Computational Brain & Behavior

Exploring the Performance Consequences of Target Prevalence and Ecological Display Designs When Using an Automated Aid.

--Manuscript Draft--

Manuscript Number:	COBB-D-20-00043R1
Full Title:	Exploring the Performance Consequences of Target Prevalence and Ecological Display Designs When Using an Automated Aid.
Article Type:	Original Paper
Funding Information:	
Abstract:	Human operators do not necessarily perform better when receiving assistance from an automated aid 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, we compared displays where the automated aid was integrated with task information in general or with another visual decision support aid. Subjects performed a speeded judgment task with the assistance of an automated aid, varying in display type, difficulty, and prevalence. Results indicated that participants performed less efficiently with the automated relative to without, and that there was no added benefit of the visual decision support in terms of response times. Both decision supports improved participant's sensitivity over no support, which may be beneficial for weakening the performance consequences of the low prevalence effect. Unexpectedly, we found that participants performance with each display was strongly dependent on when they experienced each display. It is possible participants might be using strategies that complement one display over another, depending on the condition they see first. Automated aids could be used in real world contexts to alleviate the effects of low target prevalence, however the effectiveness may depend on experience with other interfaces.
Corresponding Author:	Cara Kneeland, M.Sc Wright State University UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Wright State University
Corresponding Author's Secondary Institution:	
First Author:	Cara M Kneeland, M.Sc
First Author Secondary Information:	
Order of Authors:	Cara M Kneeland, M.Sc
	Joseph W Houpt, Ph.D.
	Kevin B Bennett, Ph.D.
Order of Authors Secondary Information:	
Author Comments:	
Response to Reviewers:	Reviewer #1: The study in this paper is sound and well-motivated, the predictions are sensible, and it's interesting and worth knowing that the display and automation supports were less beneficial than we might have expected. That's all good. A few concerns knock the manuscript back a little. The biggest problem, I think, is that the presentation of data makes the findings very difficult to follow. Many effects are described with pairwise comparisons in the text, without any visualisation, making patterns difficult to encode. On pages 21-22, for example, we read, "There was strong evidence in support of a difficulty and prevalence interaction, BF = 6.58 x 10^20. There was moderate evidence against a difference between easy and difficult conditions for both the equal prevalence condition, BF = 0.18, and the low

prevalence condition, BF = 0.15. There was strong evidence in support of a difference between equal prevalence and low prevalence conditions for the easy condition, BF = 1.13 x10 40 , and the difficult conditions, BF = 3.47 x10 ^40 . It could be that this interaction is largely driven by the large prevalence main effect."
That pattern would be much easier to encode if the means were graphed. The same is true for a number of other patterns reported throughout the paper. Graphs of all these effects would take up a lot of space, but would be very helpful to readers.

And to clarify, what I'd recommend is conventional line or bar graphs, with independent variables on the x-axis and as grouping factors, and the dependent variable on the y-axis. The current Figure 4 is useful for representing speed-accuracy tradeoffs, but doesn't make main effects and interactions very obvious.

More graphs were provided to help clarify the results. Additionally, results section was reorganized to also help organize the results to be more understandable.

A few smaller points:
- The introduction to assessment functions on page 11 is somewhat cursory. It would be helpful to have a few more sentences about how the functions work, what they measure, and how they lead to the conclusion that "the accuracy benefit of the decisional supports...may not outweigh the temporal disadvantage they produce."
More information on how to interpret the assessment functions were added.
- It appears that Bayes factors are calculated relative to the null model, but I didn't notice that stated explicitly (apologies if I overlooked it).
It was mentioned it on Page 17. However, we included an additional sentence on line 16-17 to make this more clear.
- In the discussion of prevalence effects, it might be worthwhile to consider work discussing the effects of low target prevalence on the positive predictive values of alerts. When prevalence rates are low enough, even a highly sensitive alarm may have a low positive predictive value, which presumably reduces the aid's value. Some papers that have discussed this in the human factors literature are referenced below. I don't mean to suggest that all (or necessarily any) of these papers need to be cited, just that it's important not to give the impression that automated aids might be a panacea for low-prevalence effects.
Thank you for including some papers. I have incorporated this literature into the discussion section to address this concern.

Signed,

Jason McCarley

Botzer, A., Meyer, J., Bak, P., & Parmet, Y. (2010). User settings of cue thresholds for binary categorization decisions. Journal of Experimental Psychology: Applied, 16(1), 1-15.
https://urldefense.com/v3/__https://doi.org/10.1037/a0018758__;!!On18fmf1aQ!h7afWu75gg4AcvZOqaQYEB0ygvlPgbiBzN80SjIyfZlLGiuHVnFGeTjtUVoPjoP7V9M$

Getty, D. J., Swets, J. A., Pickett, R. M., & Gonthier, D. (1995). System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. Journal of Experimental Psychology: Applied, 1(1), 19-33.

Parasuraman, R., Hancock, P. A., & Olofinboba, O. (1997). Alarm effectiveness in driver-centred collision-warning systems. Ergonomics, 40(3), 390-399.
https://urldefense.com/v3/__https://doi.org/10.1080/001401397188224__;!!On18fmf1aQ!h7afWu75gg4AcvZOqaQYEB0ygvlPgbiBzN80SjIyfZlLGiuHVnFGeTjtUVoPpgGugmA$

Sanquist, T. F., Doctor, P., & Parasuraman, R. (2008). Designing Effective Alarms for Radiation Detection in Homeland Security Screening. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38(6), 856-860.
https://urldefense.com/v3/__https://doi.org/10.1109/TSMCC.2008.2001708__;!!On18fmf1aQ!h7afWu75gg4AcvZOqaQYEB0ygvlPgbiBzN80SjIyfZlLGiuHVnFGeTjtUVoPkvttAW

o$

Reviewer #2: The authors present a study that investigates the effects of display design and target prevalence in decision making tasks that include automated aid. Contradicting predictions from the literature, evidence for the absence of display and prevalence effects was found via Bayesian analyses. Automated support slowed participants down, failed to remove participants' response biases, but did improve their decision accuracy when there were multiple sources of support. The authors delved deeper with a workload capacity analysis that showed the support was no more efficient than no support, and whatever benefits to accuracy were observed, these benefits do not justify the associated slow down.

Overall, the study provides results that contradict many of the predictions from the literature. Although this is interesting, I believe the reporting of the work needs to be improved before we can accept the findings here.

Firstly, the clarity of the analyses needs to be worked on. For example, details surrounding the Bayesian analysis and the mapping between analyses and hypotheses (at times you are testing more than is needed, e.g., 4 way interactions).

To have a clearer connection between the hypotheses and analyses, we took out the Order effect that were not hypothesized and reported them in a separate exploratory analysis section. This will hopefully improve the clarity in the main results and highlight the exploratory nature of the order effects.

Secondly, I think there needs to be further discussion surrounding all the contradicting findings reported here. The discussion focuses heavily on the order effects, which were not a prediction of the literature. I wonder if there is empirical or theoretical work that is in line with what has been found here - the lack of effect of automated aids.
More research was integrated to hopefully integrate this work with the larger literature, with regard to order effects.

Below are comments that I hope will improve the manuscript.
Gabriel Tillman
1. pg 3 "Would it be useful to define automation the first paragraph?"
Included a definition of automation on Page 3.
2. pg 5 line 2: What is the "Proximity Compatibility Principle"? For instance, EID is defined well later in the paragraph.
Elaborated on PCP on Page 5.
3. Pg 7: You stated "First, it provides a very direct mapping between visual features in the display and recommended responses. Therefore, response times should decrease." Given the use of SDT in the rest of this section, would it be worth noting that the effect you are describing is an increase in sensitivity (d') from no threshold support to threshold support conditions?
Further information on SDT is provided and this sentence was adjusted to better describe this effect.
4. Pg 8: At this point of the introduction, it is clear this task design can be interpreted under a signal detection framework, but not much writing prior to this has been dedicated to spelling out the key concepts of SDT. I think for clarity it would be good to explain some of the core principles of SDT before jumping into the effects of display design on bias and sensitivity.
SDT further defined on page 6.
5. Pg 8: It was a little odd to hypothesise about the display results in the middle of the introduction. The main reason being that there are references to the prevalence conditions which are not discussed until the following pages. It is more typical to leave hypotheses to the end of the intro given all required info would have been discussed by then.
Switched the prevalence and display design sections and moved hypotheses towards the end of the introductions.
6. Pg 8 Line 11: "Lastly, the accuracy and sensitivity (d') of responses will be a reflection of variations in the underlying domain semantics (i.e., performance will depend heavily upon the extent of differences between the mean and standard deviations of the underlying signal and noise distributions)." I am a little confused about

what is being said here. Perhaps the "accuracy and response time" of responses is more appropriate given they are the behavioural outcomes. With that change I think what is being said in this para is threshold support will increase sensitivity, decrease bias, but is not the only factor affecting accuracy and response time given the SDT distribution means and SD also change across conditions.
This sentence was reworded to improve clarity. In this case, it is not response times because SDT doesn't make predictions regarding response times. Sensitivity is measured based on behavioral outcomes.
7. Pg 9: In the target prevalence section there are no hypothesis given for the current study, much like in the previous section. There are examples, however. This makes sense given the lack of previous research with aids and target prevalence, yet it would be good to state the work is exploratory in this regard.
Made explicit that this work is exploratory in this regard. Page 4. " 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 (Wolfe & Van Wert, 2010; Peltier & Becker, 2016) to help inform this exploratory work."
8. Pg 10 line 4: "potentially misleading conclusions" some citations of examples here would be helpful I think.
Added citations. Page10 " Using mean response times collapses the entire RT distribution, resulting in a loss of information and could result in potentially misleading conclusions (e.g., Eidels et al., 2010; Yamani & McCarley, 2016)."
9. Pg 28: Perhaps it would be helpful to highlight that the measures of workload capacity will be measures of temporal efficiency. There are a lot of measures and terms introduced in this section and it would be good to clarify the operational links between them.
More information on defining what the outcomes of the workload capacity mean in terms of information processing.
10. Pg 12: "We predicted that participants would perform more efficiently with the Combined Support display format than the Automation Alone display in the difficult condition and in the low prevalence condition." Is this prediction also suggesting you don't expect differences in efficiency in the easy high prevalence condition?
Not necessarily for our predictions. We focused primarily on the difficult and low prevalence conditions.
11. Pg 12: The power analysis is conducted with respect to alpha = .05. Could this be explained in terms of the decision rule you have adopted with Bayes factors?
A footnote was included in the paper to further explain why we have a NHST power analysis and not a Bayes Factor approach. Originally, we intended to use a NHST approach, which would explain the power analysis. However, we switched to Bayes Factors after data collection as they are more informative than p values.
12. Pg 13: "To ensure participants experience a prevalence close enough to the desired prevalence, participants completed at least 800 trials." Is it possible to retrieve the presentation data and display it here as a percentage?
We added the average and standard deviation prevalence for participants.
13. Pg 13: "Following participants' responses, participants received feedback on their judgments." What feedback was provided?
We provided more detail on the feedback provided following every trial.
14. Pg 14: "If the participant relied solely on the threshold line, the participant would achieve 96% accuracy in the easy condition, and 68% in the difficult condition. This was determined by estimating the probability of observing a bar in the overlapped portions of the short and long bar length distribution, which would be incorrectly marked by the observer." Does the threshold line contain an area (due to its thickness) where the long and short bars could be confused confused because they are occluded by the threshold display? This would likely change the percentages reported here and in the following sections.
This was considered when creating the threshold on the display. The threshold was only 0.1 degrees of visual angle in height centered on 2.5 degrees from the bottom of the target bar. So participants would essentially not be able to distinguish between 2.55 and 2.45 with the threshold bar. The differences in discriminability between these values is very low. Moreover, the values occurring in that range showed up a maximum of 5 percent of the time for participants.
15. Pg 16 Line 45: In this section can the timing data of the stimulus presentation be provided? Moreover, were there any fixation crosses or masks used in the experiment?
Provided more detail on the timing of each phase of a trial.

