performance are not guaranteed, and aids often fail to improve human performance in the ways expected by designers (Parasuraman & Riley, 1997). These failures generally take one of two forms depending on the aid's reliability: *complacency and disuse*. The users' perception of the automation's reliability shapes the tendency toward complacency, appropriate use, or disuse (Singh, Molloy, & Parasuraman, 1993a, 1993b; Parasuraman & Riley, 1997). Complacency occurs when users rely on an aid more than they should. Initially, users tend to have a positive perception of an automated aid, believing that it will outperform them, even when they are told that the aid is not perfect (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Wiegmann, 2002). This perception can persist until the aid commits an error and can engender complacency in the human operator, a tendency to over trust and unquestionably rely on the aid. In one scenario, for instance, a pilot inappropriately trusted cockpit automation and failed to notice that the autopilot was failing (Sparaco, 1995). Here, the user over-trusted the automation and acted complacently (i.e., complying with each decision made by the aid). Complacency does not pose a problem as long as the aid is perfect. However, when relying on an imperfect aid, complacency will result in performance decrements. An operator behaving complacently is less likely to check the aid's decision; therefore, any mistakes made by the automation will be missed by the operator (Singh et al., 1993a). One study compared the behavior of users assisted either by perfect automation or imperfect automation.

Participants who were assisted with perfect automation behaved complacently, relying completely on the automation without sampling any other available information. The participants who were assisted by imperfect automation did not act as complacently as the alternative group. Those who behaved complacently made a greater number of commission errors, erroneous decisions to agree with the automation's diagnosis even when the aid was incorrect. When the imperfect automation failed, participants who caught the error looked at other information whereas those who failed to catch the aid's error neglected to examine other information (Bahner, Huper, & Manzey, 2008).

On the other hand, when an aid is susceptible to failure, people may begin to distrust it even if the failures occur infrequently, believing it does not work. In this case, users may rely upon the aid less than they should (Maltz & Meyer, 2001; Wiegmann & Cristina, 2000; Wiegmann, Rich, & Zhang, 2001), an effect described as disuse (Parasuraman & Riley, 1997). As an example, consider smoke detectors. Smoke detectors are designed to use a low detection threshold, such that even small amounts of smoke will activate the alarm. As a result, users perceive smoke detectors to be unreliable and, thus, disuse them (Parasuraman & Riley, 1997). Furthermore, when users believe they have detected a signal that the aid neglected to detect, they no longer trust the aid and disuse it despite being told that the aid would outperform them (Dzindolet et al., 2003). Users in this situation begin to ignore the aid's diagnoses altogether even though they know their joint performance with the aid is better than the performance of the human alone (Wiegmann & Cristina, 2000). Disuse obviously can compromise performance of the human-automation system. For example, in a study comparing user agreement rates—how often participants agree with an aid's diagnosis in light of that aid's reliability—participants who agreed with the aid most of the time were about as accurate as the aid was reliable (i.e., participants were 80% accurate when assisted by an aid that was 80% reliable) and performed better than the group who agreed with the aid less often (Wiegmann, 2002). In general, the frequency of automation errors determines the user's trust in the automation (Lee & Moray, 1994). However, the perceived unreliability of an automated aid, as determined by the number of errors it makes, can be exacerbated by a common user assumption that an aid should perform flawlessly (Dzindolet, Pierce, Beck, & Dawe, 1999).

Previous studies have found a reliability threshold below which an aid appears to lose all value to the human operator (Rovira, McGarry, & Parasuraman, 2007). According to Wickens and Dixon (2007), unaided performance is better than performance assisted by an aid of any reliability below 0.70, and thus users completely lose trust in the automation and find it useless if its reliability is below 0.70. As long as an aid has a high reliability—that is, reliability above the .70 threshold—it will enable performance better than unaided performance (Skitka, Mosier, & Burdick, 1999). But, an aid whose reliability is nearly perfect may elicit complacency while users distrust an aid with a lower reliability. Hence, there may be a possibility of a quadratic effect in which an aid with near-perfect reliability is better than both an aid with lower reliability and an aid with higher reliability.

A user's current workload level can also influence his or her willingness to utilize an automated aid (Parasuraman & Riley, 1997; Sorkin & Woods, 1985). When a human operator's workload increases, automation can potentially offset the attendant costs and even allow the operators to attend to other tasks (Parasuraman, Molloy, & Singh, 1993). But, there is mixed evidence for this effect. In a study where participants were tasked with tracking, monitoring, and managing tasks, for example, participants were told they could use automation for the tracking task. Surprisingly, the data gave no evidence of increased automation use during periods of high workload (Harris, Hancock, & Arthur, 1993). Under some circumstances, automation may actually add to users' workload. For instance, automated tasks in the cockpit may require the aircrew to monitor the automation because it, too, can make mistakes or fail (Chambers & Nagel, 1985). Workload may also increase if the human is required to engage or disengage an aid (Kirlik, 1993). The time necessary to engage or disengage the automation—which adds additional time before the user can begin the next task—may deter people from using it at all, especially when the cost of engagement transcends the benefits.

As mentioned, aviation accidents and other catastrophic events can result from users' misguided use or disuse of an automated aid (Mosier & Skitka, 1996). Therefore, an aid's design must foster appropriate use if it is to be effective in supporting human detection performance (e.g., Rovira et al., 2007). In safety-critical situations such as airport luggage screening, the costs of a missing signal can be great. The automation's design should recognize this with a liberal shift of beta, decreasing the miss rate. But, in doing so, the aid becomes false-alarm prone (Parasuraman & Riley, 1997). Too many false alarms may cause users to ignore the aid and lose trust (Gupta, Bisantz, & Singh, 2002; Horowtiz & Dingus, 1992), which can be catastrophic when the aid alerts the user to a true target.

