2019

Gauging Human Performance with an Automated Aid in Low Prevalence Conditions

Cara M. Zinn
Wright State University

GAUGING HUMAN PERFORMANCE WITH AN AUTOMATED AID IN LOW PREVALENCE CONDITIONS

A Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

by

Cara M. Zinn
B.A., Pacific Lutheran University, 2016

2019
Wright State University

Wright State University
GRADUATE SCHOOL

May 2nd, 2019

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Cara M. Zinn ENTITLED Gauging Human Performance with an Automated Aid in Low Prevalence Conditions BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

Joseph Houpt, Ph.D.
Thesis Director

Scott Watamaniuk, Ph.D.
Graduate Program Director

Debra Steele-Johnson, Ph.D.
Chair, Department of Psychology

Committee on
Final Examination

Scott Watamaniuk, Ph.D.

Kevin Bennett, Ph.D.

Barry Milligan, Ph.D.
Interim Dean, Graduate School

ABSTRACT

Zinn, Cara M. M.S. , Department of Psychology, Wright State University, 2019. Gauging Human Performance with an Automated Aid in Low Prevalence Conditions.

When receiving assistance from an automated aid, human operators do not necessarily perform better than without the automated aid. The current work explored the impact of integrating the automated aid with the task information in low prevalence conditions. Specifically, this work compares displays where the automated aid was integrated with task information in general or with more meaningful task information. Subjects performed a speeded judgment task with the assistance of an automated aid, varying in display type, difficulty, and prevalence. Results indicated that there was no effect of display type or prevalence on human temporal performance, and that the effect of low target prevalence on miss rates weakened in the context of an automated aid. Automated aids could be used in real world contexts to alleviate the effects of low target or target prevalence. Designers should consider the potential utility of automated aids for low prevalence tasks in real world applications.

Contents

1 Introduction **1**
- 1.1 Display Design . . . 2
- 1.2 Target Prevalence . . . 5
- 1.3 Current Study . . . 6

2 Method **11**
- 2.1 Participants . . . 11
- 2.2 Experimental Design . . . 11
- 2.3 Task . . . 12
- 2.4 Manipulations . . . 13
  - 2.4.1 Display . . . 13
  - 2.4.2 Prevalence . . . 14
  - 2.4.3 Difficulty . . . 15
- 2.5 Measurements . . . 15
  - 2.5.1 Workload Capacity Analysis . . . 15
  - 2.5.2 Trust in the Automation Aid . . . 16
  - 2.5.3 Demographic survey . . . 16
- 2.6 Procedure . . . 16

3 Results **18**
- 3.1 Checks . . . 18
- 3.2 Performance Analysis . . . 20
- 3.3 Low Prevalence Effect . . . 21
  - 3.3.1 Criterion Shift . . . 22
  - 3.3.2 Miss Rates and mean response times . . . 22
- 3.4 TASS Results . . . 25
- 3.5 Post Hoc Analysis . . . 26

4 Discussion **27**
- 4.1 Summary of Results . . . 27
- 4.2 Implications for display designs . . . 28
  - 4.2.1 Sufficient Integration . . . 28

4.2.2 Strategies 30
4.3 Implications for Low Prevalence Effect 30
4.4 Limitations and Future Directions 32
4.5 Conclusion 33

5 References **34**

A Appendix A **37**

B Appendix B **38**

List of Figures

1.1 Time course of trials 7
1.2 Workload Capacity classifications 8
1.3 Linear Ballistic Accumulator Model 9

2.1 Display Types 14

3.1 Violin plot of participants responding before the aid 19
3.2 Capacity Coefficients 20
3.3 Miss Rates and Mean Response Time violin plots 24
3.4 Trust and Distrust scores 26

B.1 Detailed Mean Response Time Violin plot 38

List of Tables

2.1 Order Conditions 12

3.1 Bayesian ANOVA 21

Acknowledgments

I would like to take this opportunity thank to my advisor, Dr. Joseph Houpt, as well as my committee members, Drs. Kevin Bennett and Scott Watamaniuk. They have been incredibly supportive throughout this entire project. Additionally, I would like to thank Dr.Yusuke Yamani for allowing me to use his task and base my thesis on his work. I would like to thank the Mathematical Modeling of Human Performance Lab at Wright State University for all their support. Lastly, I would like to thank my family and friends for their love and support.

Introduction

Performing efficiently with an automated decision aid would benefit human performance greatly in a variety of domains, such as in medicine (e.g., Horowitz, 2017). The goal of automated aids is to provide useful information to assist and guide human operators' decision making. However, prior research has found that human operators often misuse automated decision aids when provided (e.g., Parasuraman & Riley, 1997). Some researchers have suggested that this could be due to an ineffective display design (e.g., Yamani & McCarley, 2018) that does not encourage human operators to use the automated aid. As Horowitz (2017) has suggested, there might be an effect of target prevalence on the usage of an automated aid. For example, a human operator might observe that a malfunction occurs less than 2% of the time and be more likely to assume the automated aid's warning is a false alarm. Because malfunctions in a system can have low occurrence in the real world, it is important to examine the effects of low target prevalence on human performance with an automated aid. Little research has examined the effects of low target prevalence on human performance with an aid, much less the interaction between display design and target prevalence. Thus, the purpose of my study was to examine the effects of low target prevalence and the interaction of display design and target prevalence on human performance with an automated aid.

As defined by Parasurman and Riley (1997), automation is the execution by a machine agent (usually a computer) of a function that was previously carried out by a human An automated aid is automation that assists human operators in the execution of a function,

typically in their decision-making process (Parasurman & Riley, 1997; Dzindolet et. al, 2002). The primary distinction between automation and an automated aid is that automation largely operates independent from human users and an automated aid works with the human operator. A common example of an automated aid is automated diagnostic aids or Computer-Aided Detection that assist the radiologists in finding areas of interest in medical imagery (Horowitz, 2017). Research on Computer-Aided Detection has found little evidence that these aids improve radiologists' performance in detecting areas of interest (Horowitz, 2017). There is even some evidence to suggest that these aids could potentially harm radiologist's performance (Fenton et al., 2007; Firmino et al., 2014; Horowitz, 2017).

1.1 Display Design

Yamani and McCarley (2018) investigated whether operator's suboptimal performance with an automated aid could be a result of an ineffective display design. In Yamani and McCarley's (2018) study, they asked participants to evaluate rectangular bars as either long or short as quickly and accurately as possible with the assistance of an automated aid. The automated aid was either presented on the rectangular bar, called an integrated display, or on a separate object in the display, called a separated display. Yamani and McCarley (2018) hypothesized that participants' performance with an automated aid might be dependent on how separated the automated aid's information is from the task information. In an ideal design, the automated aid's cues fully integrate with the task information, minimizing the processing of both sources of information (Wickens & Carswell, 1995; Yamani & McCarley, 2018). For the purposes of this study, the term integration will refer to the physical integration of objects or information (e.g., co-located objects), as opposed to the concept of perceptual integration (e.g. Garner, 1976)

Yamani and McCarley's (2018) hypothesis is based on Wickens and Carswell's (1995) Proximity Compatibility Principle (PCP). PCP states that the perceptual proximity of ob-

jects on a display should be in agreement with the required processing proximity to perform the task. Perceptual proximity refers to the degree that one perceives objects in parallel (Wickens & Carswell, 1995). An example of a display with higher perceptual proximity would be one that has features of the display physically integrating (e.g. using shape and color of an object to represent different features or information). Processing proximity refers to the degree that one process both sources of information in parallel to perform the task efficiently (Wickens & Carswell, 1995). In a higher processing proximity task, the subtask of processing each source of information must be integrated in order to perform the overall task efficiently. Both the perceptual proximity of the objects on the display and the processing proximity required by the task are on a continuous scale.

In the case of human-automation teaming, the task might require a higher processing proximity because human operators must process both the stimulus information and the automated aid's information to perform efficiently. If that is the case, then operators would perform better when the automation's cue integrates with the target stimulus than when it does not. Although Yamani and McCarley's original study did not come to this conclusion, Zinn, Yamani, Houpt, and Scott-Sharoni (2018) made this observation in their reanalysis of Yamani and McCarley's (2018) data.

These results did not replicate in a separate study by Zinn, Houpt, Yamani, and Scott-Sharoni (2018). In the replication, participants performed similarly with the separated and integrated displays. This could be due to insufficient physical separation between the automated aid's information and the task information in the replication. In that study, the automated aid's cue information was at maximum three degrees of visual angle away from the task information, which might not be enough to create any cost of performance in processing the aid's cues (Zinn, Houpt, Yamani & Scott-Sharoni, 2018).