16. Pg 17: In this section the model and effect priors used should be stated. Or at least state that default priors from the package were used.
We used the default priors from the package and added a statement on that.
17. Pg 17 Line 38: "BF = 1-3" Best to use the super scripts to inform the reader the ratio is alternative over null, BF10 = 1-3
Fixed it to BF 10
18. Pg 18: The bayes factors presented here are not for the different nested models resulting from your ANOVA. I assume you are reporting inclusion Bayes factors from the model averaged results. If this is the case, I think you could justify this in the previous section, perhaps by stating that individually inspecting all models nested under the 4-way ANOVA is not feasible because the model set is too large. Therefore, instead you will calculate inclusion Bayes factors which provide the model averaged evidence of an effect. You can write inclusion Bayes factors out as BFinclusion to represent this. One final thing to be careful of is whether the Bayes factor package is excluding models that have interactions and no main effects in the calculation (resulting in unbalanced model sets) or using the "matched" models only method. This can influence the Bayes factor value and may be worth checking.
The Bayes Factor package in R can be used calculate inclusion Bayes factors, however that is not being reported here.
19. Pg 18 Line 43: Im not sure what is being reported here. You are referring to three interactions and there is some inequalities reported that I don't understand. Is it that all BF for all the remaining effects lie within that range?
Yes, however these effects are presented in the Appendix now to avoid further confusion.
20. Pg 19: "Interestingly, the strongest model was an interaction between display, order, and Prevalence" I think you are reporting a model averaged inclusion Bayes factor here for an effect, rather than a model. Moreover, given your hypothesis is an interaction term including order necessary? It seems this variable would be best included as a covariate to control for it. The dimensionality of your analysis would be reduced too without studying this variable.
We reran without the order and reported the order effects in a exploratory analysis section.
21. Pg 21 Line 1: Is it possible to provide tables or plots of the criterion data? It would be good to see what is driving the interaction between prevalence and difficulty.
Provided a plot for Figure 8 on page 23.
22. Pg 24: Could the no support response time distribution serve as the numerator for the capacity coefficient calculation for all other display options? That way you would get three different types of coefficients for the aided conditions and could compare them.
Yes the no support data was used to calculate the UCIP model for the workload capacity analysis. For the workload capacity analysis we were comparing performance between aided versus unaided.
23. Pg 27: At this point there is quite a bit of information to digest in terms of effects from the ANOVA analysis. Perhaps another approach is to point readers to the tables in the appendix for the complete analysis and only discuss the BF from results that directly speak to the hypothesis presented in the introduction. There are quite a few things tested in each ANOVA write up, but only a few hypotheses presented in the intro. For example, quite a bit goes into discussing order effects, yet these are not a main focus of the study.
Agreed. Exploratory analyses has now been separated and minimized to highlight the main takeaways. All other interactions are reported in the Appendix.
24. Pg 28 Line 47: I think this is referring to Figure 6 not 5.
Reordered the figures.
25. Pg 29: "We performed a logistic regression to see if the automation's response can be predictive of the human response." I'm not sure what was done here with this analysis, can more detail be provided?
More detail on the logistic regression was provided here.
26. Pg 41: Table 1 and 2 need the first column cleaned up so that the effects can be more easily discerned.
Tables were cleaned up and effects were bolded

Reviewer #3: Journal: Computational Brain & Behavior

Title: Exploring the Performance Consequences of Target Prevalence and Ecological Display Designs When Using an Automated Aid.
General evaluation: In this paper, the authors followed Yamani and McCarley's (2018) study and further tested the effect of automated aids on the decision efficiency by varying the display type (no decision support, automation alone, threshold alone, and combined support), difficulty (easy/ difficult), and target prevalence (equal/ low). The authors adopted the single-target self-terminating capacity to quantify the assistance of automated aids on the decision efficiency. Results showed limited to unlimited capacity, suggesting that participants performed similarly or less efficient with the aids relative to without. More importantly, the order of the test conditions interacted with the other factors, implying that the effectiveness of the automated aids may depend on how the participants experience the display. Generally, the study was well-conducted. This study addresses an important issue and could result in significant practical implications on the design principles
regarding the assistant aids. However, if not familiar with Yamani and McCarley's (2018) study, it is difficult for the readers to understand the rationale of the current study. In other words, the writing should be improved, especially for the rationale and hypotheses in the introduction, the details of the method, the results interpretation, and how the results link to the literature in the general discussion. Please see below for my specific points. I believe that with adequate revision, this work is suitable to be published at Computational Brain & Behavior.

Comments and suggestions:
1. In the abstract, the authors argued that the automated aids could be used in the real-world context to alleviate the effect of low target prevalence. I'd like to learn more about "how" to make it feasible. I suggest the authors providing detailed discussions with several real-world examples in the general discussion.
Real world examples (e.g., collision warning systems and computer aided detection systems) were included and discussion on how automated aids can and have been implemented in low prevalent conditions.
2. Figures 1 and 2 are relatively unclear to me. These two figures are very important ones for the readers to have a general idea about the concept of the study. However, the schematic representation of the figures makes me confused about the design and procedure of the experiments. The meanings of the dotted black, green, red, and blue lines are unclear. Are they one condition or two conditions? How do the participants experience the trials and how do they respond? A clear way is to plot the four display types separately for the ease of understandings. In addition, what is the SOA manipulation? If the two of the four display types are presented within a block with order being counterbalanced, do the participants know whether the aid will be presented or not? Is it possible to make a response before the presentation of the aid in the aid-present trials? For the aid-absent trials, the RT is recorded from the beginning of the trial presentation; but how about the aid-present
condition? I also read Yamani and McCarley's (2018) paper, they used an odd way to determine the RT, but how do the authors record the RT (i.e., not sure whether the random SOA being counted or not)?
More information was provided in the methods section regarding the handling of RTs. Additionally, the figures were changed to be easily interpretable.
3. Yamani and McCarley (2018) compared the decision efficiency across the integrated displays and separate displays. But, in the present study, the authors used the threshold line instead of the separate display. I am curious about the rationale of using the threshold line. How may the threshold line help the participants' decision? Why is it called "threshold"? For me, the threshold line alone does not provide any informative cues for the judgement of the length of the bar. Why do the authors hypothesize that the threshold line can help decision? How do the threshold line manipulations inform the display design? What do the results relate to Yamani and McCarley's (2018) separate display design?
More information was included to better describe the Yamani & McCarley study, as well as how it relates to the current display designs. Specifically, we highlighted the importance of the threshold and what the threshold means in terms of SDT.
4. How does the current study relate to PCP and EID? In the general discussion, the authors should highlight how the results could inform the display design.
Further discussion of PCP and EID was added in the introduction. Specifically, how the current use of EID compares to PCP approaches from Yamani and McCarley 2018.
5. P. 5 line 50-54, not sure why the authors stated the signal detection model and

	the ideal observer threshold. Does it relate to the threshold line manipulation? Yes, further information was provided to make this connection clearer 6. For the display type, is the design a one-way within-subject design or can we trat it as a nested design instead? What I mean is that aid presence is nested within the threshold presence. If using the nested-design ANOVA to analyze the data, the data can be interpreted more properly, and significant results are expected. The display design was treated as a one-way within subject design. Using a nested design did not necessarily yield different (e.g., more significant) results. 7. P.7 last paragraph and p.8 first paragraph, the rationale is unclear. The rationale for these hypotheses are clarified. 8. P.11 about the assessment function, there are four response types, not sure why the authors only focus on the correct and fast responses. Please explain. More information was provided about the fast and slow assessment functions. Only the slow assessment functions are reported as they are not different from the fast results. A footnote is provided to explain this. 9. The method sections should be clarified. More details were provided in the methods section. 10. There is lack of details for the fPCA analysis. How do you decide two PCs? What are the two PCs for? Please also discuss about how fPCA could help data interpretation. More detail was added to describe the fPCA analysis. 11. For the results part, please present the descriptive statistics in tables or figures for all the combinations of factors. Descriptive statistics reported in Appendix 12. Why shall the authors regard the order as an important factor? Please state the rationale in the introduction. Are these unexpected results? Order effects were considered unexpected, and therefore the order results were moved to a separate section of the results that highlights them as a post hoc exploratory analysis 13. The authors present the signal detection analysis. But, they didn't introduce the rationale in both introduction and method sections. Further information regarding signal detection theory and how it relates to this study was provided in the introduction. 14. The quality of the capacity plots needs to be improved. Assuming that most readers are not familiar with the capacity analysis, please provide more explanations regarding how to interpret the results and their implications. In addition, the legends of present/absent and threshold present/absent are confusing. More information on how to read the plots were added. 15. For the sensitivity results, you are discussing about hits and misses without showing the data. Hits, Misses, FA, and CR are all reported in appendix Signed by Cheng-Ta

Running head: EFFECT OF DISPLAY AND PREVALENCE WITH AN AUTOMATED AID

Exploring the Performance Consequences of Target Prevalence and Ecological Display Designs When Using an Automated Aid.