However, factors beyond the aid's design—costs and payoffs, rewards and punishments, knowledge, and other factors idiosyncratic to an individual user—influence automation use (Lee & See, 2004). In the airport luggage screening example, screeners might be punished for every piece of luggage containing a weapon or suspicious object that they failed to pull off the belt for inspection. To decrease his miss rate, a user might decrease the judgment parameter, beta. On the other hand, due to the rare occurrence of a weapon in a piece of luggage, screeners may increase their response threshold. In doing so, they are conservatively choosing which pieces of luggage they will manually inspect.

When users have information about the reasons the aid might make a mistake, they have greater trust in the aid when compared to users who lack the knowledge of the aid's imperfection (Dzindolet et al., 2003). An understanding of the way in which the aid works—how it makes a decision—fosters better trust in an automated aid. Users relied on an aid of lower reliability as much as an aid with higher reliability when this information was provided to them. But, the cost in doing so leads to inappropriate reliance on the aid whose reliability was lower. Additionally, users increased their reliance on an aid when they were provided with information about how well they and their aid performed separately rather than providing them only with the aid's diagnosis. Participants who only see the aid's diagnosis, without feedback on their own performance or the aid's, may begin to rely on themselves only (e.g., participants will make a judgment on a particular experimental trial, view the aid's judgment, and see that the aid is obviously wrong in its judgment).

*Graded vs. binary aids*

The ideal solution might seemingly be to design all automated aids to be perfectly reliable, but this is impossible to do. Diagnostic aids are designed to perform a signal detection task, and performance is limited by the strength of the signal (Green & Swets, 1966; Wickens & Hollands, 2000). Since the strength of the signal is often beyond a designer's control, other techniques—for example, manipulations of an aid's interface design—may be the only way to improve performance of a human-automation system. One way to improve aid utilization might be to employ graded aids. Graded alerts render diagnoses on a scale of confidence levels, reflecting a spectrum of probabilistic information or uncertainty (Bisantz, Finger, Seong, &

Llinas, 1999), when making a judgment about the state of the environment. They can provide a judgment to the user with an associated level of certainty (e.g., high confidence, low confidence, and a neutral standing in a three-level graded aid), or can convey the urgency of a situation on a spectrum ranging from completely safe (target absent) to imminent danger (target present) with judgments in the middle to express some danger. Confidence ratings can be derived, as in the standard signal detection model, by dividing the evidence axis into three or more regions using multiple response criteria, rather than simply dividing the evidence axis with a single response criterion into regions corresponding to yes and no judgments. The potential benefit of using a graded aid is that it provides extra information to users about the state of the system, allowing them to make a better informed decision (Sorkin, Kantowitz, & Kantowitz, 1988; Woods, 1995). Additionally, when users must attend to several tasks or have a high workload, this extra information about the likelihood of a target allows can help users decide where best to allocate their attention (e.g., if the graded aid conveys a high likelihood of danger, the user may decide to attend to avoiding the danger whereas the user may decide to continue attending to the current task should the aid convey a low likelihood of the target).

Unfortunately, studies testing the value of graded aids relative to binary aids have produced ambiguous results. Some evidence has endorsed the use of graded aids. In a study examining decision making during the interaction between people and an integrative cockpit display, for example, participants who used a likelihood alert made more accurate decisions than those who used a binary aid, especially with a high workload (Bustamante, 2008). In another study, Andre and Cutler (1998) use graded aids in a navigation task during which an operator was asked to come as close as possible to a target without actually hitting it. The target's location was estimated with a circle that encased the target, and the circle's size changed as a function of the uncertainty with regard to the target's location. The circle helped operators to avoid hitting the target because it provided them with a level of uncertainty regarding the target's location. Gupta et al. (2002) used auditory graded alerts in a simulated driving task: in the binary-aid condition, each alert was equally loud to convey a potential skid or collision, but in the graded-aid condition, alerts differed in loudness depending on the urgency of the situation. Driving performance was better when graded alerts were used not only because they provided drivers with information about the likelihood of a skid or a collision, but also because the binary alert caused drivers to react very quickly, accelerating the car right after the alarm.

Other work, unfortunately, has found less value to graded aids. In one case, three different levels of a false alarm-prone alarm and a miss-prone alarm were provided, and each one differed with regard to the level of imminent danger such that the lowest grade of the alert was the word "OK" inside of a green box (Clark, Peyton, & Buastamante, 2009). A binary aid was used for comparison; only two grades—target-present or target-absent—were provided. The likelihood alert fostered better decision making, but only with the false alarm-prone aid because the miss-prone aid reduced operator reliance (Clark et al., 2009). The likelihood alert in this study fostered better trust as is reflected by the increase in user response to correct judgments made by the aid. Other work found that people relied more on a binary, false alarm-prone aid than on a likelihood, false alarm-prone aid (Stanton, Ragsdale, & Bustamante, 2009), and some studies have found no benefits at all for graded alerts (Wickens & Colcombe, 2007).

Why have studies' graded aids produced somewhat contradictory findings? One aspect of a graded aid's design that might affect operator performance is the separation between response criteria used to divide the evidence axis into different confidence levels. The current experiment illustrates this possibility, using two response criteria to separate the signal detection evidence axis into regions corresponding to *yes*, *no*, and *uncertain* diagnoses. Evidence values that fall above both response criteria produce confident *yes* (target-present) diagnoses and those that fall below both criteria produce confident *no* (target-absent) diagnoses. The evidence values that fall between the two response criteria are rendered as uncertain judgments. The separation between criteria influences the number of diagnoses rendered by the aid and the confidence with which each diagnosis is made (Figure 1). A diagnosis that falls above the two response criteria that are separated by many standard deviations will be extremely accurate as compared to a diagnosis that falls just above the criteria that are separated by fewer standard deviations. This is because the area under the signal curve is small when the criteria are placed further apart from one another, and the diagnosis is made at the very end of the curve where the probability of noise is nearly zero. McCarley (2009) found evidence for a role of criterion placement in determining the value of graded alerts. Participants in his experiment were more willing to use the graded aid than the binary aid in a simulated baggage screening task. However, they were more willing to use a graded aid that rendered a judgment more often than an aid of equal sensitivity that gave a judgment less often. This result suggests that the placement of an aid's response criteria is an important factor in designing an aid that will foster appropriate use and improve human detection performance. The current study explores this issue further, providing a comparison among three graded aids of varying response criteria to determine if any differences exist among them and if an optimal response exists.