In addition to the separability of the design, there might be a limitation in Yamani and McCarley's (2018) theoretical approach to the display design. As stated by Bennett and Flach (2011), it is important to consider the constraints of the task in evaluating display

designs. Constraints are the factors that limit the possible ways that a person can complete a task and are independent of the way the person chooses to complete the task. In Yamani and McCarley (2018), the task was to judge whether a bar was long or short and the constraints are the length of the current rectangular bar and the length of the bar that lies between the distributions of long and short bar lengths. There is little to no change between the designs in Yamani and McCarley (2018) and Zinn, Houpt, Yamani, and Scott-Sharoni (2018) in their ability to address the constraints of the task. The key difference between the designs in Yamani and McCarley (2018) is the accessibility of the aid. In other words, Yamani and McCarley (2018)'s results only demonstrated that a cue further from the task information would be less accessible and more likely to lead to performance decrements than a closer cue. This might explain why there was a difference in results between Yamani and McCarley (2018) and the replication done by Zinn, Houpt, Yamani, and Scott-Sharoni (2018). In Zinn, Houpt, Yamani, and Scott-Sharoni (2018), there was less of a separation between the automated aid's cue and the task information in the separated display and thus there was little change in the accessibility of the aid.

A more meaningful comparison would be of designs that either address or does not address the constraint of the task. With a design that address the constraint of the task, the automated aid would integrate with more meaningful task information as opposed to general task information. Unlike the integration in Yamani and McCarley (2018), this comparison would be moving beyond just the continuous level of physical integration of information and comparing how the automated aid integrates with the task. This comparison would look at the categorical integration of the automated aid with meaningful task information and with general task information. An assumption made with this sort of display is that the meaningful information that the automated aid is integrating with is actually meaningful for completing the task. This assumption could be rejected with further analysis of what is meaningful task information for the given task. This would require some task analysis and consulting task experts.

1.2 Target Prevalence

As mentioned earlier, the prevalence of a target might be an important factor to consider when analyzing human performance with an automated aid. Horowitz (2017) noted an important distinction between research conducted in the lab with automated aids and the real-life application of those aids (e.g., Computer-Aided Detection). In the real world, targets for Computer Aided Detection often occur less than five percent of the time. In the lab and with training automated aids, targets occur much more frequently than in some real-world cases (Horowitz, 2017). By analyzing the effect of target prevalence, the current paper might provide a better depiction of human performance with an automated aid with rare targets. Because there is little to no research done on the effect of prevalence on individuals' performance with an automated aid, the current paper looks to previous research on the effect of prevalence on individuals' unassisted performance (e.g., Peltier & Becker, 2016; Wolfe & Van Wert, 2010).

The low prevalence effect refers to when participants are more likely to miss rarely-present target than if the target was present more often (Peltier & Becker, 2016; Wolfe & Van Wert, 2010). Researchers have examined this effect in visual search tasks (e.g., Peltier & Becker, 2016), but this effect could be generalizable to other types of tasks, such as discrimination tasks (e.g., Swets, Tanner, & Birdsall, 1961). Wolfe and Van Wert's Multiple Decision Model (2010) described the causes of this low prevalence effect in visual search tasks. Their model posits that the low prevalence effect is because of a decrease in the participants' threshold to quit the search, and a shift in participants' decision criterion in low target prevalent conditions. In other words, participants are more likely to quit their search prematurely (i.e., before they fixate on all objects in the space) and are less likely to identify an object as a target (Peltier & Becker, 2016; Wolfe & Van Wert, 2010).

Previous research on the causes of the low prevalence effect informs the potential effect of low prevalence on an individuals' performance with an automated aid. For example, an individual with a low quitting threshold might not expend the effort to fully process the

automated aid's cues with the task information. A shifted decision criterion might influence how receptive an individual is to the advice of an automated aid. For example, in a low prevalence condition, an individual might be less receptive to the advice of an automated aid when the operator has a more conservative decision criterion. Because of these potential effects, creating an effective design that encourages the use of the aid in these low prevalence conditions would be essential to form an efficient team between the human and the automated aid.

1.3 Current Study

Using the same task as in Yamani and McCarley (2018), the current project explores the effect of display design and target prevalence on human usage of an automated aid. The display designs used in this study vary in the degree that the automated aid integrates with a key constraint of the task (i.e., the length of the bar that lies between the long and short distributions). Figure 1.1 shows the task for each display type. Display A is the integrated display from Yamani and McCarley (2018). Display B is a new design that includes a reference bar, representing the length of the bar that lies between the long and short distributions. To categorize performance with an automated aid, the current work will use workload capacity analysis (Houpt et al., 2014), as also used in Yamani and McCarley (2018), to gauge the efficiency of human performance with the assistance of an automated aid.

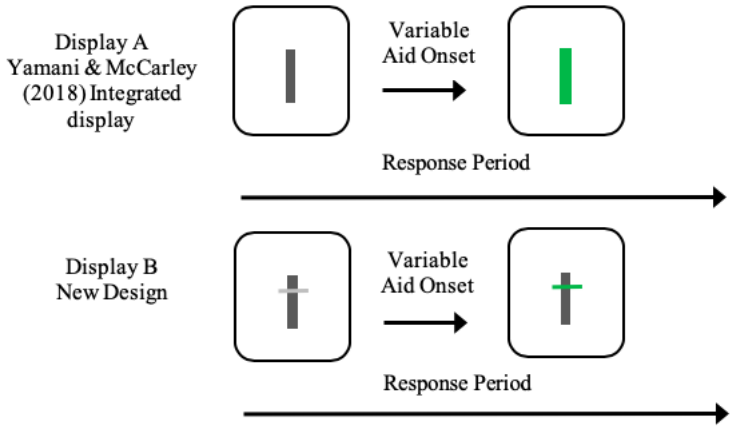


Figure 1.1: A time course of a trial with the different each display type.(Adapted from Yamani & McCarley, 2018) Participants are presented with the vertical bar and after some variable onset, the automated aid appears. Participants respond at any point after the onset of the vertical bar.

Workload capacity analysis is a part of the Systems Factorial Technology framework, which describes how one processes multiple channels of information (Townsend & Nozawa, 1995). Workload capacity analysis describes the temporal efficiency with which one processes multiple channels of information compared to processing the channels independently, using both response times and accuracy (Houpt et al., 2014). In the case with this study, workload capacity analysis can analyze the temporal efficiency of having an automated aid compared to not having the automated aid (Yamani & McCarley, 2018; Zinn, Yamani, Houpt, & Scott-Sharoni, 2018).

Participants' performance with the automated aid was compared to their performance without the automated aid, also known as their baseline performance in the task. A participant is performing at limited capacity when his/her performance is worse than his/her baseline. A participant is performing at unlimited capacity when his/her performance is equal to his/her baseline. A participant is performing at super capacity when his/her performance is better with the automated aid than his/her baseline. For the purposes of this study, I used a nonparametric and parametric measure of workload capacity. Using both measurements will allow for consistency in analysis with Yamani and McCarley (2018) as well as explore other parameters that may influence performance.

The nonparametric measure of workload capacity is the single-target self-terminating capacity coefficient (C_{STST}). The C_{STST} is a ratio of the cumulative reverse hazard functions, *K*, at time, *t*, from the aided and unaided trials. The cumulative reverse hazard function is a transformation of the response time distribution of those trials. The equation for the C_{STST} is :

$$C_{STST}(t) = \frac{K_{unaided}(t)}{K_{aided}(t)}$$

Using this ratio, I can make classifications as to how the participant is performing with the automated aid. Workload capacity classifications are shown in Figure 1.2. The benefit to using the capacity coefficient instead of standard analyses of temporal performance, like mean response-time comparisons, is that workload capacity analysis uses entire response time (RT) distributions. By using the entire RT distribution, workload capacity analysis avoids potentially misleading or ambiguous conclusions about a system's performance (Eidels et al., 2010).

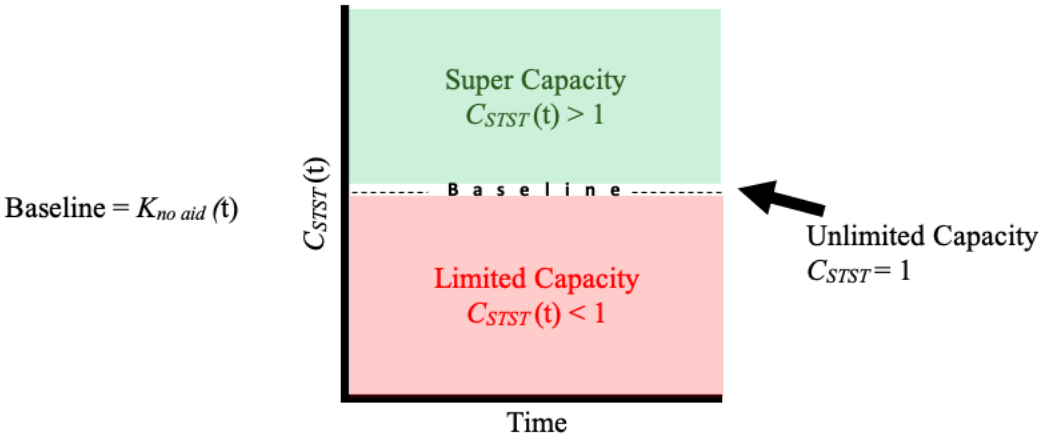


Figure 1.2: The zones for characterizing a system in terms of temporal efficiency, using the Single Target Self-Terminating (STST) Capacity Coefficient.