Cara M. Kneeland[1*], Joseph W. Houpt[2], and Kevin B. Bennett[1]

[1] Wright State University, Department of Psychology. Dayton, OH

[2] University of Texas at San Antonio, Department of Psychology, San Antonio, TX

[*] Corresponding Author

Email: zinn.10@wright.edu

Abstract

Human operators do not necessarily perform better when receiving assistance from an automated aid 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, we compared displays where the automated aid was integrated with task information in general or with another visual decision support aid. Subjects performed a speeded judgment task with the assistance of an automated aid, varying in display type, difficulty, and prevalence. Results indicated that participants performed less efficiently with the automated relative to without, and that there was no added benefit of the visual decision support in terms of response times. Both decision supports improved participant's sensitivity over no support, which may be beneficial for weakening the performance consequences of the low prevalence effect. Unexpectedly, we found that participants performance with each display was strongly dependent on when they experienced each display. It is possible participants might be using strategies that complement one display over another, depending on the condition they see first. Automated aids could be used in real world contexts to alleviate the effects of low target prevalence, however the effectiveness may depend on experience with other interfaces.

Keywords

Human-automation interaction, Systems Factorial Technology, Prevalence, Ecological Display Design

Introduction

Automation is often implemented with hopes of improving human performance, by allowing the human to conserve cognitive effort (Parasuraman & Riley, 1997; Wickens & Dixon, 2007) or by supporting the processing of information (Parasuraman, Sheridan, & Wickens, 2000). Automation in this case is defined as a machine agent that performs a task that was previously and can still be completed by a human, such as automated teller machines (Parasuraman & Riley, 1997). However, people often misuse the automation (e.g., Parasuraman & Riley, 1997), resulting in poor performance with an automated aid. Researchers have speculated various factors that influence the human's use of the automation including the display design (e.g., Yamani & McCarley, 2018) and aid's reliability (e.g., Wickens & Dixon, 2007). Horowitz (2017) suggested that the prevalence of a target may also influence the use of an automated decision aid, such as in the case of computer aided detection. For example, an individual might observe that a target occurs less than 2% of the time then be more likely to assume the automated aid's cue is a false alarm. Malfunctions and targets can occur infrequently in the real world, and it is important to examine the effects of infrequent targets on human performance with an automated aid. The purpose of this study was to examine the effects of infrequent targets and the interaction of display design and target frequency on human performance with an automated decision aid.

Target Prevalence

The prevalence of a target is 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. In many real world applications, targets for computer aided detection occur less than five percent of the time. In the lab and with training automated aids, targets occur much more frequently which can inflate the aids false alarm rate (Horowitz, 2017). By analyzing the effect of target prevalence, the current paper may provide a better depiction of human

performance with an automated aid when targets are rare. Because there is little to no research done on the effect of prevalence on individuals' performance with an automated aid, the current paper draws from previous research on the effect of prevalence on individuals' unassisted performance (Wolfe & Van Wert, 2010; Peltier & Becker, 2016).

The low prevalence effect refers to when participants are more likely to miss rarely-present target than if the target was present more often (Wolfe & Van Wert, 2010; Peltier & Becker, 2016). 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) posits that the low prevalence effect in visual search is because 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 (Wolfe & Van Wert, 2010; Peltier & Becker, 2016) when targets are rare. In their model, these effects are driven by a lowered threshold for quitting and decision criterion that favors correct rejections (even at the expense of missed targets).

This previous research on the causes of the low prevalence effect indicates that 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. Moreover, a highly sensitivity aid that performs well in the lab may be less valuable in the real world with low prevalent targets, because of the inflated false alarms that these sensitive aids produce (e.g., Botzer et al., 2010; Getty et al., 1995). Because of these potential effects, creating an effective design that encourages the use of the aid in these low prevalence conditions is essential to forming an efficient team between the human and the automated aid.

Display Design

Yamani and McCarley (2018) investigated the extent to which the human's inefficient performance with an automated decision aid could be attributed to issues in display design. Their task simulated a manufacturing environment where each trial involved speeded judgments of whether to attempt production based on the amount of raw materials provided. A bar representation was provided in all four experimental conditions; participants were instructed to accept shorter bars for production and to reject longer bars. Participants were also provided with the output of an automated aid which provided a binary recommendation (accept or reject) in the form of a color cue.

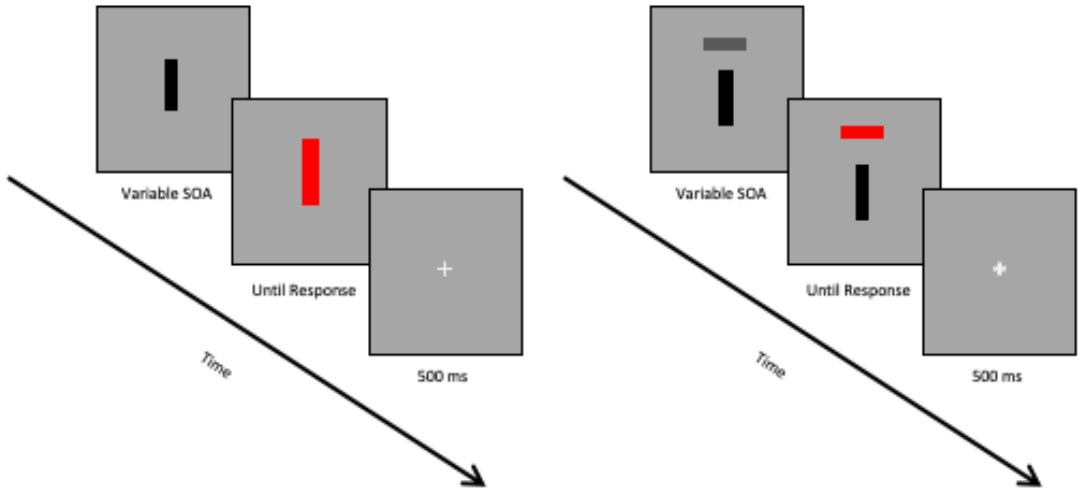
Yamani and McCarley (2018) evaluated four alternative displays (see Figure 1). In this task, participants were shown the vertical bar first with the automated aid having a variable onset. Following each trial, participants received feedback on whether they were correct or incorrect with either a white "+" or a red "x". In the integrated design (left panel of Figure 1) the height of a vertical bar represented the amount of raw materials provided. In the aided version of this display, the color of this vertical bar was used to represent the recommendation of the automated aid (green – accept; red – reject). In the separated display format (see right panel of Figure 1) the vertical bar graph remained the same, but a second horizontal bar graph was added which either changed color to provide a recommendation.

Yamani and McCarley (2018) framed issues in display design from the theoretical perspective of Wickens and Carswell's (1995) Proximity Compatibility Principle (PCP). Briefly, PCP is one approach to display design that is concerned with the agreement between the perceptual proximity of the objects on the display with the necessary processing proximity required by the task. For example in a divided attention task, participants would benefit from having high perceptual proximity (e.g., objects closer together) to support their processing of both objects in parallel. Likewise, for a selective attention task, participants would benefit from having low perceptual proximity (e.g., objects placed further apart) to

support their selective processing of one source of information.

Figure 1

Task from Yamani and McCarley (2018) showing all display formats



Note. This shows the task from Yamani and McCarley (2018) with the integrated display (left panel) and the separated display (right panel). Participants in this experiment saw both displays with and without the aid.

Yamani & McCarley (2018) hypothesized that human performance with an automated decision aid is a divided attention task because participants are encouraged to attend to both the automation and the vertical bar. In this case, human performance should be improved with the integrated display (because the automation's cue was physically integrated with the task information) relative to the separated display (because there are two representations physically separated in space) according to PCP. Their results did not appear to support the hypothesis (however, see Zinn et al., 2018, for a reanalysis of their data). Yamani and McCarley (2018) speculated that an alternative approach to display design that focused more on the display semantics and its mapping of the automation cues to the underlying physical properties of the stimulus set might have proven more successful; they cited ecological interface design (Bennett & Flach, 1992) as one example of an alternative

approach.

Ecological interface design (EID) is a theoretical perspective on display and interface design (e.g., Bennett and Flach, 2011) which emphasizes complex interactions between domain semantics, graphical representations, and user cognitive / perceptual capabilities and limitations. The principles of EID are used to develop graphical decision support which transforms decision making from an activity which requires complicated cognitive processes to one which requires simple perceptual processes. This is accomplished by developing graphical representations that provide one-to-one mappings between the higher-order visual properties of a display and the critical semantic properties of a work domain. For the current study, we designed an alternative display for the Yamani & McCarley task using the principles of EID.

The domain semantics of the Meyer (2001) and Yamini and McCarley (2018) task are essentially defined by the concepts of classic signal detection theory (Green & Swets, 1966). Signal detection theory is used to measure a person's ability to distinguish a signal from noise. According to signal detection theory, the presence of a target and the presence of a non-target (or in many cases the absence of a signal) are represented by two normal distributions. The degree of the offset between the two distributions (d') indicates how well a person can distinguish the target from the non-target (e.g., high separation between the two distributions means higher sensitivity). Under this model, we can also measure a person's criterion or bias towards responding signal or noise (e.g., higher criterions indicates more response bias for noise and lower criterions indicate more bias towards signal).

The ecological display was designed to provide a visual representation which directly reflects these domain semantics and provides a salient perceptual cue to aid decision making. The fundamental task constraints in Yamani & McCarley (2018) are jointly determined by the properties (i.e., mean and standard deviation) of a signal distribution (i.e., as defined for the short bar) and an overlapping noise distribution (i.e., as defined for the long bar). The

ecological display incorporates a horizontal line which is located at the midway point between the means of these two distributions (see Figure 2). In signal detection terms, the location of this line corresponds to the location of the "Ideal Observer's Threshold" when signal and noise responses are equally likely. In other words, this threshold indicates the recommended criterion for distinguishing long and short bars with equal prevalance. This display format provides a very direct form of decision support. Practically speaking, when targets and non-targets are equally likely, if the horizontal line intersects the bar graph (as illustrated in Figure 2, lower panel) then the participant should provide a "no" response. Alternatively, if it lies above the bar graph then the participant should provide a "yes" response.