### *Skill acquisition with graded and binary aids*

An unavoidable concern in providing human operators with automated aids is the possibility of the aid's failure. Having the assistance of an automated aid, while often useful, might hurt people's ability to perform a task unaided (Wickens & Hollands, 2000), or to develop skills in the unaided task (Glover et al., 1997). As a result, performance may suffer if the aid fails or otherwise becomes unavailable to the operator (Goh, Wiegmann, & Madhavan, 2005). On the other hand, some research has found that once an automated aid was removed from the experiment, performance of operators who had previous experience with an automated aid was better than those who did not have experience with one (Goh et al., 2005). The same work suggests that using automated cueing to direct people's attention to a possible target improves detection performance and can be used as a training mechanism for visual search tasks such as luggage screening.

As yet, there is a lack of research exploring how differences in the effects of graded aids and binary aids on human users' learning in a signal detection task. Graded aids might produce poorer learning than binary aids because graded aids do not provide a recommendation as often as the binary aids (e.g., feedback on 82% of the task may be less useful than feedback on the entire task). Feedback provided on a proportion of the task does not give as much guidance and, hence, may not produce learning nearly as well as learning produced from the feedback provided during the entire task. Alternatively, too much assistance (providing a recommendation all of the time) may worsen learning because users may rely on the recommendations and fail to check their performance (Schmidt & Bjork, 1992). The users may fail to learn the task, rendering the binary aid useless for learning. Users will develop longer-lasting skills when they are allowed to make their own mistakes during training rather than complying with the aid on every trial. The graded aids might produce better learning since they provide a recommendation on only some of the trials. Schmidt and Bjork (1992) suggest some feedback during training to improve the user's performance, but not too much where the user does not learn.

# CHAPTER 3: EXPERIMENT

## 3.1. Introduction

The purpose of this study was to determine where the response criteria of the graded aids need to be placed for best performance and whether graded aids produce greater learning effects than binary aids. Participants completed a signal detection task framed as a medical decision making task, in which they were instructed to search a sample of simulated human cells on the computer screen for an abnormal-looking cell. Some participants received an automated diagnosis regarding the presence or absence of an abnormality, while the other participants completed the task unassisted. Four forms of automated aids were employed. All four were equal-variance Gaussian signal detection systems (MacMillan & Creelman, 2005) with a $d'$ of 2. One provided a binary judgment, making a target-absent or target-present diagnosis on each of the 200 experimental trials. This aid possessed a single response criterion that was placed at the point of unbiasedness, in the center of the signal and signal plus noise distributions. The other three automated aids each made graded judgments, offering either a confident target-absent, a confident target-present, or a neutral diagnosis each trial. One graded aid had two response criteria that were separated by 0.75 standard deviations, position equal distances to either side of the point of unbiasedness, such that the aid provided a diagnosis on 82% of all trials with a predictive value of .90 (i.e., diagnoses were 90% accurate). On the remaining 18% of the trials the aid gave a neutral diagnosis, forcing the participant to effectively perform the task unassisted. The second graded aid had two response criteria separated by 1.5 standard deviations, again centered about the point of unbiasedness, generating a diagnosis on 64% of all trials with a predictive value of 0.94; on the remaining 36% of the trials the aid gave a neutral diagnosis. The final graded aid had two response criteria separated by 2.25 standard deviations, centered on the point of unbiasedness, generating a diagnosis on 47%, of all the trials with a predictive value of 0.96; on the remaining 53% of the trials, the aid gave a neutral diagnosis. During the first 150 (aid-present) trials, aided participants worked with their assigned automated aid. During the last 50 (aid-absent) trials, aided participants were stripped of their assigned aid and left to perform the task unaided.

## 3.2. Methodology

### 3.2.1. Participants

One hundred and fifteen undergraduate students (mean age = 19.54, 45 males) at the University of Illinois participated for course credit.

### 3.2.2. Stimuli

Stimuli were images of random polygons generated using Posner and Keele's (1968) prototype-distortion procedure (see Smith, Redford, Gent, & Washburn, 2005, for use of similar stimuli in a visual search task). Each stimulus image contained five filled polygons, drawn in gray with 50% transparency. Each polygon was drawn at a random location, and individual items were free to overlap. All five polygons within an image were distortions of the same prototype. Two versions of each stimulus image were created, a target-absent version and a target-present version. Within the target-absent image, all five polygons were moderate (level 2; Posner & Keele, 1968) distortions of a common prototype. Within the target-present image, one polygon was distorted from the prototype more highly (level 7.7) than the other four. Aside from this difference, the target-absent and target present versions of an image were identical. In total, 200 pairs of target-present/target-absent images were generated, each pair using a different prototype.

### 3.2.3. Procedure

After reading and signing a consent form, participants read instructions on the computer before performing the 200-trial visual search task. The instructions asked each participant to imagine that he or she was a doctor screening cell samples for abnormalities. The instructions explained that the participant would see a sample of five objects each trial, and defined an abnormality as an object that differed in shape from the other objects. Participants were instructed that an abnormality would be present on 50% of all trials, and because of random variation in shape, normal cells would occasionally appear to be abnormal and abnormal cells would occasionally appear to be normal. Instructions in the aided conditions also explained that the participant would perform the task with the assistance of a computerized aid and warned participants that the aid is imperfect. The instructions also told participants that because the aid uses sensors that are different from the human visual system, it may detect some targets that users miss and miss some targets that users might detect.