The parametric measure of workload capacity used the linear ballistic evidence accumulator (LBA) model, as shown in Figure 1.3 (Eidels et al., 2010). This allows changes in bias to be measured in addition to performance in a speeded task. Observing bias helped

determine whether there is an effect of prevalence on individual's bias against responding to the rare target, as found in previous literature. Briefly, an LBA model assumes that evidence accumulates from an initial point, A, linearly at a speed given by a drift rate drawn from a normal distribution with a mean, v, and standard deviation, s, until it exceeds a threshold, b. Base time, or the time it takes for early perceptual processing, response selection and execution, is treated as a constant, t_0. The workload capacity analysis uses the draft rates, comparing the relative magnitudes of drift rates from both aided, v_A, and unaided trials, v_{UA}. If v_A is greater than v_{UA} then it is a super capacity system. If v_A is less than v_{UA} then it is a limited capacity system. If v_A is the same as v_{UA} then it is an unlimited capacity system (Eidels et al., 2010).

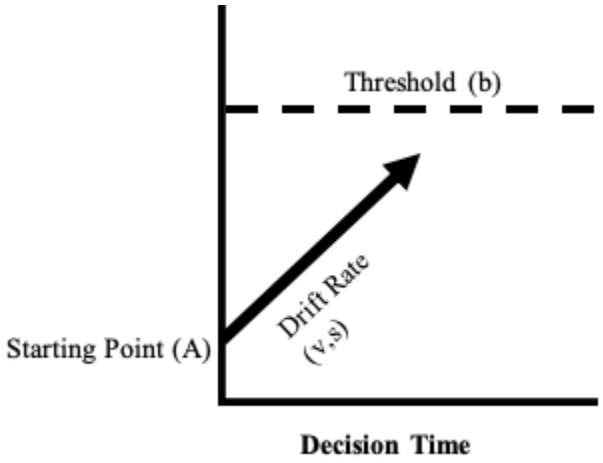


Figure 1.3: A single linear ballistic accumulator with parameters labeled.

In summary, exploring the effect of target prevalence is important to capture a more realistic depiction of individuals' performance with an automated aid in low target prevalence conditions. The current work measured human performance with an automated aid with varying display types, target prevalence, and task difficulty. Including task difficulty helped to improve the fit of the LBA parameters. I hypothesized that participants would perform more efficiently using the display that integrates the automated aid's information with meaningful information (Display B) than one that does not (Display A). Additionally, I predicted that participants will perform more efficiently with Display B than Display A

in the difficult condition and in the low prevalence condition. This would be because the meaningful integration would be beneficial in situations where there is more uncertainty in accurately discriminating the bar length categories, as would be expected in the low prevalence condition and difficult condition.

Method

2.1 Participants

Sixty-two undergraduate students from an introductory psychology course at a mid-sized Midwestern university (Age: M = 20.40 years, SD = 4.59; 6 subjects excluded; 56 participants' data used; 37 female) participated in this study. Participants received course credit for their participation. A power analysis on pilot data (n = 4) indicated that a sample size of 60 participants would be sufficient to detect a medium effect of (d = .4) with 95% power at a statistical significance level of .05. Participants were excluded for low accuracy (less than 55% accurate) and incomplete data.

2.2 Experimental Design

This study used a 2x2x2 mixed design (Display Type x Prevalence x Task Difficulty). Participants saw all combinations of display types (Display A and Display B) and task difficulty conditions (Easy and Difficult) as a within subjects treatment. The order of conditions was counterbalanced across participants. The motivation behind counterbalancing is to ensure that there is no confound of ordering on their performance with each display type. Participants received one order of difficulty (i.e., Easy then Difficult or Difficult then Easy) and one order of condition (i.e., Display A then Display B or Display B then Display A).

This ordering system was for simplicity in creating the orders and to ensure that participants only needed to receive descriptions on the display once before starting that half of the experiment, as opposed to multiple times throughout the experiment. Available ordering conditions are shown in Table 2.1. Each participant experienced only one level of target prevalence (Low or Equal), which is a between-subjects treatment. This was to avoid potential carry-over effects of bias into the different prevalence conditions.

1st Condition	2nd Condition	3rd Condition	4th Condition
Display A (Easy)	Display A (Difficult)	Display B (Easy)	Display B (Difficult)
Display A (Difficult)	Display A(Easy)	Display B (Difficult)	Display B (Easy)
Display B (Easy)	Display B (Difficult)	Display A (Easy)	Display A (Difficult)
Display B (Difficult)	Display B (Easy)	Display A (Difficult)	Display A (Easy)

Table 2.1: Possible ordering conditions

2.3 Task

The task for this study was consistent with Yamani and McCarley's (2018) task. Participants made speeded judgments on whether to attempt production based on the amount of raw materials provided. The length of the vertical bar on the display represented the amount of raw materials provided. Participants were to accept shorter bars for production and reject longer bars as quickly and accurately as possible. Each trial began with the onset of the vertical bar. I sampled the length of the bar for each trial from a Gaussian distribution with mean of either 2.2 degrees of visual angle and 2.8 degrees of visual angle for short and long bars respectively. The standard deviation of the Gaussian distribution and sampling proportion of short and long bar lengths varied, depending on condition (see next section). Participants responded whether they would attempt production as quickly and accurately as possible. Following participants' responses, participants received feedback on their judgments. I placed the bar randomly on the display with the center of the bar placed either one

degree up or down and either one degree left or right from the center of the display. This was to prevent participants from directly comparing the current bar to the previous bar.

Participants saw all four combinations of display type and difficulty, in varying orders. For all conditions, they completed the task with and without the assistance of an automated aid. Participants had the automated aid on alternating blocks of trials, either starting on an unaided block or an aided block. The motivation for having the automated aid on alternating blocks was to ensure that there was no effect of order on whether all the automated aid trials occurred before or after the unaided trials. For trials with the automated aid, participants saw a color cue to represent the aid's suggestion. Participants saw a dark green cue for a short bar recommendation and a bright red cue for a long bar recommendation. A sample from an exponential distribution with a mean of 676 milliseconds determined when the onset of the cue occurred each trial. This was to prevent participants from anticipating the automated aid's cue (Yamani & McCarley, 2018). The aid's reliability was at 95% to encourage the use of the aid while simulating imperfection.

2.4 Manipulations

2.4.1 Display

For the aided conditions, participants saw the automated aid's cue on the stimulus rectangle itself (Display A) or on a separate horizontal rectangular indicator on top of the stimulus rectangle (Display B). The displays are presented in Figure 4. This indicator represented the middle point between the means of the short and long bar distributions (2.5 degrees in length). On each trial, the indicator was 2.5 degrees from the bottom of the vertical bar and moved with the vertical bar between trials. To check that performance is due to the presence of the automated aid on the separate indicator and not due to the indicator itself, participants completed the task with the indicator alone in the unaided trials of the Display

B conditions.

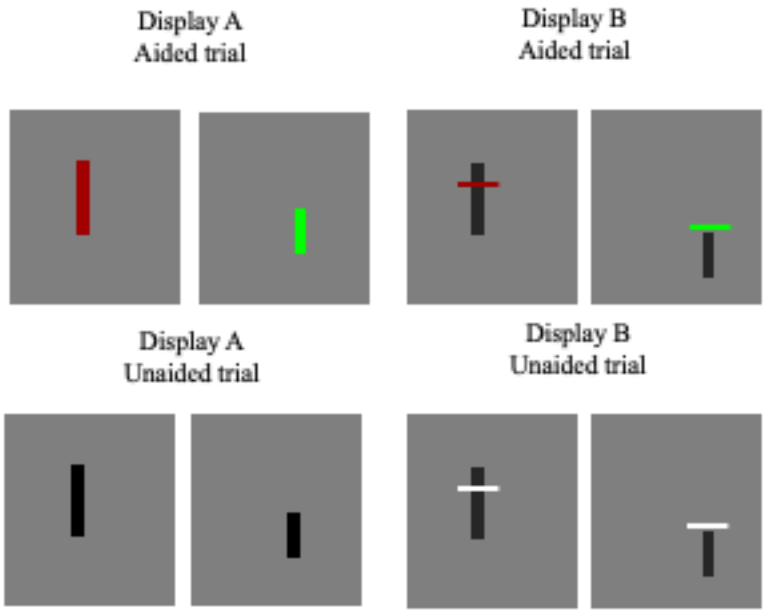


Figure 2.1: All aided and unaided display types for each long and short bar conditions. Note that the reference bar is present in the unaided version of Display B.