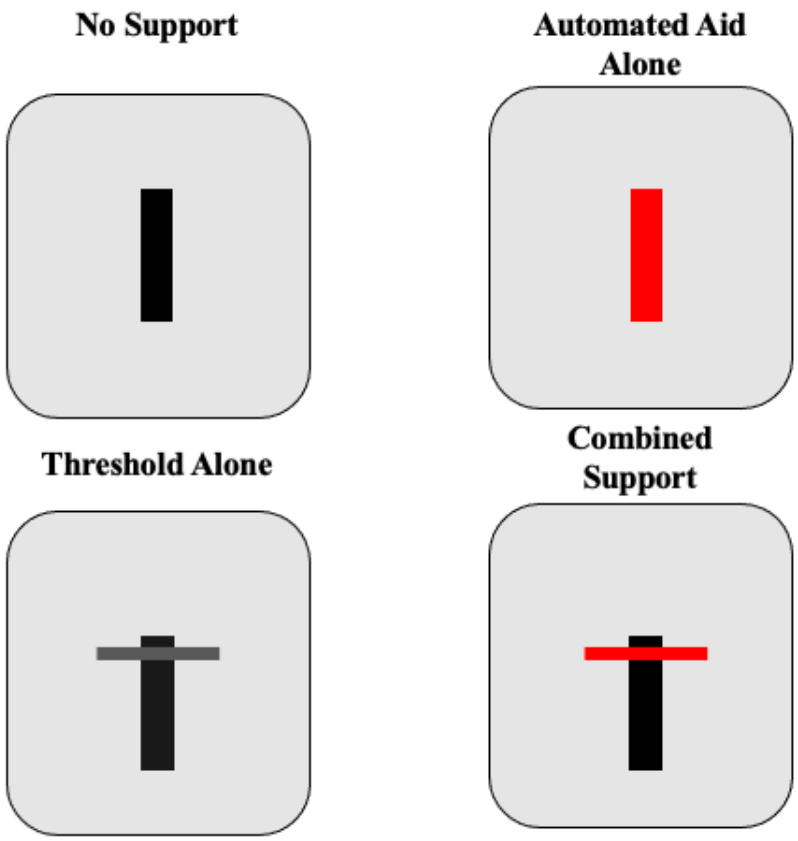
Thus, an explicit design goal of EID has been accomplished. The activities needed to produce a response have been transformed from those which require limited cognitive resources (e.g., mental comparisons in working memory involving internal representations of prototypical bar lengths) to those which leverage high perceptual resources (e.g., perceptual discriminations to determine whether or not the line intersects the bar graph). The task becomes a simple perceptual task (i.e., judging the bar's height relative to the location of the threshold line) as opposed to a complicated cognitive decision making task (i.e., comparing the bar height to the person's memory of long and short bars). Integrating the automated aid with this visual decision support may prove to be more effective in improving user performance with an automated aid than integrating the automated aid more generally with the task information. We will refer to this generally as the Threshold Alone display format.

In the present study we evaluated four varieties of decision support formed by a factorial combination of two display formats (bar graph or Threshold) and two levels of automated aiding (on or off), shown in Figure 2. The *No Support* (NS) display consists of the bar graph alone. The *Automation Alone* (AA) display integrates the automated aid's recommendation by color-coding the bar graph. These two displays replicate the Integrated

design displays of Yamani and McCarley (2018). The *Threshold Alone* (TA) display adds the threshold indicator line to the basic bar graph display. The Combined Support (CS) display incorporates both forms of decision support by color-coding the threshold indicator line to reflect the recommendation of the automated aid . In addition to display design, we also manipulated the prevalence of the short bar length and the difficulty of the task.

Figure 2

All 4 display designs for the current study



Note. This contains represents all 4 displays compared in the current work.

Current Study

The current project explores the effect of display design and target prevalence on human usage of an automated aid using the same task as in Yamani and McCarley (2018). The display designs used in this study vary in the degree that the automated aid integrates

with underlying properties of the stimulus set (i.e., the length of the bar that lies between the long and short distributions).

We hypothesize that the inclusion of the threshold support will improve performance to varying degrees. First, it provides a direct mapping between visual features in the display and recommended criterion for distinguishing long and short bars. Because this allows the participants to rely on perception more than memory, response times should decrease and sensitivity (d') should increase. Second, participants should be less biased using the Threshold Alone display format. The location of this line provides a visual representation of the exact decision criterion that should be adopted to minimize both the probability of misses and false alarms (assuming equal prevalence and cost for each of these types of error). It is important to note that minimizing bias would be most appropriate for the equal prevalence condition, but possibly not for the low prevalence condition. In the low prevalence condition, it can be sensible and adaptive for a participant to have a stronger bias towards the more prevalent outcome than the less prevalent outcome (e.g., Rich et al., 2008). However, this bias towards the more prevalent outcome can lead to more misses of the less prevalence outcome (e.g., Horowitz, 2017). Lastly, performance (e.g. accuracy and sensitivity) will depend upon the differences between the means of the underlying signal and noise distributions, which are determined by the difficulty conditions.

We hypothesized that participants would perform more efficiently using the display that integrates the automated aid's information with the threshold line (Combined Support) than the display that does not (Automation Alone). Further, we predicted that participants would perform relative efficiency of the Combined Support display would be accentuated in the difficult and low prevalence conditions. We predicted this interaction because the integration of the automated aid recommendation and the threshold line in the Combined Support display would be most beneficial in situations where there is more uncertainty in distinguishing between the bar length categories, as would be expected in the low prevalence

condition and difficult condition.

Workload Capacity

To categorize performance with an automated aid, this study used 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. The benefit to using the workload capacity analysis instead of standard analyses of temporal performance, like mean response-time comparisons, is that workload capacity analysis uses entire response time (RT) distributions. Using mean response times collapses the entire RT distribution, resulting in a loss of information and could result in potentially misleading conclusions (e.g., Eidels et al., 2010; Yamani & McCarley, 2016). By using the entire RT distribution, workload capacity analysis avoids potentially misleading or ambiguous conclusions about a system's performance (Eidels et al., 2010).

Workload capacity analysis is a part of the Systems Factorial Technology framework (Townsend & Nozawa, 1995; Houpt et al., 2014; Little et al., 2017). 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 et al., 2018). There are two measures of workload capacity: the capacity coefficient and the assessment functions. Capacity coefficient focuses solely on temporal performance of correct trials and is more amenable to inferential statistics than the assessment functions. Assessment functions provide a joint analysis of response time and accuracy. Assessment functions vary in terms of whether correct or incorrect responses are used, and whether measures are based on cumulative distributions or survivor functions (Townsend & Altieri, 2014; Donkin, Little, & Houpt, 2014).

Specifically, we used the single-target self-terminating (STST) stopping rule for our workload capacity analysis. Previously, Yamani and McCarley (2018) used both the OR and the AND stopping rule capacity coefficients to describe temporal performance before and after the onset of the automated aid. We used the STST stopping rule (Blaha, Townsend, Kneeland, & Houpt, Under Review) because it compares human performance with the automated aid to a null model (e.g., processing the task information independent of the aid). With the OR and the AND stopping rule would require making further assumptions about their strategy for decision making (e.g., processing either task information or automated aid's information, or processing both information).

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, and describes the probability that the participant has not responded at the given time, *t*.

The equation for the C_{STST} is :

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

Participants' performance with the automated aid was compared to their performance without the automated aid. The baseline for comparing performance was the predicted performance of an unlimited-capacity, parallel, and independent (UCIP) model. In other words, information is processed independently and in parallel, with no additional cost or benefit to processing each individual source of information. A participant is performing at limited capacity when information is processed less efficiently ($C_{\text{STST}}(t) < 1$) than his/her baseline predicted by the UCIP model. A participant is performing at unlimited capacity when the participant is processing information as efficient as his/her baseline ($C_{\text{STST}}(t) = 1$). A participant is performing at super capacity when the participant is processing information more efficiently ($C_{\text{STST}}(t) > 1$) than his/her baseline.

Assessment functions provide a joint analysis of response time and accuracy, and in this case indicate whether there is a change in the speed and accuracy of processing information with an aid compared to without. Assessment functions vary in terms of whether correct or incorrect responses are used, and whether the response time distribution or the survivor function is used (called fast and slow, respectively). For this study, we are interested in the correct responses, as they are more informative regarding whether participants are performing more efficiently with the aid compared to without. Typically, using cumulatived distribution funcions of response time are more appropriate for modeling maximum time tasks, and survivor functions are more appropriate for minimum time tasks (Townsend & Wenger, 2004). Because STST is neither a minimum or maximum time task, neither function is necessarily more or less appropriate. Moreover, both the fast and slow assessment functions contain similar information as they are both estimated from response time distributions (either by using the response time distribution directly or transforming this distributions into survivor functions). Similar to Zinn et al. (2018), we used the correct-slow assessment function[1], written as:

$$A_{CS}^{STST}(t) = \frac{\log[P_{aided}(\text{Correct})] - H_{aided|Correct}(t)}{\log[P_{unaided}(\text{Correct})] - H_{unaided|Correct}(t)}$$

where $P_{aided}(\text{Correct})$ is the probability of a correct response when aided, and $H_{aided|Correct}(t)$ is hazard function at time, t, for correct aided trials. We used both the capacity coefficient and the assessment functions to provide a more comprehensive analysis of workload capacity.

For the assessment functions, interpretations are similar to that of the capacity coefficient. A participant is performing at limited capacity when the participant is processing information less efficiently and less accurately (e.g., $A_{STST}(t) < 1$) than his/her baseline. A

[1] We checked to make sure the fast assessment functions contained similar results, as expected considering both approaches are reliant on response time distributions. We only report the slow results for simplicity and because it does not contain any different results from the fast assessment functions.

participant is performing at unlimited capacity when the participant is processing information as efficient and as accurately as his/her baseline (e.g., $A_{\mathrm{STST}}\ (t) = 1$). A participant is performing at super capacity when the participant is processing information more efficiently and more accurately (e.g., $A_{\mathrm{STST}}\ (t) > 1$) than his/her baseline

Method

Participants

A power analysis on pilot data (n = 4) indicated that a sample size of 60 participants would be sufficient for 95% power at a statistical significance level of .05.[2] 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. Participants were excluded for low accuracy (less than 55% accurate) or incomplete data.