The participant initiated each trial with a key press. The trial began with a 200 ms text message from the aid ("ABNORMAL," "NORMAL," or "Waiting for sample…"), followed by a 200 ms blank screen and a two-second presentation of the stimulus image (see Figure 2). After the stimulus image was removed, the participant was prompted to report whether he or she had detected an abnormality, pressing "1" on the keyboard to report "no" and "3" to report "yes." Afterwards, he or she was asked to provide a 1 to 3 confidence rating—where 3 represents high confidence—on his or her judgment. There were a total of 200 trials. The stimulus image for each trial was selected randomly without replacement from the set of 150 target-absent/target-present pairs, and the target-present image was presented with a probability of .50 each trial.

3.2.4. Design

Each participant was randomly assigned to one of five experimental conditions (unaided control; binary aid; graded aid with separation of 0.75 SD; graded aid with separation of 1.5 SD; graded aid with separation of 2.25 SD); each condition had 23 participants. To analyze learning effects, data were broken into aid-present (trials 26-150) and aid-absent blocks (trials 151-200) and then compared; the first 25 trials were treated as practice and excluded from analysis.

3.3. Results

3.3.1 Analysis

Raw data were transformed into S prime values. S prime is a nonparametric measure of sensitivity that is more reliable than′ and more sensitive than the area under the ROC (Balakrishnan, 1998). It is similar to d′ in that it estimates the mean differences between the two signal and signal plus noise distributions, but it doesn’t assume a normal distribution for these two curves (MacMillan & Creelman, 2005). Because it is more sensitive, the S prime values were obtained before analyses. Data values of the area under the ROC curve and the S prime data values were consistent.

3.3.2 Aid-assisted Trials

Data were submitted to a univariate ANOVA with automated aid condition as a between-subjects factor. The F-test was significant for trials 26 through 150 when participants in the aided conditions had the opportunity to use the aid, $F(4,110) = 5.20$, $p = .001$ (refer to Figure 3). Orthogonal contrasts were constructed to make the following planned comparisons: mean of all automation aided conditions versus mean of the unaided condition; mean graded aid conditions to mean of the binary aid condition; linear effects of separation between criteria for the graded aids, and quadratic effects for graded aids. The first contrast was constructed to determine if all automated aids were better than the unaided condition. Because the experimental interest lied in performance differences between the two kinds of automated aids—binary and graded—the second contrast looked to see whether graded aids improved human detection performance better than the binary aid. One of the experimental goals was to find the best response criteria placement for best graded aid performance, so the third contrast looked at any linear differences among these aids: perhaps as the two response criteria move farther apart from one another, for example, performance worsens. Alternatively, perhaps there is an optimal amount of response criteria separation such that performance of aids whose criteria were closer together or farther apart were worse; this was the question addressed by the fourth contrast. The only statistically significant contrast compared all automated aids to no aid, $F(1,110) = 17.01$, $p < .001$, $r$ =.36, where *r* (effect size) is the simple correlation between subject scores and the coefficients of the groups to which they belong (Rosenthal, Rosnow, & Rubin, 2000). The second contrast was not significant, $F(4,110) = 2.08$, $p = .15$, $r = .13$, and there were no linear, $F(4,110) = 1.58$, $p = .21$, r = .11, or quadratic effects, $F(4,110) = 0.16$, $p = .69$, r = .03.

Because users may rely differently upon graded aids that differ in predictive value, it is important to examine the frequency with which people relied upon each of the aids. Agreement rates for trials on which the aid rendered a diagnosis served as a measure of how criterion placement affected the participant’s willingness to act on an aid’s judgments. This would determine if participants assisted by an aid with a higher predictive value agreed with the aid when it made a correct diagnosis more often than participants assisted by an aid with lower predictive value. Data from trials in which the aid gave a correct diagnosis and the participant agreed with the aid were submitted to a univariate ANOVA with automated aid condition as a between-subjects factor. The same planned orthogonal contrasts used to compare all of the aid-assisted trials were used to compare the graded aids to the binary aid and test for linear and quadratic effects. Although the results are not significant (all *p*s > .05), participants trended towards trusting the graded aids more than the binary aids. There were not sufficient data to examine performance for neutral-judgment trials in the automation-aided conditions; the binary aid renders a diagnosis on every trial, the aids with 0.75 standard deviations and 1.5 standard deviations separating their response criteria made a diagnosis on most of the trials.

A pair of additional analyses was conducted to ensure that potential effects of aid format were not obscured by variance due to learning throughout the course of an experimental session. Although automation-assisted participants were given a description of the aids before beginning the experimental task, they might have required a significant number of trials to familiarize themselves with the aids' behavior and to establish and calibrate their trust in the aids' judgments. Their willingness and ability to properly utilize the aids might therefore have changed over the course of the experimental session. To minimize the variance due to learning, an analysis repeated the analyses described above, but using only the last 50 of the aid-assisted trials. The pattern of results was the same as that seen in the analysis trials 26-150.

### 3.3.3 Unassisted Trials

The F-test for trials 151-200 when the automated aids were taken away from participants is not significant, although all automated aids performed better than no aid (Figure 4). All of the automated aids produced performance that was numerically better than the unaided condition, and the graded conditions were not significantly different from the binary condition. However, the graded aids did not hinder learning.