If the participant relied solely on the reference bar from Display B (e.g., responded short if it was below the bar and long if it was above the bar), the participant would achieve 96% accuracy in the easy condition, and 68% in the difficult condition. If they relied on the reference bar and the automated aid in combination, the participant would achieve 99.8% accuracy for the easy condition and 98.5% accuracy for the difficult condition.

2.4.2 Prevalence

The sampling frequency of long bar lengths varied depending on the prevalence condition (Low or Equal). For a low prevalence condition, a sample from the long bar distribution occurred around 10% of the time. In an equal prevalence condition, a sample from the long bar distribution occurred around 50% of the time. Because of the random sampling of the bar lengths, participants might not see long bars at exactly 10 percent or 50 percent

of the time. To ensure participants experience a prevalence close enough to the desired prevalence, participants completed at least 800 trials.

2.4.3 Difficulty

For each trial, bar lengths are sampled from a Gaussian Distribution with either a mean at 2.2 for short bars and 2.8 for long bars. The standard deviation of both distributions determined the amount of overlap between the short and long bar lengths. With a larger overlap between the long and short bar length distributions, it is more difficult to distinguish long and short bar lengths. Therefore, I set the standard deviation of the bar length distributions to be larger for the difficult trials ($SD = 0.3$) than the easy trials ($SD = 0.15$).

2.5 Measurements

I collected participants' response times and whether the participant was correct for each trial. Additional information I collected was the onset of the automated aid, the judgment of the automated aid, and the length of the vertical bar for each trial.

2.5.1 Workload Capacity Analysis

To measure the participants' performance with the automated aid, I used two different measures of workload capacity. The first measure is using the difference between fitted drift rates from a Linear Ballistic Accumulator model in the aided and unaided conditions. For the LBA model, starting point, threshold, and drift rate were free parameters across the within subject conditions (Display type x Task Difficulty x Presence of aid). Each participant had a total of 26 parameters (8 x drift rate, 8 x threshold, 8 x starting point, base time, and standard deviation) for each decision type (long or short stimuli and long or short responses). The second measure is using capacity coefficients, which also uses response

times and accuracy.

2.5.2 Trust in the Automation Aid

In addition, participants completed the Trust in Automated Systems Scale (TASS, Jian, Bisantz, & Drury, 2000) to evaluate participants' trust and distrust in the automated aid. This questionnaire provided information about the participants' attitudes towards the automated aid, which might influence participants' usage of the automated aid. Results from this scale helped to rule out potential alternative hypotheses that could also explain some of the results of the experiment (e.g. that an individual is underperforming with the automated aid because of his/her distrust in the aid as opposed to the manipulations). This scale contained 12 items (total α = .72) which measured both trust in automation (7 items; α = .93) and distrust in automated (5 items; α = .83) (Dolgov & Kaltenbach, 2017). Participants rated the degree to which they agreed with the provided statements on a scale from 1 (not at all) to 7 (extremely). An example statement from the measurement of trust was " The system provides security " and in the measurement of distrust was "The system is deceptive" (see Appendix A for full survey). Average scores of the trust and mistrust items created the trust and mistrust measurements.

2.5.3 Demographic survey

Participants completed a demographic survey that collected age, gender, race/ethnicity, primary language, and class standing.

2.6 Procedure

Participants first completed the demographic survey with the informed consent. Participants then completed a training session and an experimental session on two separate days

(about 2 days apart; max 10 days apart). In the training session, participants completed 16 blocks of 50 trials (800 trials in total) without the automated aid. The goal of the training session was to allow participants to practice the task and learn the short and long bar length and prevalence. In the experimental session, participants completed 32 blocks of 50 trials (1600 trials in total). On alternating blocks in the experimental session, participants received cues from an automated aid to assist them. Following the experimental session, participants completed the TASS questionnaire.

Results

I analyzed workload capacity scores (drift-rates and capacity coefficients) in R (R Core Team, 2019) with Bayesian analyses from the BayesFactor package (Rouder & Morey, 2012; Rouder, Morey, Speckman & Province, 2012). The benefit of using Bayes Factors (BF) as opposed to traditional NHST testing is that Bayes Factors can represent evidence in support of the null hypothesis or the alternative hypothesis. Labels provided in Jeffreys (1961) informed the labels used in this study. Weak evidence in support of the alternative hypothesis is a BF = 1-3 (BF = 0.3 - 1 for the null hypothesis). Moderate evidence in support of the alternative hypothesis is a BF = 3-10 (BF = 0.1 - 0.3 for null hypothesis). Strong evidence in support of the alternative hypothesis is a BF $>$ 10 (BF $<$ 0.1 for null hypothesis).

3.1 Checks

To ensure that there was no effect of the ordering of conditions, I compared participants' capacity coefficients based on their order condition, using a Bayesian t-test. The results indicated that there was moderate evidence against an effect of order on performance, BF = 0.11.

Because the automation's cues had different luminance levels (e.g., dark green or bright red), I compared participants' drift rates between cue colors using a Bayesian t-test. This was to check whether participants processed the two cue types differently. I refit

a subset of the data to the LBA model that only included trials with one display type (i.e., Display A), one difficulty level (i.e., Easy), and when the automation was present (200 trials per participant). Drift rate was the only parameter free between trials with green and red cues. There was weak evidence against the effect of luminance and colors differences between the cue colors, BF = 0.62.

Because participants could respond prior to the automated aid in the aided trials, I checked how often participants responded before the automated aid. Across all conditions, participants responded before the automated aid on average 94.83 rows (*SD* = 36.00). Figure 3.1 shows the density of the number of trials where participants responded before the automated aid by condition. Based on the plots and on the results of a Bayesian ANOVA, there is strong evidence to suggest an effect of prevalence on the number of trials a participant responds before the automated aid, BF $>$ 100.

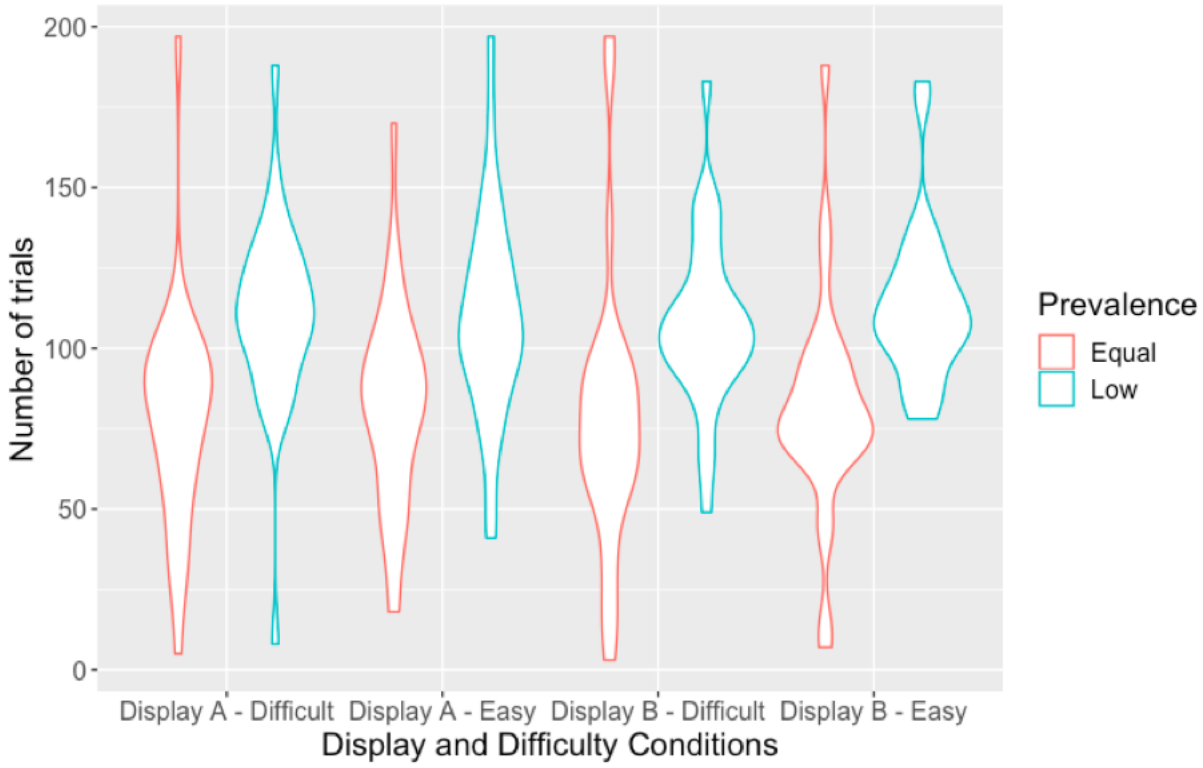


Figure 3.1: Violin plots of trials participants responded before the automated aid per condition.