Task

The task used in this study was similar to that of Yamani and McCarley, shown in Figure 1. 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 instructed 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. We 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

[2] This may seem a little confusing given that we are primarily using Bayesian approaches but have a frequentist approach to the power analysis. Originally, we intended on using null hypotheses significance tests (NHST) for the analyses presented here, which is why we used a power analysis to achevive sufficient power for a significance level of 0.05. However, we decided that a Bayesian approach to our analyses would be more informative than an NHST approach after data collection.

of visual angle for short and long bars respectively. Participants viewed the stimulus from a seating distance of 90 cm without the chin rest which informed the degrees of visual angle calculations.

The standard deviation of the bar length distributions and sampling proportion of short and long bar lengths varied, depending on the difficulty condition and prevalence condition. With a larger overlap between the long and short bar length distributions, it is more difficult to distinguish long and short bar lengths. Therefore, we 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). 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. On average, participants in the low prevalence condition saw the long bars 10.07 % of the time (SD = 0.01) and 49.73 % (SD = 0.01) for the equal prevalence condition.

Following participants' responses, participants received feedback on their judgments by providing either a red "x" for incorrect or a white "+" for correct. We 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.

Decision Support

Two forms of decision support were developed for this experiment. The threshold display format represented the statistical constraints of the task. A horizontal indicator line was placed at 2.5 degrees of visual angle (i.e., the middle point between the means of the

noise and the signal plus noise distributions) from the bottom of the bar graph (see Figure 1b). If the participant relied solely on the threshold line, the participant would achieve 96% accuracy in the easy condition, and 68% in the difficult condition. This was determined by estimating the probability of observing a bar in the overlapped portions of the short and long bar length distribution, which would be incorrectly marked by the observer.

The second form of decision support simulated an automated aid. 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. This was to prevent participants from anticipating the automated aid's cue (Yamani & McCarley, 2018). The aid's reliability was set at 95% across difficulty conditions to encourage the use of the aid while simulating imperfection.

These two forms of decision support (present, absent) were combined factorially to produce four support conditions: No Support (bar only), Threshold Alone display (bar plus indicator line), Automation Alone (bar with color cue), or Combined Support (bar with color cue and threshold line). These different display conditions are shown in Figure 2. If they relied on the threshold line 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. This was determined by estimating the joint probability of observing a bar in the overlapped portion of the bar length distributions and the automated aid being incorrect, in which both decision supports fail for the user.

Procedure

Participants first completed the demographic survey with the informed consent. Participants completed a training session and an experimental session on two separate days (about 2 days apart; max 10 days apart). Participants completed 16 blocks of 50 trials (800 trials in total) in the training session without the presence of automated aiding (i.e., No Support and Threshold Alone conditions). Participants experienced the aided and unaided

conditions on alternating blocks. In other words, within a single block, participants completed 25 trials in one aid condition, then 25 trials with the other aid condition. The order of the aid conditions randomly sampled initially and carried out through all blocks.

Participants completed 32 blocks of 50 trials (1600 trials in total) in the experimental session. The four decision support conditions (within-subjects factor) were paired for presentation based on the display format. The No Support and Automation Alone conditions had a single bar graph; the Threshold Alone and Combined Support conditions had a bar graph plus the threshold line. Participants completed the first 8 blocks with a pair of displays and one level of difficulty (Easy or Difficult; within-subjects factor). The displays were alternated (order randomly determined) between the 8 blocks. The next 8 blocks were completed with the same pair of displays and the second level of difficulty. The final 16 blocks were completed with the second pair of displays and the same presentation order for the two levels of difficulty. Note that the pair of displays within an 8-block sequence differed only with regard to the presence or absence of the automated aid. The presentation order for the two display pairs and the two levels of difficulty were counterbalanced; this counterbalancing scheme was replicated (see Figure 3) for both levels of Target prevalence (Low or Equal; between-subjects). Participants were randomly assigned to one of the resulting 8 groups.

Each trial began with the presentation of the vertical bar. The center of the bar was placed one degree up (or down) and one degree left (or right) from the center of the display (randomly determined). A sample from an exponential distribution with a mean of 676 milliseconds determined when the onset of the automated aid's cue occurred each trial. Participants indicated their judgement by clicking the right mouse button for short and the left mouse button for long. Participants then received feedback for 500 milliseconds before the onset of the next trial. We collected participants' response times and whether the participant was correct for data analysis.

Data Analysis

We analyzed correct mean response times and capacity coefficients z scores (Houpt & Townsend, 2012) in R (R Core Team, 2019) with Bayesian analyses from the "BayesFactor" package (Rouder & Morey, 2012; Rouder, Morey, Speckman & Province, 2012). Mean response times were of the humans response alone, and not adjusted based on the onset of the aid. Default priors from the "BayesFactor" package was used. The capacity coefficient z scores are representations of the entire RT distribution into a single score that can be used for inferential statistics (Houpt & Townsend, 2012). Assessment functions on the other hand are less amenable to inferential statistics and rely on the use of functional principal component analysis (fPCA) to observe group level effects (Burns, Houpt, Townsend, & Endres, 2013). Briefly, functional principle component analysis is used to identify major deviations of individual functions from the overall mean function. Individual functions are assigned a value that signifies the degree to which a given deviation is characteristic of the present function.

We used Bayes Factors instead of traditional F tests. The benefit of using Bayes Factors (BF_{10}) as opposed to traditional NHST testing is that Bayes Factors can represent evidence in support of the null hypothesis or the alternative hypothesis. All models presented in the result section compare an alternative model (e.g., a model with a main effect of display) to a null model (e.g., a model assuming no effect). Labels provided in Jeffreys (1961) informed the labels used in this study. Weak evidence in support of the alternative hypothesis is a BF_{10} = 1-3 (BF_{10} = 0.3 - 1 for the null hypothesis). Moderate evidence in support of the alternative hypothesis is a BF_{10} = 3-10 (BF_{10} = 0.1 - 0.3 for null hypothesis). Strong evidence in support of the alternative hypothesis is a $BF_{10} > 10$ ($BF_{10} < 0.1$ for null hypothesis).

Figure 3

All possible order conditions in the current work.

Decision Support:
No Support (NS) -- bar graph; Automation Alone (AA) -- bar graph and automated aiding
Threshold Alone (TA) -- bar graph and threshold line; Combined Support (CS) -- bar graph, threshold line and automated aiding

Task Difficulty: Easy, Difficult

Prevalence	Block 1-8	Block 9-16	Block 17-24	Block 25-32
Low	TA / CS Easy	TA / CS Difficult	NS / AA Easy	NS / AA Difficult
	NS / AA Easy	NS / AA Difficult	TA / CS Easy	TA / CS Difficult
	TA / CS Difficult	TA / CS Easy	NS / AA Difficult	NS / AA Easy
	NS / AA Difficult	NS / AA Easy	TA / CS Difficult	TA / CS Easy
Equal	TA / CS Easy	TA / CS Difficult	NS / AA Easy	NS / AA Difficult
	NS / AA Easy	NS / AA Difficult	TA / CS Easy	TA / CS Difficult
	TA / CS Difficult	TA / CS Easy	NS / AA Difficult	NS / AA Easy
	NS / AA Difficult	NS / AA Easy	TA / CS Difficult	TA / CS Easy

Note. This shows the order conditions across the within subjects treatments (display and difficulty) as well as the between subjects treatments (prevalence). Within a set of blocks, participants alternated between aided and unaided displays (e.g., TA/CS or NS/AA).

Results[3]

Mean Response Time analysis

We ran a 3-way ANOVA including display (No Support, Automation Alone , Threshold, and Combined Support), prevalence (high and low), and difficulty (easy and difficult). Mean response times across conditions are shown in Figure 4. There was strong evidence in favor of a main effect of prevalence, $BF_{10} = 2.01 \times 10^5$, in which participants in the low prevalence condition had faster mean response times ($M = 0.46$ seconds, $SD = 0.15$)

[3] All Descriptive Statistics and ANOVA results are presented in the Appendices

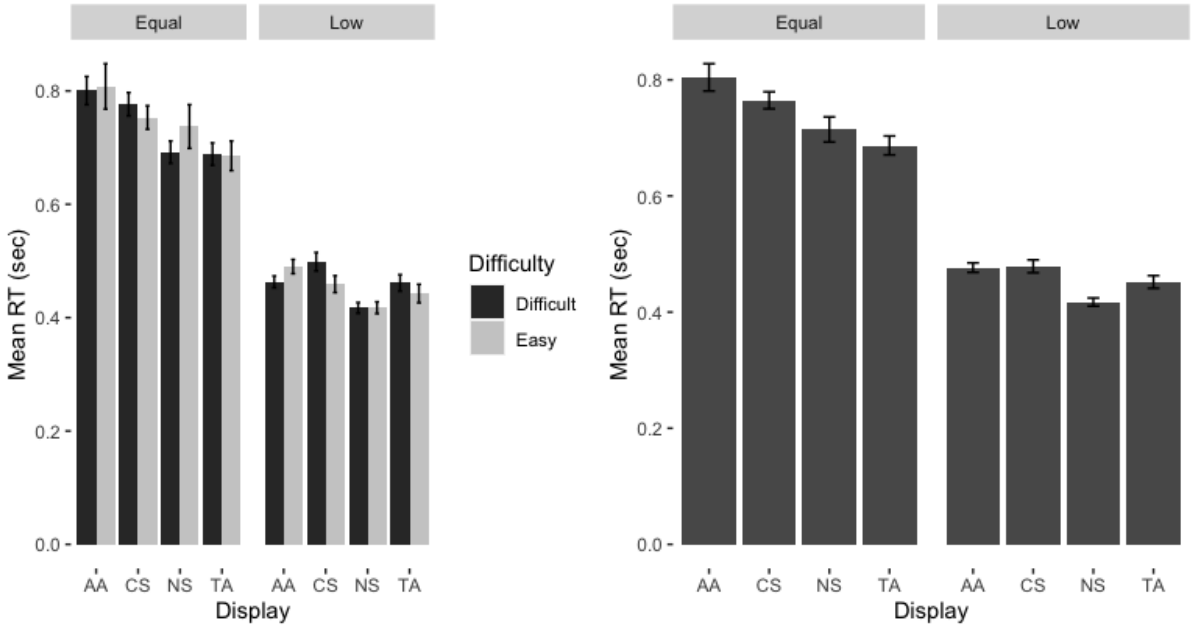
than participants in the equal prevalence condition (M = 0.74 seconds, SD = 0.27). This is unsurprising given prior research in the low prevalence effect, in which participants are faster to make a judgment than those in higher prevalence conditions.