### 3.4. Discussion

Generally, detection performance was better when humans were assisted by automated aids than without assistance as shown by the first contrast. The lack of statistical significance for the second contrast comparing all graded aids to the binary aid suggests that the binary aid performed as equally well as the graded aids, although the graded aids trended toward significance. The third contrast tested for linear effects of the graded alerts while the fourth contrast tested for quadratic effects. Neither contrast was statistically significant. This suggests that none of the graded aids improved human performance more than the others: there was no linear trend such that the graded aid whose response criteria were separated by 0.75 standard deviations improved human performance the most, followed by the aid whose criteria were separated by 1.5 standard deviations and, lastly, the aid whose criteria were separated by 2.25 standard deviations (which would have improved human performance the least). Likewise, there was not an optimal criteria separation such that the aid whose criteria were separated by 1.5 standard deviations showed the greatest improvement to human performance while the remaining two aids showed the least improvement to human performance.

Additionally, automated aids may improve learning. Detection performance was better for users who were previously assisted by an aid than for users who have only performed the task manually. Although none of the aids were significantly better than the unaided condition, the aids trended toward significance, suggesting that any kind of automated aid may help learning and certainly does not hinder performance.

The similarity of the last 50 aid-assisted trials to all 125 aid-assisted trials suggests that participants utilized the aid similarly throughout the aid-assisted trials. Had the participants taken a number of trials to figure out how to use the aid or decided to disuse the aid at a given point in the experiment, there would have been a pattern of effects across all trials, and the last 50 aid-assisted trials would have shown different effects from the total (150) number of trials.

# CHAPTER 4: GENERAL DISCUSSION

Past work has produced contradictory evidence for the benefits of graded diagnostic automation as compared to automation producing binary judgments. The current experiment asked whether the placement of the graded automation's response criteria would modulate the automation's effectiveness as an aid to human decision makers. The results show that the assistance of an automated aid, binary or graded, can improve human performance in a signal detection task. However, graded aids trended toward improving performance more than the binary aid. These results imply that participants were using the graded aids similarly. Because the participants trended towards agreeing with the graded aids when the aids were correct more often than the binary aid when it was correct, the additional information provided by each graded aid—a range of diagnoses rather than a single, yes or no diagnosis—may have helped the participants more, though the effects fell shy of statistical reliability.

In other circumstances, or by other measures, graded automated alerts may offer more substantial benefits over binary aids. Although the current experiment did not measure workload, past work has found that graded aids can reduce workload (Stanton, Ragsdale, & Bustamante, 2009). This may be an important benefit even if the graded aid does not improve detection performance, *per se*. In a true medical scenario, for instance, a physician who must examine a sample of human cells may want automated assistance if he has a high patient load or feels fatigued from long hours. Operators who must maintain vigilance in a monitoring task such as air traffic control, screening luggage at the airport, or screening human cells for abnormalities might also be best to use a graded alert (Clark, Peyton, & Bustamante, 2009), since it provides information about how likely a target is to appear. Alternatively, there may be conditions under which binary aids would be more useful than graded aids, for example, under high time stress.

When it gives a neutral judgment, the graded aid requires the human operator to make the decisions on her own, which might slow down performance. Under high time stress, it might therefore be useful to give a binary recommendation, assuming that the binary judgments are sufficiently reliable. By analogy, status displays—those providing information regarding the status of the system to the human—may be more useful in certain scenarios than command displays—those ordering the human to perform a particular action (Wickens & Hollands, 2000). A status display is better than a command display if the information provided is not reliable. However, a command display is better than a status display when the human is under high time stress, because the command display makes a decision that the human would otherwise need to make.

Learning may benefit from the use of an automated aid, either binary or graded. As evidenced by the current experiment, experience performing the task with an aid produced improvements in unaided performance that were at least as good as the improvements seen from unaided practice on the task. Perhaps the reason is that participants used the aid's assistance to process and make a diagnosis from the information given to them and learn to differentiate between abnormal and normal cells to detect the abnormalities as suggested by Goh et al. (2005). More notably, the graded aids produced learning roughly as strong as that resulting from the binary aids. Providing a substantial number of neutral diagnoses during the learning trials thus did not seem to make the graded aids less effective in helping the human operators develop skill in the detection task.

Perhaps the way in which the automated aids' presented their diagnoses in the current experiment shaped the way the participants used the aids and their ability to detect abnormalities. Thus, a potential follow-up study could look at the differences between text message diagnoses and cued location diagnoses. The text message diagnoses in the current experiment served as an indirect cue, alerting the user to the possibility of an abnormality in the image (Goh, Wiegmann, & Madhavan, 2005) without drawing attention to the specific object that was the high-likelihood target. Perhaps a direct cue, highlighting or placing a ring around the one likely target, would change user behavior and detection performance. A direct cue may cause attentional tunneling (Maltz & Shinar, 2004), but may be more helpful than an indirect cue since it provides further information about the anomaly's location. Previous research has found that users had greater reliance and better performance when assisted by a direct cue as opposed to a text message despite equal reliability (Wiegmann, McCarley, Kramer, & Wickens, 2006). Another study investigated the use of binary and graded direct-cue aids in a search task (St. John & Manes, 2002). The graded version of the direct cue provided a range of information to the user about the likelihood of a target in a particular location, which would direct the user's search. However, this aid was only beneficial when it was reliable, and the graded direct-cue aids complicated the user's search path. However beneficial the direct cue, factors such as target salience and the frequency with which a user expects to see a target will affect detection performance (Yeh & Wickens, 2001).