3.2 Performance Analysis

Visual inspection of the group level' capacity coefficients (shown in Figure 3.2), revealed that participants performed at all different levels of capacity (super, unlimited and limited). Most participants performed at unlimited capacity. This means that most participants performed similarly for aided and unaided trials. The quality of the fit for the LBA parameters on participants' data varied, resulting in less reliable parameter estimates. Potential reasons why there could have been poor fit for some of the participants were that some participants had bi-modal response time distributions, and poor starting parameter values. As a result, my interpretations relied more on capacity coefficients than drift rate estimations.

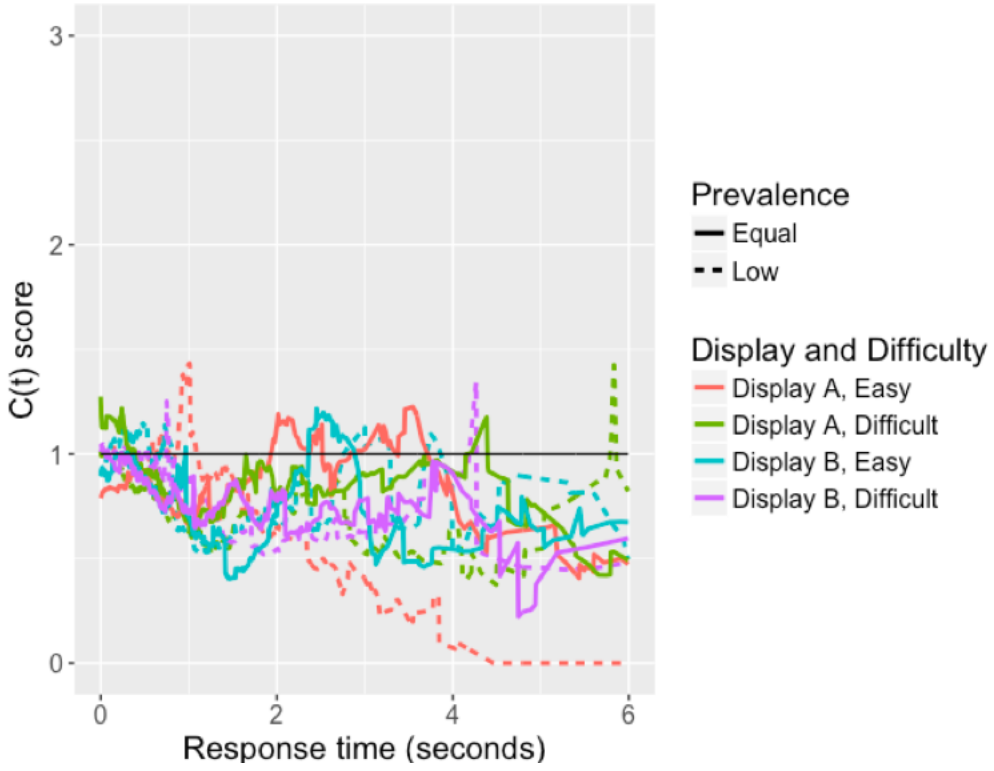


Figure 3.2: Group Level capacity coefficients over time for each condition.

To measure whether there were differences in participants' performance with the automated aid across display, difficulty, and prevalence conditions, I ran a 3-way Bayesian ANOVA for both drift rates and capacity coefficient z scores, C_z. Bayes Factors from this

ANOVA are displayed in Table 3.1.

	Drift Rate BF	C_z BF	Interpretation
Display	0.15	0.28	Moderately against
Difficulty	0.11	0.18	Moderately against
Prevalence	0.14	0.71	Moderately to weakly against
Display x Prevalence	< 0.01	0.16	Strongly to moderately against
Display x Difficulty	< 0.01	0.01	Strongly against
Prevalence x Difficulty	< 0.01	0.04	Strongly against
Display x Difficulty x Prevalence	< 0.01	< 0.01	Strongly against

Table 3.1: Results of three-way Bayesian ANOVA on drift rates and capacity z scores (C_z) with interpretations

There was moderate evidence against an effect of display type on drift rates, BF = 0.15, and capacity coefficients, BF = 0.28. This contradicted my hypothesis that participants would perform more efficiently with the meaningful display as opposed to the general display. There was strong evidence against an effect of display and difficulty conditions on drift rates, BF $<$ 0.01, and capacity coefficients, BF = 0.01. This contradicted my hypothesis that participants would perform better with Display B in the difficult condition. There was moderate to strong evidence against an effect of display types and prevalence on drift rates, BF $<$ 0.01, and capacity coefficients, BF = 0.16. This contradicted my hypothesis that participants would perform more efficiently with Display B than Display A in conditions with more uncertainty.

3.3 Low Prevalence Effect

I compared the bias or distance between the starting point and threshold for participants with low prevalence and equal prevalence conditions using a Bayesian t-test. This was to see if there was a change in bias based on prevalence, as would be expected based on the previous prevalence literature. There was moderate evidence to suggest there was

no change in bias between low prevalence (M = 7466852.0, SD =51269458.0) and equal prevalence (M = 559031.2, SD = 5096025.0) conditions, BF = 0.17. This could suggest that there might have not been a strong manipulation of target prevalence. To further investigate this, I evaluated whether the low prevalence effect was present in this study similar to previous literature. Specifically, I evaluated whether there was a shift in participant's criterion between prevalence conditions and whether there was higher miss rates in the low prevalence condition than the equal prevalence condition. Additionally, I looked at overall mean response time differences across prevalence conditions.

3.3.1 Criterion Shift

To determine whether there was a shift in the participants' criterion, I analyzed participant's data with signal detection analysis. I found that there is evidence in support of an effect of prevalence on participant's criterion of distinguishing long and short bars (BF $>$ 100). Participants of the equal prevalence condition (M = 1.10, SD = 0.41) had a lower beta than the beta for the low prevalence condition (M = 12.40, SD = 12.83). A model of only prevalence was the best model compared to models that included conditions of display and presence of the automated aid.

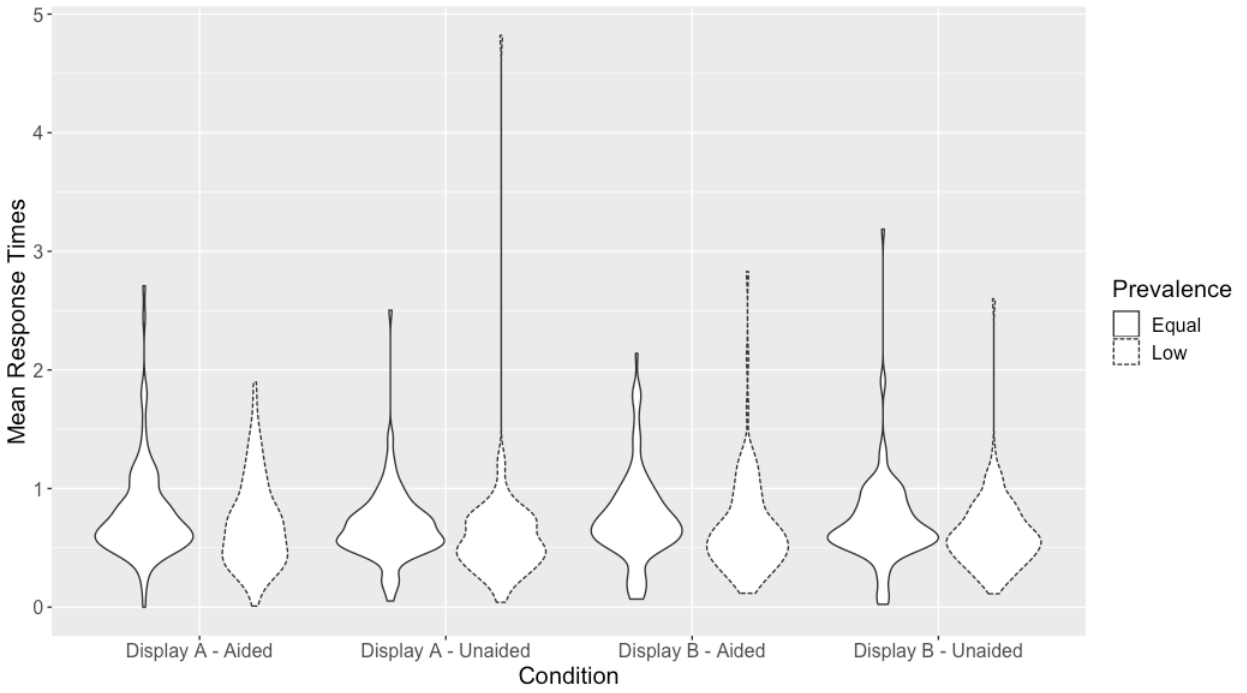
3.3.2 Miss Rates and mean response times