Additionally, there was strong evidence in favor of a main effect of display, BF_{10} = 171 and an interaction between display and prevalence , BF_{10} = 4.77 x 10^{6}. Figure 4 focuses on mean response times across display and prevalence conditions. Participants with the Combined Support display (with aid and threshold line) seemed to perform faster (M = 0.77 seconds, SD = 0.30) than participants with the Automation Alone (M = 0.80 seconds, SD = 0.32) for the equal prevalence condition only. However, this effect had moderate evidence against it in a post hoc Bayesian t test, BF_{10} = 0.23. There was also moderate evidence against a difference between the threshold alone (M = 0.68 seconds, SD = 0.19) and no support conditions (M = 0.70, SD = 0.24) , BF_{10} = 0.22. This does not support our hypothesis that the threshold provides a substantial temporal benefit to performance. From Figure 4, it seems that the combined support display and the automated alone conditions were generally slower than the threshold alone and no support conditions, and this was moderately supported by a post hoc t test, BF_{10} = 3.68. This demonstrates that the automation led to slower response times than no automation across prevalence conditions.

Figure 4

Mean Response Times across Conditions

A. B.



Note. Display conditions presented here are Automation Alone (AA), Combined Support (CS), No Support (NS) and Threshold Alone (TA). The bars represent 95 % confidence intervals. The left figure (A) shows mean response times across display, difficulty, and prevalence conditions. The right figure (B) shows mean response times across display and prevalence conditions.

Signal Detection Analysis

Similar to the mean response time analysis, we ran two 3-way ANOVA including display (No Support, Automation Alone , Threshold, and Combined Support), prevalence (high and low), and difficulty (easy and difficult) on different Signal Detection parameters (sensitivity, d', and criterion, c).

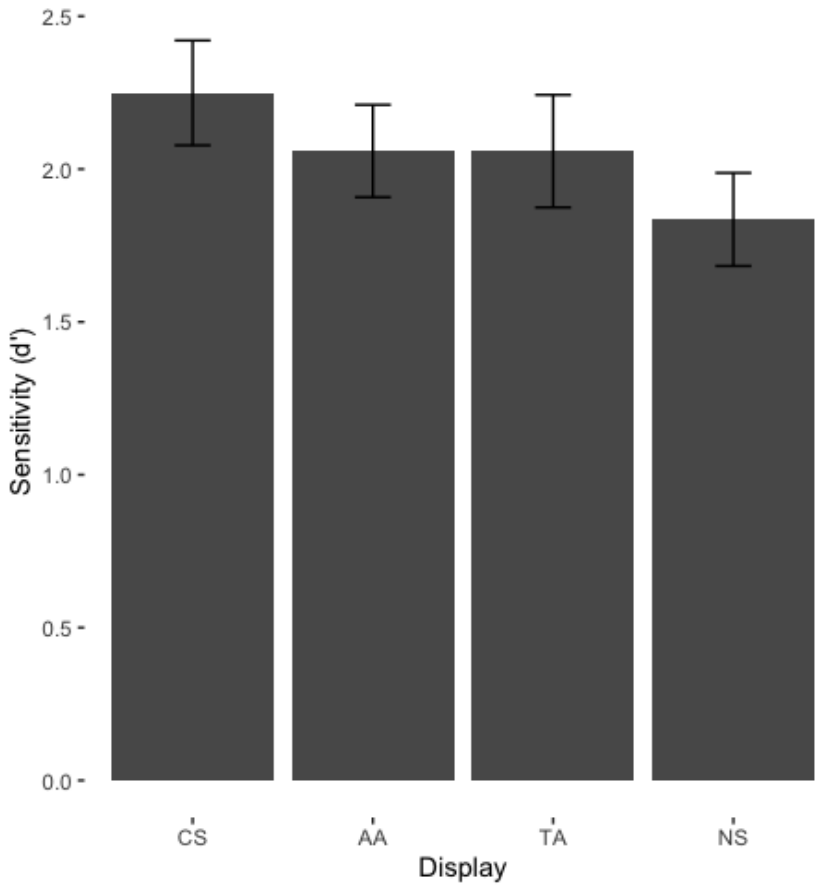
Sensitivity. There was moderate evidence in favor of an effect of display for sensitivity, $BF_{10} = 7.41$. Figure 5 shows sensitivity across display conditions. Participants had higher sensitivity with the combined support display ($M = 2.14$, $SD = 0.14$) than no support ($M = 1.73$, $SD = 0.15$) condition, $BF_{10} = 68.2$. There was weak evidence against a difference between combined support and the threshold alone ($M = 1.81$, $SD = 0.15$), $BF_{10} = 0.49$, and the automation alone condition ($M = 1.96$, $SD = 0.12$), $BF10 = 0.48$. There was

also weak evidence against a difference between the threshold alone condition and the no support condition, $BF_{10} = 0.72$, and weak evidence in favor of a difference between the automation alone condition and the no support condition, $BF_{10} = 1.42$. This suggests that having multiple decision supports improve users' sensitivity substantially compared to no support, but may not substantially improve performance compared to having one form of support. This did not support our hypothesis that having these multiple decision supports would improve participant's sensitivity over having one decision support.

Figure 6 shows the tradeoffs between speed and sensitivity across the display conditions. This plot shows that although displays with the automated aid resulted in slower responses, these displays improved participant's sensitivity. Moreover, this plot shows the benefit in sensitivity that participant receive with the combined support display does not seem to come at the cost of slower response times.

Figure 5

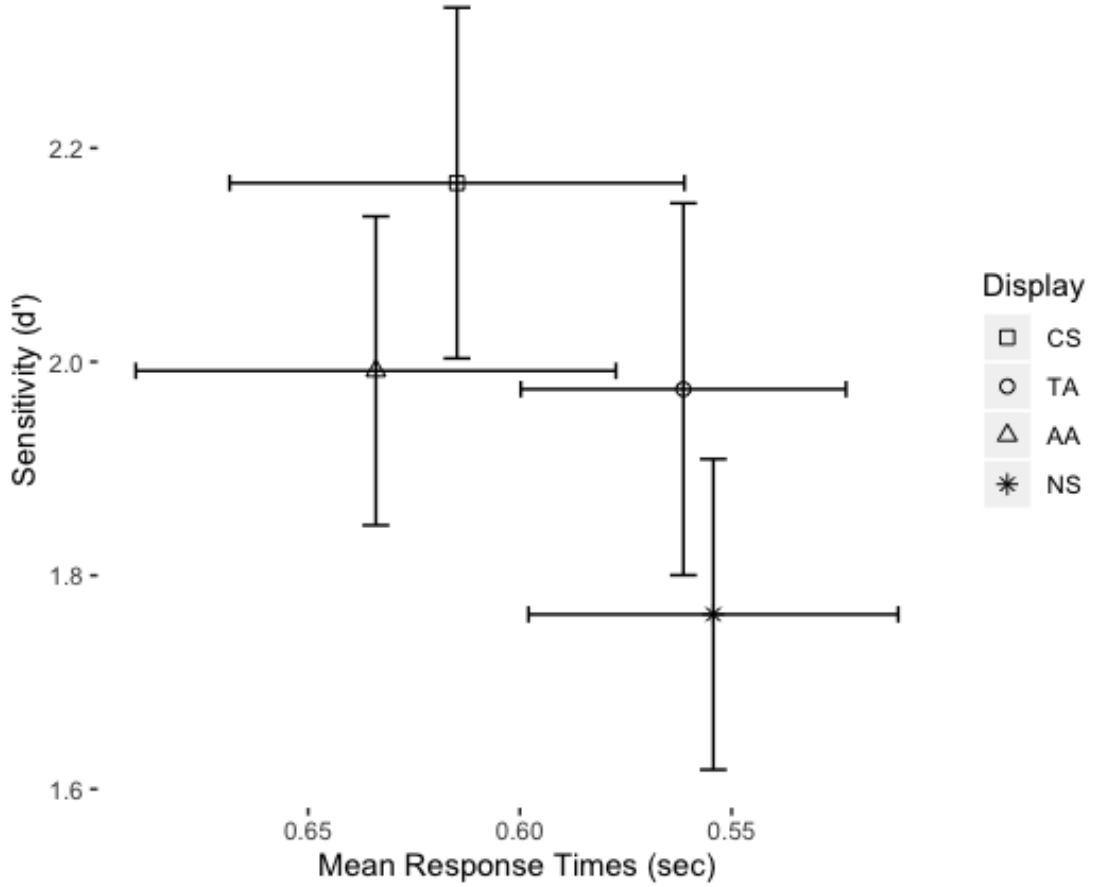
Sensitivity across display conditions



Note. Display conditions presented here are Automation Alone (AA), Combined Support (CS), No Support (NS) and Threshold Alone (TA). The bars represent 95 % confidence intervals.

Figure 6

Mean Response Times by Sensitivity across display conditions



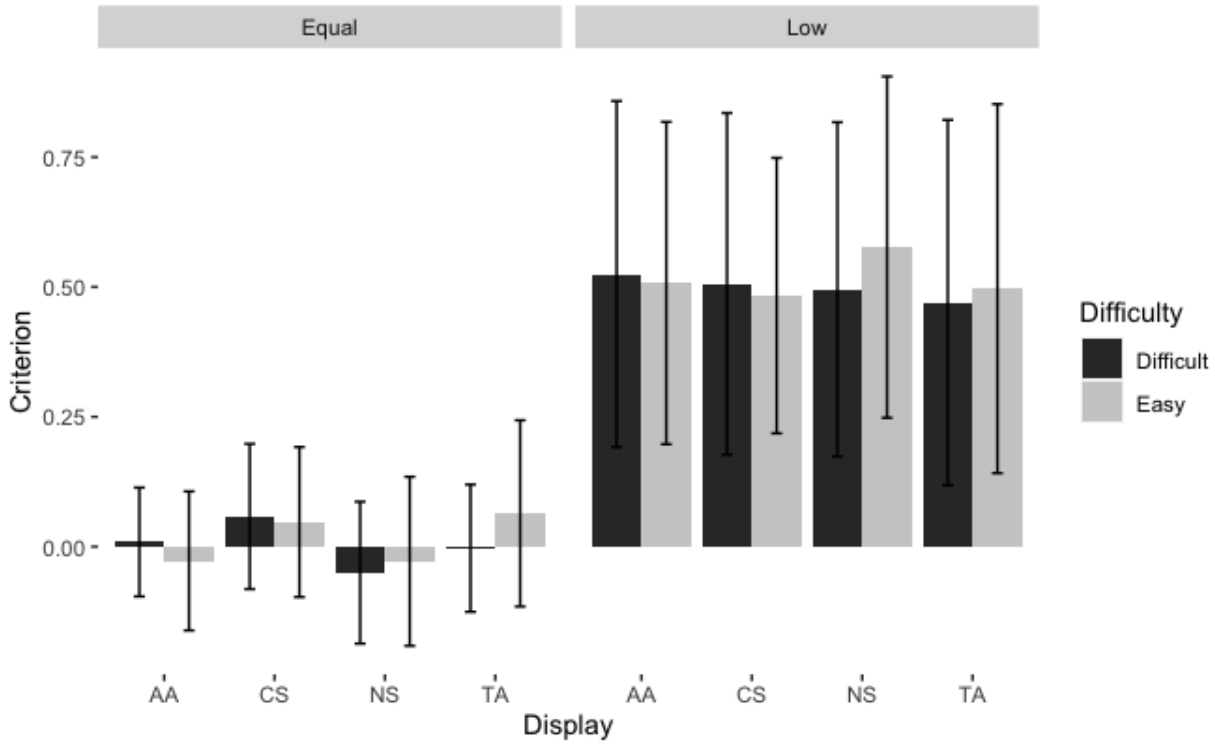
Note. Mean RT is displayed along the x axis with sensitivity displayed along the y axis for each display condition. Better overall performance (e.g., faster RT and higher sensitivity) is towards the upper right side of the graph, while worse overall performance (e.g., slower RT and lower sensitivity) is towards the bottom left side of the graph. The bars represent 95 % confidence intervals.