Alternatively, perhaps the automated aids' judgments biased the participants' judgments before they had the opportunity to view the stimuli in the current study, and presenting the stimuli before the aid's judgments might change user behavior and performance. Perhaps the graded aids will be significantly better than the binary aid in improving human detection performance because the human will have made her own judgment and use the graded aid merely as confirmation with an associated level of confidence. The user might render the binary aid useless because it neglects to provide likelihood information and serves only as a confirmation or contradiction to the user's judgment.  Therefore, a future study could look at the differences in user behavior as the order of information changes (e.g., presenting the image of cells first before providing the aid's diagnosis).  There is evidence that the way in which information is provided affects an operator's decision bias such that operators recalled the most recently presented information best (Ashton & Ashton, 1990; Perrin, Barnett, Walrath, & Grossman, 2001).  However, Balzer and colleagues (1992) found that when the aid's diagnosis was presented after users made a diagnosis themselves, the aid does not significantly improve the user's detection performance.  But, this study employed a binary aid.  Perhaps assistance in the form of likelihood information will improve detection performance.

The current research was designed to investigate the differences between three graded aids to determine the optimal response criterion for signal detection performance and which form of automated aid, binary or graded, would produce better learning using a simulated medical scenario.  Although the evidence for an optimal criteria placement was weak, this research indicated that automation-assisted performance is better than unaided performance in a signal detection task and in learning.  This has implications for designing training tasks or for the use of automation in a learning scenario.  The decision to use a binary versus a graded aid will be dependent upon the context of its use as supported by previous research.

# References

Andre, A. D., & Cutler, H. A. (1998). Displaying Uncertainty in Advanced Navigation Systems. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, 31-35. Santa Monica, CA: Human Factors and Ergonomics Society.

Ashton, R. H., & Ashton, A. H. (1990). Evidence responsiveness in professional judgment: Effects of positive versus negative evidence and presentation mode. *Organizational Behavior and Human Decision Processes, 46,* 1-19.

Bahner, J. E., Huper A., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies, 66*, 688-699.

Balzer, W. K., Sulsky, L. M., Hammer, L. B., & Sumner, K. E. (1992). Task information, cognitive information, or functional validity information: which components of cognitive feedback affect performance? *Organizational Behavior and Human Decision Processes, 53*, 35-54.

Balakrishnan, J. D. (1998). Some More Sensitive Measures of Sensitivity and Response Bias. *Psychological Methods, 3*(1), 68-90.

Bisantz, A. M., Finger, R., Seong, Y., & Llinas, J. (1999). Human Performance and Data Fusion Based Decision Aids. In *Proceedings of the 2nd International Conference on Information Fusion*, 918-925

Bustamante, E. A. (2008). Implementing Likelihood Alarm Technology in Integrated Aviation Displays for Enhancing Decision-Making: A Two-Stage Signal Detection Modeling Approach. *The InternationalJournal of Applied Aviation Studies, 8*(2), 241-260.

Chambers, N., & Nagel, D. C. (1985). Pilots of the future: Human or computer? *Communications of the Association for Computing Machinery*, *28*, 1187-1199.

Clark, R. M., Peyton, G. G., & Bustamante, E. A. (2009). Differential Effects of Likelihood Alarm Technology and False-Alarm vs. Miss Prone Automation on Decision Making. In *Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting*, 249-252. Santa Monica, CA: Human Factors and Ergonomics Society.

Dzindolet, M. T., Peterson S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58,* 697-718.

Dzindolet, M. T., Pierce, I. G., Beck, H. O., & Dawe, I. A., (1999). Misuse and disuses of automated aids. In *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting*, 339-343. Santa Monica, CA: Human Factors and Ergonomics Society.

Glover, S. M., Prawitt, D. F., & Spilker, B. C. (1997). The Influence of Decision Aids on User Behavior: Implications for Knowledge Acquisition and Inappropriate Reliance. *Organizational Behavior and Human Decision Processes, 72*(2), 232-255.

Goh, J., Wiegmann, D. A., & Madhavan, P. (2005). Effects of Automation Failure in a Luggage Screening Task: A Comparison Between Direct and Indirect Cueing. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting,* 492-496. Orlando, FL: Human Factors and Ergonomics Society.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* Oxford, England: John Wiley.

Gupta, N., Bisantz, A. M., & Singh, T. (2002). The effects of adverse condition warning system characteristics on driver performance: an investigation of alarm signal type and threshold level. *Behavior & Information Technology, 21*(4), 235-248.

Harris, W., Hancock, P.A., & Arthur, E. (1993). The effect of task load projection on automation use, performance, and workload. In *Proceedings of the 7th International Symposium on Aviation Psychology*, 890A-890F. Columbus: Ohio State University.

Horowitz, A. D., & Dingus, T. A. (1992). Warning signal design: a key human factors issue in an in-vehicle front-to-rear end collision warning systems. In *Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting*, 1011-1013. Atlanta, GA.

Kirlik, A. (1993). Modeling Strategic Behavior in Human-Automation Interaction: Why an “Aid” Can (and Should) Go Unused. *Human Factors, 35*(2), 221-242.

Lee, J.D., & Moray, N. (1992). Trust, control strategies, and allocation of function in human-machine systems. *Ergonomics, 35*, 1243-1270.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators’ adaptation to automation. *International Journal of Human-Computer Studies, 40*, 153-184.

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors, 46*, 50-80.

MacMillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user’s guide*. Mahwah, NJ: Erlbaum.

Maltz, M., & Meyer, J. (2001). Use of warnings in an attentionally demanding detection task. *Human Factors, 43*(2), 217-226.

Maltz, M., & Shinar, D. (2004). Imperfect vehicle collision warning systems can aid drivers. *Human Factors, 46*, 257-366.

McCarley, J. S. (2009). Response Criterion Placement Modulates the Benefits of Graded Alerting Systems in a Simulated Baggage Screening Task. In *Human Factors and Ergonomics Society Annual Meeting Proceedings, 53*(17), 1106-1110.

Mosier, K.L., & Skitka, L.J. (1996). Human decision makers and automated decision aids: made for each other? In: Parasuraman, R. and Mouloua, M. (Eds.)*, Automation and Human Performance: Theory and Applications* (pp.201-220). Hillsdale, NJ: Erlbaum.