There was strong evidence to support an effect of prevalence on mean-response times, BF $>$ 100, and on participant's miss rates, BF $>$ 100. Specifically, the low prevalence condition had a faster mean response time (M = 0.50, SD = 0.18), and a higher miss rate (M = 0.58, SD = 0.25) than the equal prevalence condition (Mean RT: M = 0.73, SD = 0.29; Miss Rate: M = 0.17, SD = 0.09). This provided evidence in favor of a sufficient manipulation of prevalence. Additionally, the faster mean response time in the low prevalence condition explains why participants in the low prevalence condition responded before the automated

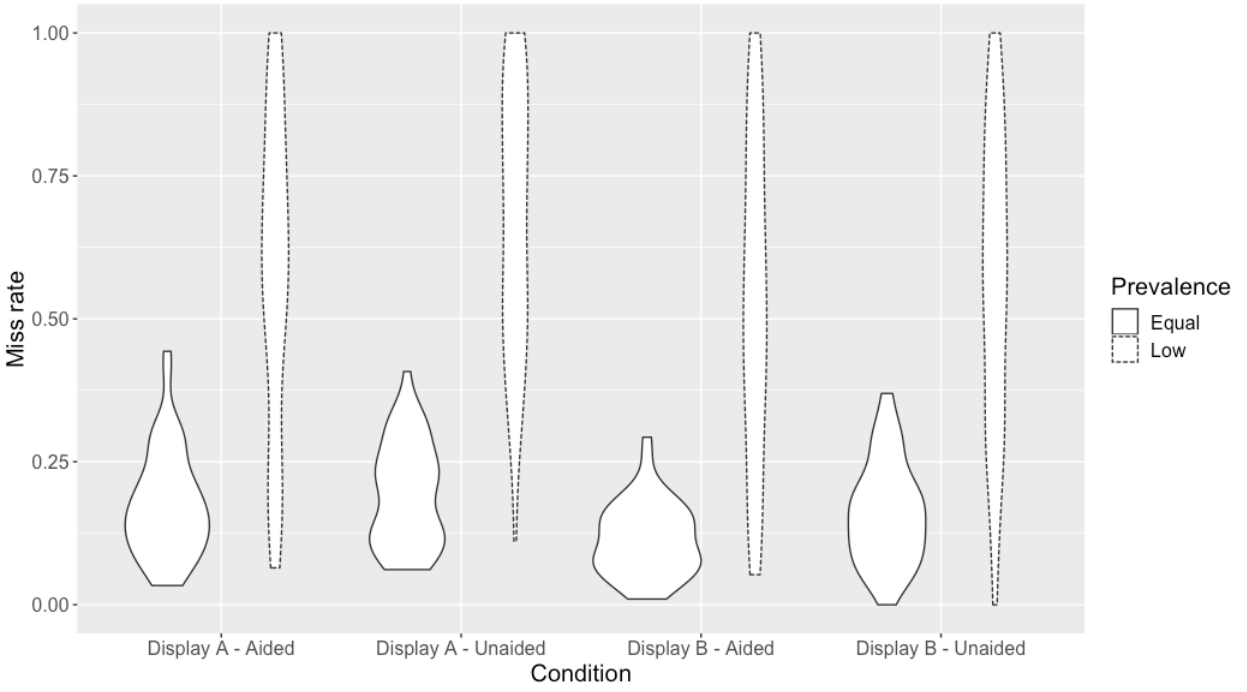
aid more often than participants in the equal prevalence condition.

To determine whether there was a change in the low prevalence effect when the automated aid was present or not present, I observed the interaction between the effect of the automated aid and prevalence. There was strong evidence in favor of an effect of the automated aid and prevalence on mean response times, BF $>$ 100, and participant's miss rates BF $>$ 100. For the low prevalence condition, participants had slower mean response times when aided (M = 0.53, SD = 0.20) as opposed to unaided (M = 0.48, SD = 0.15). Participants in the low prevalence condition had a lower miss rate when aided (M = 0.55, SD = 0.27) as opposed to unaided (M = 0.61, SD = 0.25). For the equal prevalence condition, participants had slower response times when aided (M = 0.77, SD = 0.32) as opposed to unaided (M = 0.69, SD = 0.24). Participants in the low prevalence condition had a lower miss rate when aided (M = 0.15, SD = 0.09) as opposed to unaided (M = 0.19, SD = 0.10).

Lastly, I observed whether there was a difference in the low prevalence effect in terms of display type and the automated aid. There was strong evidence in favor of an interaction between prevalence, automated aid, and display type in terms of miss rates, BF $>$ 100 and mean response times, BF $>$ 100. Figure 3.3 shows the density distributions of mean response times (a) and miss rates (b) across aided and unaided trials, for low and equal prevalence.



(a) First sub-figure



(b) Second sub-figure

Figure 3.3: Miss rate (b) and Mean response times (a) across conditions.

In the unaided trials, participants in the low prevalence condition had a slightly higher miss rate with Display A (M = 0.67, SD = 0.23) and slightly lower response times with

Display A (M = 0.41, SD = 0.13) than Display B (Mean RT: M = 0.45, SD = 0.13; Miss rates: M = 0.56, SD = 0.26). Participants in the equal prevalence condition had a slightly higher miss rate with Display A (M = 0.19, SD = 0.09) and no difference in response times between Display A (M = 0.66, SD = 0.25) and Display B (Mean RT : M = 0.66, SD = 0.22; Miss rate: M = 0.16, SD = 0.09). In the aided trials, participants in the low prevalence condition had a slightly higher miss rate with Display A (M = 0.58, SD = 0.27) and no difference in response times between Display A (M = 0.48, SD = 0.18) and Display B (Mean RT: M = 0.48, SD = 0.15; Miss Rate: M = 0.53, SD = 0.27). Participants in the equal prevalence condition had a slightly higher miss rate with Display A (M = 0.17, SD = 0.06) and slightly higher response time with Display A (M = 0.77, SD = 0.33) than Display B (Mean RT : M = 0.74, SD = 0.31; Miss Rate: M = 0.11, SD = 0.09). A more detailed plot of the mean response times by condition is in Appendix B.

3.4 TASS Results

To determine whether participants' trust or distrust in the automated aid influenced their performance with the aid, I regressed participants' trust and distrust scores from the TASS survey on their performance with the automated aid. Figure 3.4 shows the relationship between trust and distrust scores between participants. Results indicated that participant's trust (M = 3.13, SD = 1.10) and distrust (M = 3.79, SD = 1.01) were predictive of their performance with the automated aid (BF $>$ 100). There was weak evidence to suggest that participants distrust alone was particularly influential when compared to the full model of trust and distrust, BF = 0.27. However, when I added distrust to the full model with display, difficulty, and prevalence there was strong evidence against this model, BF $>$ 0.01, showing evidence in favor of a model of just display, difficulty and prevalence.

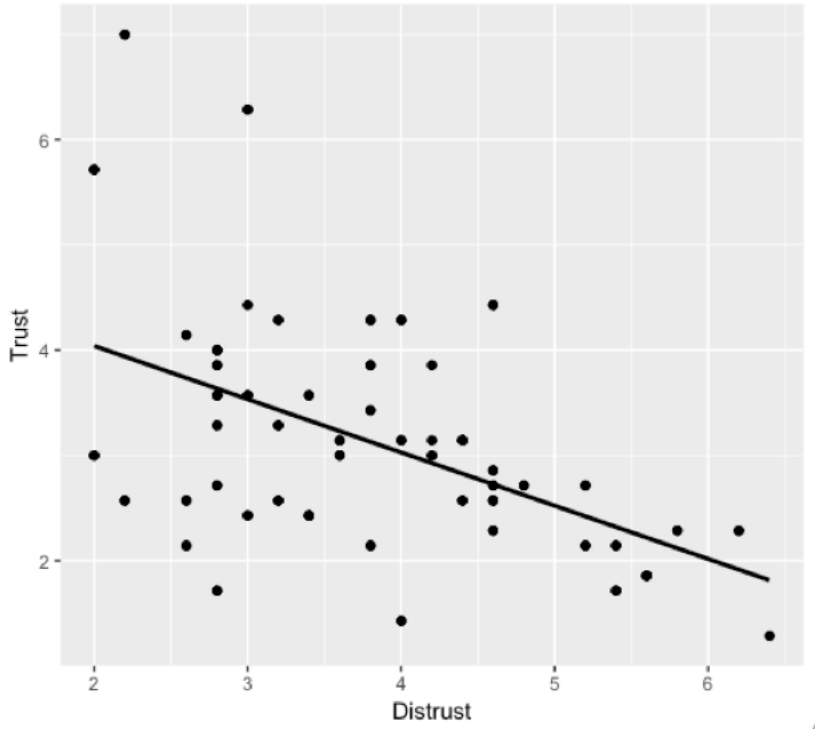


Figure 3.4: Shows each individual's score of trust by distrust, with a line of best fit to show the direction of the relationship between trust and distrust.