Criterion. We investigated whether there is a shift in participant's criterion in the low prevalence condition as predicted by prior research on the low prevalence effect. Figure 7 shows mean criterion values across display, prevalence, and difficulty conditions. There was strong evidence in favor of an effect of prevalence $BF_{10} = 1.87 \times 10^{19}$. Upon further

analysis, we found that individuals in the low prevalence condition (M = 0.57, SD = 0.28) had higher criterion (i.e.., higher bias towards responding short than long) than the equal prevalence conditions (M = 0.02, SD = 0.14), BF10 = 3.65 x10^{82}. This indicates a prevalence effect similar to prior research, in which individuals had a shifted criterion or bias in favor of the more prevalence response.

Figure 7

Criterion values by display, difficulty and prevalence conditions



Note. Display conditions presented here are Automation Alone (AA), Combined Support (CS), No Support (NS) and Threshold Alone (TA). The bars represent 95 % confidence intervals.

There was strong evidence supporting an interaction between display and prevalence, BF_{10} = 4.67 x 10^{16}. For the equal prevalence condition, participants had a higher positively biased criterion with the combined support display (M = 0.07, SD = 0.14) than participants in the no support condition (M = -0.03, SD = 0.15), $BF10$ = 26.91, and the automation alone condition (M = 0.002, SD = 0.12), BF_{10} = 3.18. Participants also had a higher criterion with the threshold alone condition (M = 0.04, SD = 0.15) than with the no support display but

this was weakly supported, $BF_{10} = 2.49$. There was moderate evidence against a difference between the combined support display and the threshold alone display, $BF_{10} = 0.28$. For the low prevalence condition, there was moderate evidence against any differences between the display types, $0.19 < BF_{10} < 0.25$. This does not support our hypothesis that the threshold leads users to be less bias and instead shows that users may be more biased with the threshold as in the case with the equal prevalence condition.

Workload Capacity Analysis

Capacity Coefficient. Workload capacity in this study is measuring processing efficiency with the automated aid compared to without. In other words, workload capacity is dependent on using both aided and unaided conditions in a single measure. We estimated the processing efficiency of the automation alone relative to no support. Likewise, we estimated the processing efficiency of the combined support relative to the threshold alone condition. Because two display formats are used to estimate the workload capacity, display will be in terms of threshold present (combined support/threshold alone) or threshold absent (automation alone/no support). To measure whether there were differences in participants' performance with the automated aid across threshold (present/absent), difficulty, and prevalence, we ran a 3-way Bayesian ANOVA on capacity coefficient z scores, C_z.

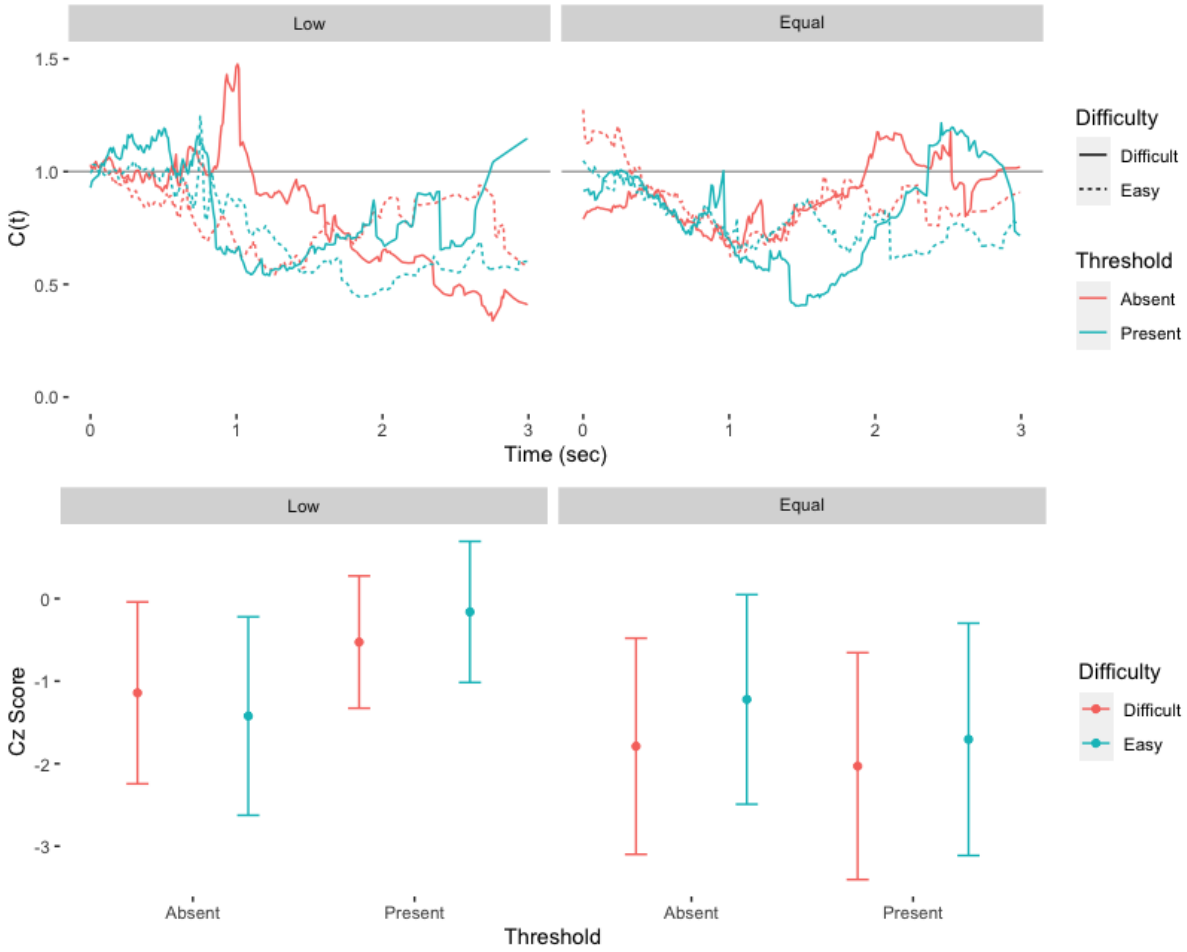
Figure 8 shows both the capacity coefficient functions and the capacity z scores for the threshold, prevalence, and difficulty conditions. The capacity coefficient functions show how efficient participants are processing information with the automated aid compared to a horizontal baseline (e.g., with the automated aid). As mentioned previously, functions near the baseline indicate similar processing with the automation compared to without, referred to as unlimited capacity. Functions lying above the baseline indicate more efficient processing of information (super capacity) and those below indicate less efficient processing of information. When visually inspecting these functions, it is best to focus on the areas of the function where a majority of the response times lie (e.g., the mean responses times) because

they provided the best estimation of capacity. In case of this study, this area lies around 500 milliseconds.

Visually inspecting the capacity coefficient functions and capacity z scores, participants are performing at unlimited ($C(t) = 1$ or $C_z = 0$) to limited ($C(t) < 1$ or $C_z < 0$) capacity. This indicates that participants are performing similarly, if not less efficiently with the automated aid relative to without. This also supports the mean response time results, which found that participants performed slower with the automated aid than without.

Figure 8

Capacity Coefficient functions, C(t), and C_z Scores across Threshold, Difficulty, and Prevalence conditions.



Note. The top graph shows the capacity coefficient functions, $C(t)$, across conditions. The bottom graph shows capacity z scores, C_z scores, across conditions. The bars represent 95 % confidence intervals.

There was moderate evidence against an effect of the threshold line on capacity coefficients, $BF_{10} = 0.26$. This contradicted our hypothesis that participants would perform more efficiently with Combined Support display than with the Automated Aid alone. There was strong evidence against an effect of the threshold line and difficulty conditions on capacity coefficients, $BF_{10} = 0.01$. This contradicted our hypothesis that participants would perform better with the Combined Support display in the difficult condition. There was moderate evidence against an effect of the threshold line and prevalence on capacity coefficients, $BF_{10} = 0.33$. This contradicted our hypothesis that participants would perform more efficiently with the Combined Support display than with the automated aid alone in conditions with more uncertainty.

These results showing a similarity between displays with and without the threshold are unsurprising given the mean response time results showed that there might not be a substantial benefit to the combined support relative to the automation alone, and likewise between the threshold alone and no support displays.

Assessment Functions

To measure whether there were differences in participants' performance with the automated aid across threshold, difficulty, and prevalence conditions, we ran a 3-way Bayesian ANOVA on fPCA component scores from the assessment functions. fPCA of the assessment functions was calculated using the "sft" package in R (Houpt et al., 2014). Using a scree plot, we identified that 2 components were sufficient to describe the variability in the data. With the fPCA of the assessment functions, there were 2 components that describe the early and mean response times. For the mean response times, there were similar results to the capacity coefficient results. There was moderate evidence against an effect of display, $BF_{10} = 0.12$. Additionally, there was strong evidence against an interaction of display and prevalence, $BF_{10} = 0.005$, and an interaction of display and difficulty, $BF_{10} = 0.006$. For the early response times, there was similarly moderate evidence against an effect of display, BF_{10}

= 0.30. There was strong evidence against an interaction of display and prevalence, BF_{10} = 0.06, and an interaction of display and difficulty, BF_{10} = 0.009. These results, similar to the capacity coefficient results, contradicts our hypotheses.