Parasuraman, R. (1987). Human-computer monitoring. *Human Factors, 29,* 695-706.

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors, 39*(2), 230-253.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced “complacency.” *International Journal of Aviation Psychology, 3,* 1-23.

Parasuraman, R., Sheridan, T.B., & Wickens, C.D. (2000). A Model for Types and Levels of Human Interaction with Automation. *IEEE Transactions on Systems, Man,and Cybernetics—Part A: Systems and Humans, 30*(3), 286-297.

Perrin, B. M., Barnett, B. J., Walrath, L., & Grossman, J. D. (2001). Information Order and Outcome Framing: An Assessment of Judgment Bias in a Naturalistic Decision-Making Context. *Human Factors, 43*(2), 227-238.

Posner, M. I., & Keele, S. W. (1968). On the Genesis of Abstract Ideas. *Journal of Experimental Psychology, 77*(3), 353-363.

Riley, V. (1997). What avionics engineers should know about pilots and automation. In *Proceedings of the Digital Avionics Systems Conference,* 123-127.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and Effect Sizes in Behavioral Research*. New York, NY: Cambridge University Press.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task. *Human Factors, 49*(1), 76-87.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*(4), 207-217.

Sheridan, T. B. (2002). *Humans and Automation: System Design and Research Issues*. Santa Monica, CA: Human Factors and Ergonomics Society.

Singh, I. L., Molloy, R., & Parasuraman, R. (1993a). Automation-induced "complacency": Development of the complacency-potential rating scale. *International Journal of Aviation Psychology, 3,* 111-121.

Singh, I.L., Molloy, R., & Parasuraman, R. (1993b). Individual differences in monitoring failures of automation. *Journal of General Psychology, 120*, 257-373.

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

Smith, J. D., Redford, J. S., Gent, L. C., & Washburn, D. A. (2005). Visual Search and the Collapse of Categorization. *Journal of Experimental Psychology, 134*(4), 443-460,

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal detection analysis of group decision making. *Psychological Review, 108*, 193-203.

Sorkin, R. D., Kantowitz, B. H., & Kantowitz, S. C. (1988). Likelihood alarm displays. *Human Factors, 30*(4), 445-459.

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

Sparaco, P. (1995, January 30). Airbus seeks to keep pilot, new technology in harmony. *Aviation Week and Space Technology*, 62-63.

St. John, M., & Manes, D. I. (2002). Making unreliable automation useful. *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*, 332-336.

Stanton, N. S., Ragsdale, S. A., & Bustamante, E. A. (2009). The Effects of System Technology and Probability Type on Trust, Compliance, and Reliance. *In Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting*, 1368-1372. Santa Monica, CA: Human Factors and Ergonomics Society.

Wickens, C. D., & Colcombe, A. M. (2007). Dual-task performance consequences of imperfect alerting associated with a cockpit display of traffic information. *Human Factors, 49,* 839-850.

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

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: a synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201-212.

Wiegmann, D. A. (2002). Agreeing with Automated Diagnostic Aids: A Study of Users' Concurrence Strategies. *Human Factors, 44*(1), 44-50.

Wiegmann, D. A., & Cristina, F. J., Jr. (2000). Effects of feedback lag variability on the choice of an automated diagnostic aid: A preliminary predictive model. *Theoretical Issues in Ergonomic Science, 1*, 139-156.

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: the effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science*, *2*(4), 352-367.

Wiegmann, D., McCarley, J., Kramer, A., & Wickens, C. D. (2006). Age and automation interact to influence performance of a simulated luggage screening task. *Aviation, Space, and Environmental Medicine, 77*, 825-831.

Woods, D. D. (1995). The alarm problem and directed attention in dynamic fault management. *Ergonomics, 38*(11), 2371-2393.

Yeh, M., & Wickens, C. D. (2001). Display Signaling in Augmented Reality: Effects of Cue Reliability and Image Realism on Attention Allocation and Trust Calibration. *Human Factors, 43*(3), 355-365.

# APPENDIX A: FIGURES AND TABLES

Figure 1. The difference in accuracy of a diagnosis for an aid whose response criteria are separated by fewer standard deviations (left) compared to an aid whose response criteria are separated by more standard deviations (right). The red sphere represents a diagnosis made by an automated aid.

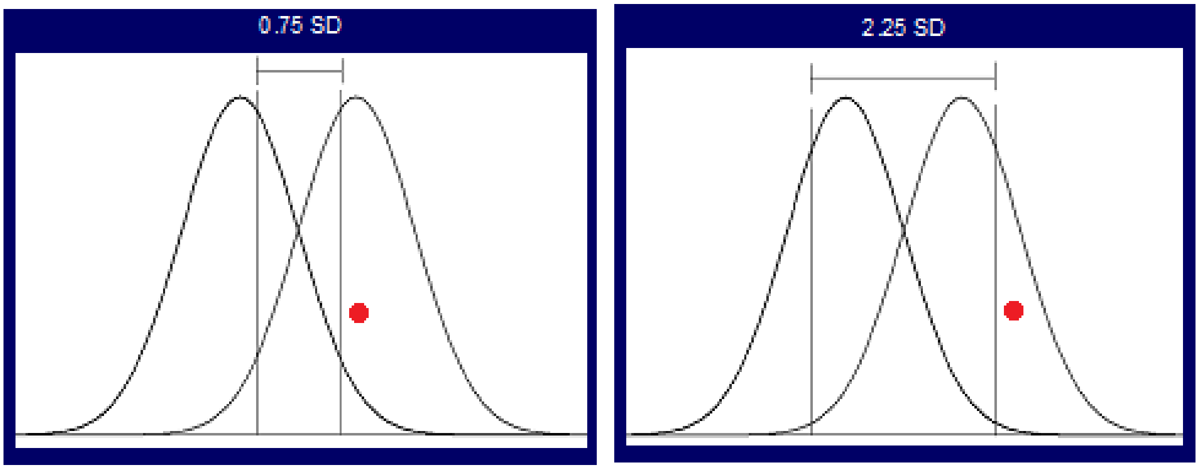


Figure 2. The experimental procedure sequence of events.