3.5 Post Hoc Analysis

One may be interested in whether participants even received any bonus in performance with the reference bar. This would inform us on whether this information is actually useful and people are using the information. To examine this, I looked at whether there was a difference in terms of accuracy and mean response times between the two display types for the unaided trials only. I found that there was evidence against a difference between the Display A (Mean RT: $M = 0.54$, $SD = 0.27$; Accuracy: $M = 0.84$, $SD = 0.12$) and Display B (Mean RT: $M = 0.56$, $SD = 0.22$; Accuracy: $M = 0.85$, $SD = 0.14$) in terms of both mean response times and accuracy. This might suggest that our assumptions about this information may be wrong.

Discussion

The purpose of this study was to examine the effects of display design and target prevalence on human performance with an automated aid. I found that there was little to no evidence in support of an effect of display design or a main effect of prevalence on human performance. Upon further examination, I found that participant's miss rates were lowest in the low prevalence condition when participants were provided with an automated aid on a meaningful display. This could suggest that the low prevalence effect could be alleviated in the context of an automated aid with a meaningful integration. These results provided insights on the relationship between target prevalence and automated aids. For example, the implementation of an automated aid could weaken the effects of low prevalence in the real world. In terms of display design, these results indicated that integrating the automated aid with more meaningful task information might not significantly improve performance with the aid.

4.1 Summary of Results

Most participants operated at unlimited capacity, meaning they performed similarly between aided and unaided trials. However, there was wide variability in capacity functions between participants, which could indicate different strategies between participants. There was little to no evidence to suggest that participants performed differently between the two display types or the two prevalence levels, which contradicted my hypotheses. The effect of

target prevalence was present in the unaided trials but was weaker in the aided trials (e.g., lower miss rates). The effect of prevalence was weakest when participants used Display B and had an automated aid. This suggests that the effect of prevalence might weaken with the implementation of the automated aid with a display design that integrates the automated aid with meaningful task information. Lastly, I found that trust and distrust were predictive of human performance with the automated aid, which is consistent with prior research on trust in automation (e.g., Lee & See, 2004)

4.2 Implications for display designs

The current findings indicate that the integration of an automated aid's cue with a meaningful feature (e.g., the reference bar) of the display did not significantly improve or weaken performance. This is not consistent with the conclusions drawn from Zinn, Yamani, Houpt, and Scott-Sharoni (2018), who suggested that human performance would benefit from a display with a more integrated design. There are two explanations regarding display design that arise from the results of the study regarding whether there is only a need for a sufficient integration and the role of strategies in determining the effectiveness of the display.

4.2.1 Sufficient Integration

The little to no difference in performance between display types could indicate that integration with meaningful task information may not yield any benefits to performance. The automated aid's cues may only need to be sufficiently integrated with general task information in order to have any performance benefits and that further specificity in the placement of the automated aid may not yield additional benefits. In other words, a display only needs to sufficiently integrate the automated aid with the task information to be effective. In this study, both display designs could have had sufficient integration of the automated

aid, resulting in little difference between performance in the display designs. In terms of Proximity Compatibility Principle (PCP) discussed earlier, these results suggests that information in a divided attention task may only need to be generally perceptually proximal to the task information. The specificity in the placement of the automated aid's information may not be relevant.

However, PCP does not address the role of context on human performance with an automated aid, like a triadic approach would (Bennett & Flach, 2011). The context in the case of this study can include the underlying reality of the task. Because the automated aid's information does not come from the vertical bar, but from an independent source of information, this suggests that there is an underlying true reality that informs both the vertical bar and the automated aid on the display. The mapping between this reality and the display is important for performing well with the task. It could be that there is a similar mapping of the underlying reality of the task with the display between both display types. This would explain why participants are not necessarily performing differently between the two display types, because both displays are not fully addressing this mapping of the underlying reality of the task.

It is possible that the reference bar was not meaningful information for the task. This would explain why there was no effect of display type on the unaided conditions. There was evidence against participants performing differently between the display designs, suggesting that the reference bar may not be useful. It could also be the case, that participants did not understand what the reference bar meant or how they may use it. Further replications of this study would need to pay close attention to how the reference bar is presented to the participants and confirm whether the participants understand the meaning behind the reference bar. Additionally, future research should further explore the mapping of the underlying reality of the task on the display type and how that may influence performance with the automated aid.

4.2.2 Strategies

As mentioned earlier, participants varied in their workload capacity measures. This wide variability in performance could be indicative of different strategies, which might compliment one display over another. As discussed in Meyer (2001), participants who use the automated aid more might benefit more from a more meaningful display than general display, because the automated aid is more valuable in the context of the reference bar. Participants who decided to ignore the automated aid more, might perform better with a more general display, because the automated aid is not on a key piece of information and could be more easily ignored. With participants adopting a mix of strategies, it becomes difficult to determine the effect of display design on performance. Because the current study did not evaluate all possible strategies in the tasks, I cannot determine whether participants had different strategies or whether there was an effect of strategy on performance. Future research should identify potential strategies and explore the relationship between strategies and performance with different display types.

4.3 Implications for Low Prevalence Effect

There are serious consequences of the low prevalence effect in the real world, from a radiologist missing a cancer diagnosis to security letting dangerous objects through their checkpoint. Because the low prevalence effect has serious consequences in the real world, researchers have been interested in methods of alleviating this effect. Horowitz (2017) reviewed the main avenues that researchers have explored to alleviate the low prevalence effect. Primarily researchers have been interested in manipulating participants' feedback, providing bursts of high prevalence trials, and manipulating the payoff matrix of their responses (Horowitz, 2017). Researchers have manipulated feedback by showing participants false feedback to represent a higher prevalence than what is true (e.g., providing feedback that represents 50% target prevalence as opposed to the true 20% prevalence). In the bursts

method, participants experience bursts of high prevalence trials in between low prevalence conditions. Finally, researchers have tried to manipulate the payoff participants receive when hitting and missing targets to encourage participants to hold a more liberal criterion (e.g., participants may be more willing to say target present than target absent if the payoff matrix indicates that hits are 100 points and misses are -900 points). From these methods, researchers found that the low prevalence effect was weaker but not eliminated entirely (Horowitz, 2017).

The present research provided evidence for a less explored avenue of alleviating the low prevalence: the use of an automated aid. As demonstrated in this study, the low prevalence effect might weaken when participants have an automated aid to assist them in their decisions. Moreover, the low prevalence effect may weaken when the automated aid is integrated with the constraints of the task in the display. As shown in the results, participants had the lowest miss rates in the low prevalence condition with the reference bar (Display B) than without (Display A) in the aided condition.

Some research has focused on the use of cues to alleviate the effects of prevalence with similar results. Russell and Kunar (2012) investigated the use of attentional cueing to weaken the effect of prevalence in a visual research task. Though Russell and Kunar (2012) found that participants performed better with the cues, they still observed an effect of prevalence. The findings of this study and of Russell and Kunar's (2012) study are not unlike previous findings using other methods (e.g., manipulating feedback, the payoff matrix or providing bursts of high prevalence). The low prevalence effect has been stubborn and has persisted even with these solutions. However, unlike the previous methods, using an automated aid is more practical for real world situations. Providing false feedback, or bursts of high prevalent targets, is not feasible for radiologists or airport security. Moreover, manipulating the payoffs for an individuals' decisions might have adverse effects in the real world (e.g., radiologists might start over-diagnosing patients). Providing an automated aid with a more effective design is more feasible and beneficial for real-world situations, and

future research should explore how automated aides can further weaken the low prevalence effect.

4.4 Limitations and Future Directions

There are several limitations to this study. First, the automated aid's cue in Display A covers a larger space (the entire bar) than in Display B that only covers the smaller reference bar. Because of this difference in presented size of the cue, there may be a difference in salience of the cue. This difference in salience might negate the potential beneficial effects of Display B. One way to test the salience between the two display designs would be to compare the response times of participants responding to the cues only (e.g. responding whether the automated aid says long or short instead of the participants responding whether the bar is short or long). If participants are faster at responding to the automated aid with Display A than Display B, then we might determine Display A to be more salient than Display B. Future designs for this research should control for cue salience across both designs.

Also, the task in this study is simple, making it difficult to generalize to real world applications. The task might be too simple for participants to need the automated aid to perform well. The automated aid might not be valuable for participants, and this may explain why this study found no difference between the conditions. Because performance is evaluated in terms of the automated aid, a useless automated aid could explain why we did not find any performance differences between the conditions. For future research, I hope to replicate these designs with a more complex task that encourages participants to rely more on the automated aid, and further evaluate the value of the automated aid in the task (i.e. would participants received a substantial performance benefit from using the automated aid).