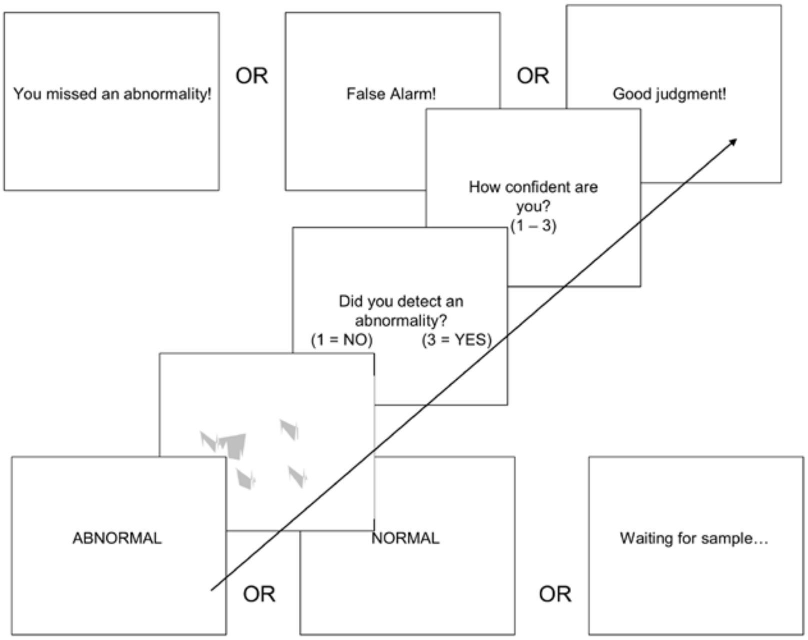

You missed an abnormality!
OR
False Alarm!
OR
Good judgment!
How confident are you? (1 – 3)
Did you detect an abnormality? (1 = NO) (3 = YES)
ABNORMAL
OR
NORMAL
OR
Waiting for sample…

Figure 3. Group means for each aided condition during assisted trials (trials 26-150).

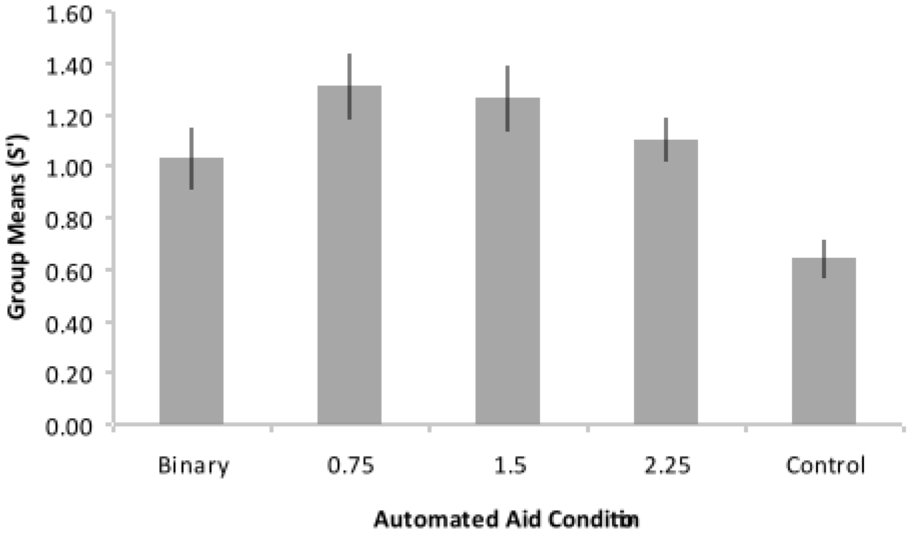
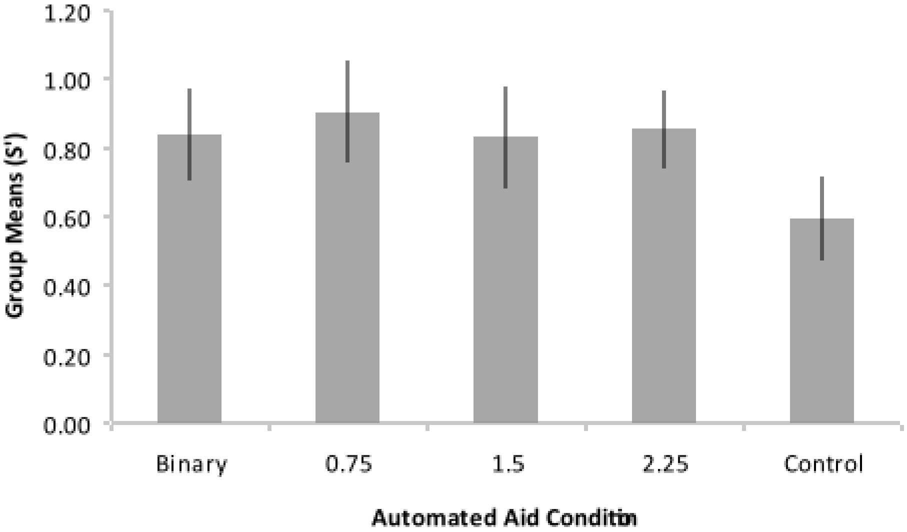


Figure 4. Group means for each aided condition during unassisted trials (trials 151-200).


Group Means (S')
Binary
Control
Automated Aid Conditon