As mentioned before, I intend to further explore the relationship between strategies

and display design. I would like to compare displays that integrate with a single source of information as opposed to multiple sources. This would address whether placement of the automated aid's cue could influence strategy selection. Lastly, visual cues are not the only type of cues that a participant could use. Future research should consider looking at the effectiveness of multiple types of cues, such as auditory and multi-sensory cues.

4.5 Conclusion

I sought to investigate the effects of prevalence and display type on human performance with an automated aid. The results indicated little to no evidence in support of an effect of display design or prevalence on participants temporal performance with an automated aid. However, there was an effect of prevalence and automated aid on participant's miss rates and criterion. Based on these results, I suspect that there is no added bonus in performance with integration of the automated aid's cue with meaningful task information. Moreover, the use of an automated aid might be a practical solution to alleviate the effects of low prevalence in real world situations. Designers should consider the potential influence of target prevalence and available strategies when designing the interface for automated aids.

References

Bennett, K. B., & Flach, J. M. (2011). *Display and interface design: Subtle science, exact art.* CRC Press.

Dolgov, I., & Kaltenbach, E. K. (2017). Trust in Automation Inventories: An Investigation and Comparison of the Human-Computer Trust and Trust in Automated Systems Scales. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting 61*(1).1271-1275.

Eidels, A., Donkin, C, Brown, S.D., & Heathcote, A. (2010) Converging measures of workload capacity. *Psychonomic Bulletin & Review 17*(6). 763-771.

Fenton, J. J., Taplin, S. H., Carney, P. A., Abraham, L., Sickles, E. A., D'Orsi, C., Berns, E.A., Cutter,G., Hendrick, R.E., Barlow, W.E., & Elmore, J. G. (2007). Influence of computer-aided detection on performance of screening mammography. *New England Journal of Medicine, 356*(14), 1399-1409.

Firmino, M., Morais, A. H., Mendoa, R. M., Dantas, M. R., Hekis, H. R., & Valentim,R. (2014). Computer-aided detection system for lung cancer in computed tomography scans: review and future prospects. *Biomedical Engineering online, 13*(1), 41.

Garner, W. R. (1976). Interaction of stimulus dimensions in concept and choice processes. *Cognitive Psychology, 8*(1), 98-123.

Houpt, J. W., Blaha, L. M., McIntire, J. P., Havig, P. R., & Townsend, J. T. (2014). Systems factorial technology with R. *Behavior Research Methods, 46*(2), 307-330.

Horowitz, T. S. (2017). Prevalence in visual search: From the clinic to the lab and back again. *Japanese Psychological Research, 59*(2), 65-108.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford University Press, Clarendon Press

Jian, J., Bisantz, A., Drury, C., & Llinas, J. (1998). *Foundations for an Empirically Determined Scale of Trust in Automated Systems*(No. CMIF198). Center for Multisource Information Fusion, Buffalo, NY.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50-80.

Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human Factors, 43*, 563-572.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230-253.

Peltier, C., & Becker, M. W. (2016). Decision processes in visual search as a function of target prevalence. *Journal of Experimental Psychology: Human Perception and Performance, 42*(9), 1466.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, 47, 877-903.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012), Default Bayes Factors for ANOVA Designs. *Journal of Mathematical Psychology*, 56, pp. 356374

Russell, N. C. C., & Kunar, M. A. (2012). Colour and spatial cueing in low-prevalence visual search. *Quarterly Journal of Experimental Psychology, 65*, 13271344.

Swets, J. A., Tanner Jr, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68*(5), 301.

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology, 39*, 321-359.

Wickens, C. D., & Carswell, C. M. (1995). The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors, 37*(3), 473-494.

Wolfe, J. M., & Van Wert, M. J. (2010). Varying target prevalence reveals two dissociable decision criteria in visual search. *Current Biology, 20*(2), 121-124.

Yamani, Y., & McCarley, J. S. (2018). Effects of Task Difficulty and Display Format on Automation Usage Strategy: A Workload Capacity Analysis. *Human Factors, 60*(4), 527-537.

Zinn, C.M., Houpt, J.W., Yamani, Y., & Scott-Sharoni, S. (2018) Assessment Function Analysis of Human-Automation Team Performance: A reanalysis of data from Yamani and McCarley (2018). *Proceedings of the Human Factors and Ergonomics Society Annual Meeting.*

Zinn, C.M., Yamani, Y., Houpt, J.W., & Scott-Sharoni, S. (2018). Gauging human-automation team efficiency with assessment functions. *Poster presented at Society of Mathematical Psychology annual meeting, Madison, Wisconsin.*

Appendix A

TASS Survey

You will be shown 12 statements for evaluating trust between people and automation. For each statement, please click on the point that best describes your feeling of trust or your impression of the automated aid on a scale from 1 (not at all) to 7 (extremely). Please click any button to start.

1. The system is deceptive
2. The system behaves in a secretive manner
3. I am suspicious of the systems intent, action or outputs
4. I am wary of the system
5. The systems actions will have a harmful or injurious outcome
6. I am confident in the system
7. The system provides security
8. The system has integrity
9. The system is reliable
10. I can trust the system
11. The system is dependable
12. I am familiar with the system

Appendix B

Mean Response Time Figure

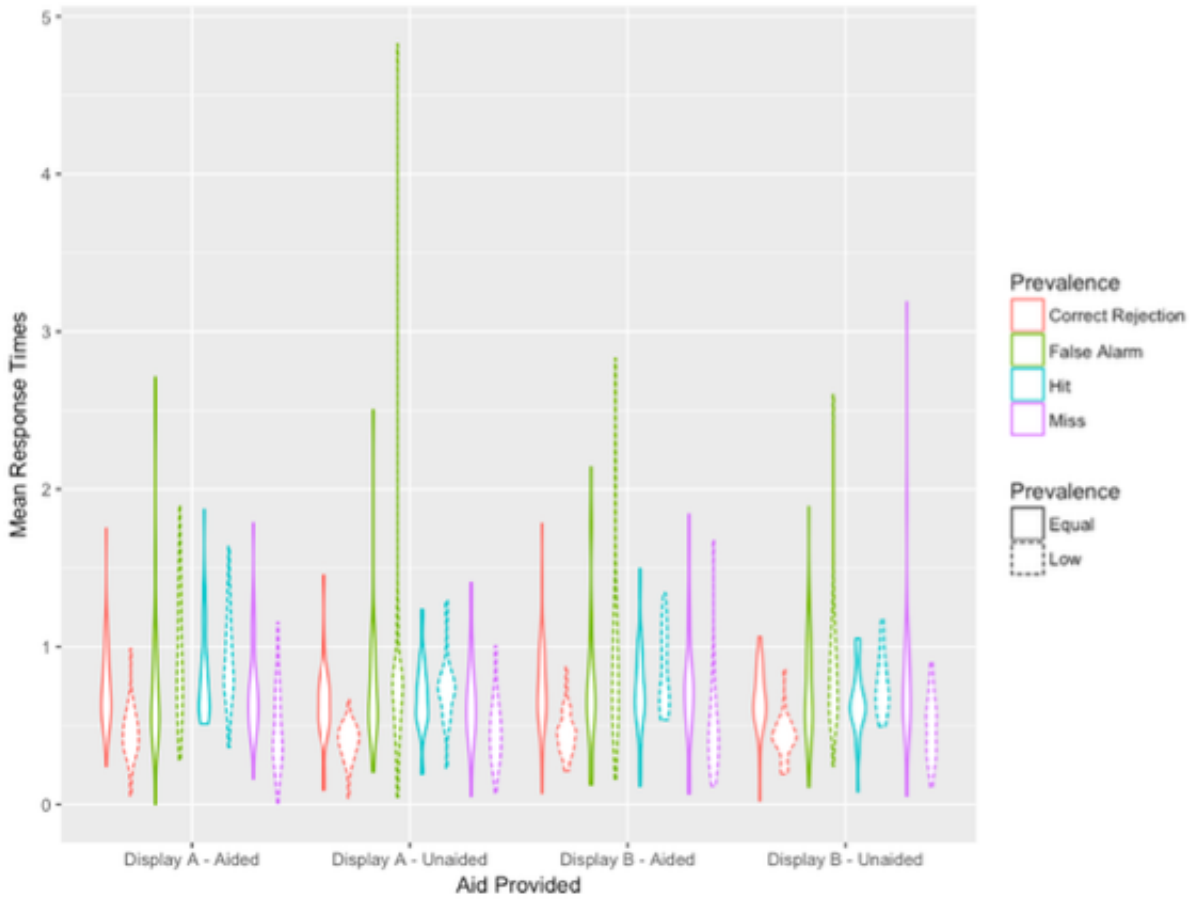


Figure B.1: Violin plots of mean response times for all response types by aid, prevalence, and display conditions.