Visual inspection of the assessment functions is similar to that of the capacity coefficient functions. The assessment functions, shown in Figure 9, provide an interesting contrast to the mean response time and sensitivity results. For the equal prevalence condition, participants are performing worse with the automated aid, and there seems to be no difference with or without the threshold support. This diverges from the sensitivity results that indicated participants were preforming better with the decisional supports relative to no support. In the equal prevalence condition, the slower response times from the decision supports outweighs the potential gain of accuracy, creating worse overall performance. For the low prevalence condition, it seems that participants perform slightly more efficiently with the automated aid when it is paired with the threshold (combined support) than without (automation alone). However, this was not indicated by the fPCA analysis, which could be due to a wide variability in automated aid use. This is further explored later in our *post hoc* analyses. This could mean that there is a potential, though unsubstantial, gain of having both decisional supports over automation alone. Further research is needed to explore these possibility.

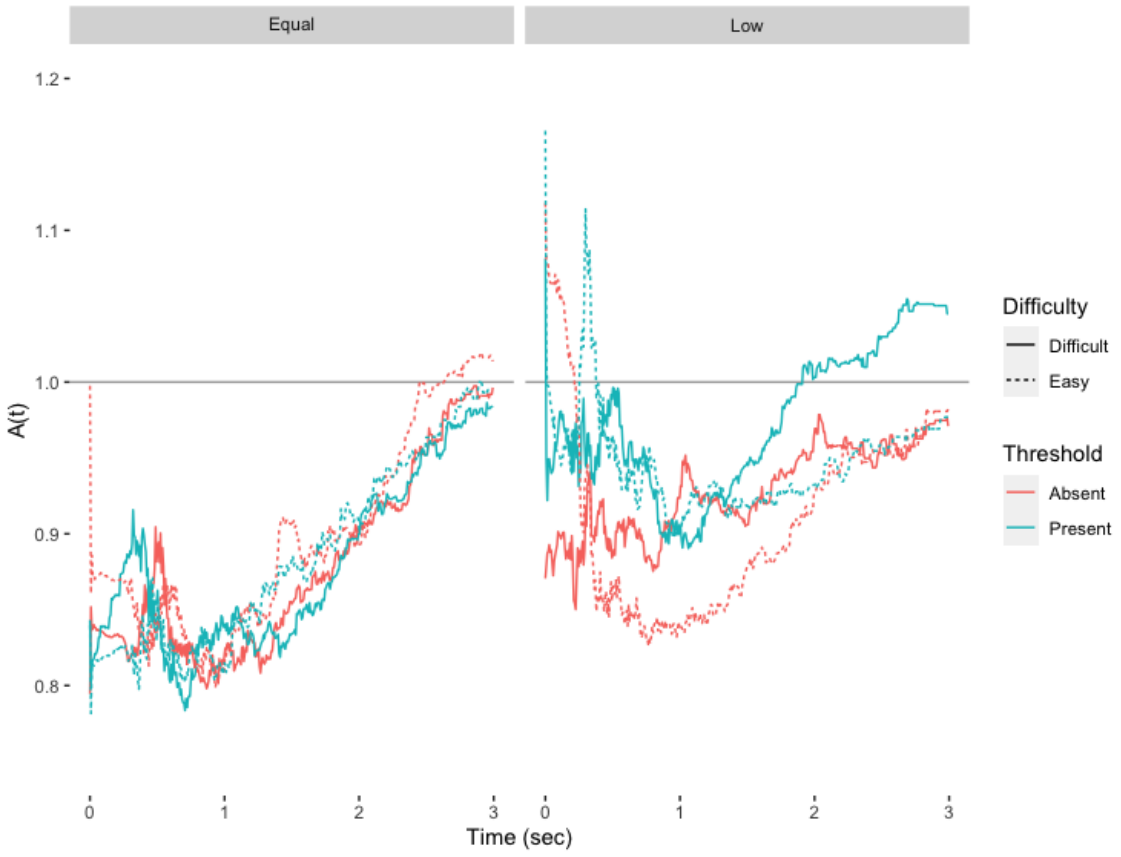
Post-Hoc Exploratory Analysis

Automation Use. We performed a logistic regression to see if the automation's response can be predictive of the human response after accounting for the stimulus. This was done by fitting model that use the automation's response and the signal to predict the person's response using the "brms" package in R (Burkner, 2017). Bayes factor was computed using the marginal likelihoods of this model compared to the marginal likelihoods of a null model (e.g., model with signal alone predicting the person's response). We found that the model with both the automation and the signal was more predictive than a signal

alone model, $BF_{10} > 1x10^{10}$. This means that participants on the group level had to have been using the automated aid throughout the experiment. However there is some variability in its predictiveness across participants, estimated standard deviation of group level effects = 1.06. This indicates that there is likely some variability in participant's use of the automated aid which could influence our ability to detect these effects explored in this study.

Figure 9

Group-level Assessment Functions, A(t), across Threshold, Difficulty, and Prevalence conditions.



Note. This figure shows group level assessment functions across difficulty, prevalence, and threshold conditions. The horizontal line at A(t) = 1 represents the UCIP baseline, where the assessment functions lying below the line indicate limited capacity performance.

Order Effects. One thing we considered was whether the order of conditions may have influenced participants' performance with different display designs. We ran several 4 way ANOVAs to include the effect of order on our mean response time analysis, signal detection analysis, and workload capacity, which are all reported in the Appendix B.

Interestingly, for the capacity coefficient z scores, there was strong evidence in support of an interaction between order and the presence of a threshold line, BF_{10} = 34.15, and even stronger support for a 3 way interaction of order, threshold line, and prevalence, BF_{10} = 54.29. Likewise, for the assessment functions there was strong evidence for both an interaction between order and threshold, BF_{10} = 3648.56 and strong evidence in favor of a 3 way interaction of display, order, and prevalence, BF_{10} = 164.63.

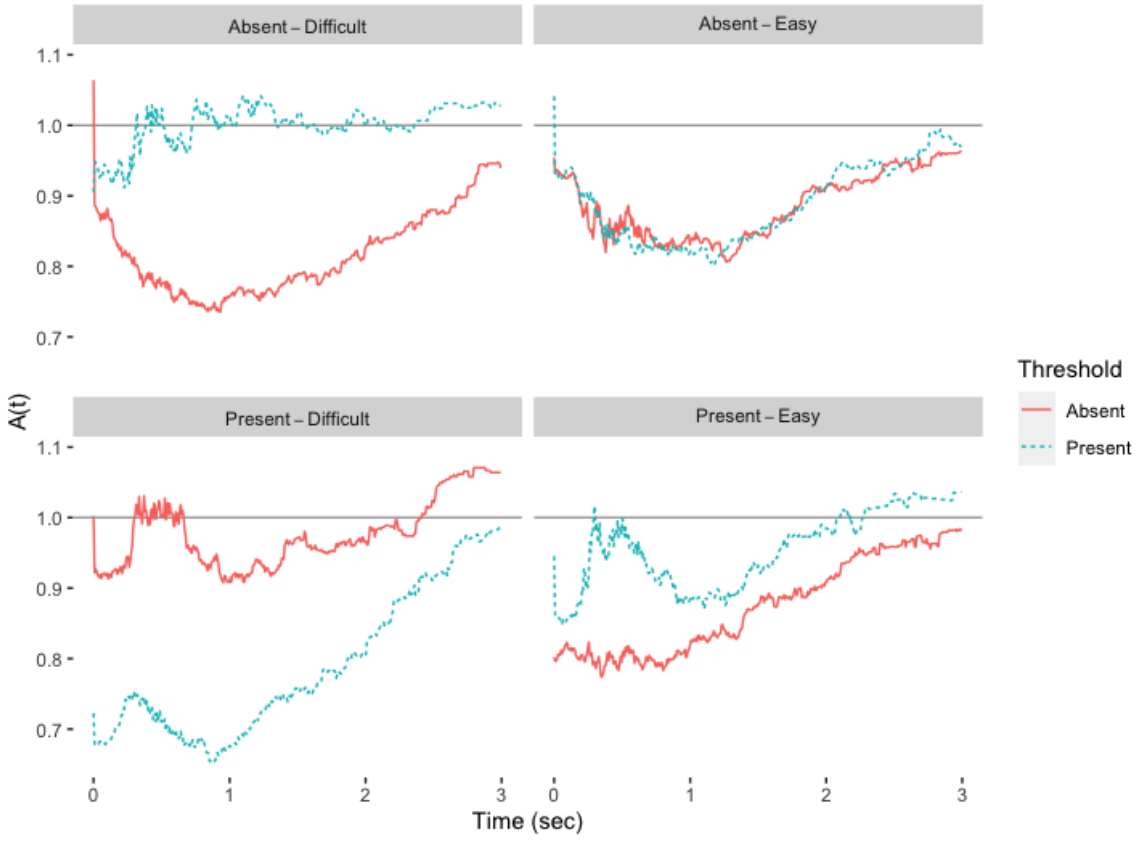
Visual inspection of the group level' capacity coefficients (shown in Figure 8), revealed that at the group level, participants generally performed at unlimited to limited capacity (at or below the baseline). It seems for participants that started with a difficult condition (either with the threshold line present or absent) performed at or above the baseline with the second display they saw. For example, participants who did not receive the threshold line initially (Equal: $M = -2.79$ C_z score, $SD = 3.17$; Low: $M = -1.10$ C_z score, $SD = 1.93$) performed less efficient without the threshold line than with the threshold line present (Equal: $M = -0.04$ C_z score, $SD = 2.35$; Low: $M = -0.52$ C_z score, $SD = 2.11$). Likewise, participants who received the threshold line first performed better without the threshold line (Equal: $M = -0.34$ C_z score, $SD = 2.87$; Low: $M = 0.78$ C_z score, $SD = 2.50$) than with (Equal: $M = -3.68$ C_z score, $SD = 3.24$; Low: $M = -0.04$ C_z score, $SD = 1.65$). This could be indicative of learning effects. However, we do not see similar effects with participants who receive the easy conditions first.

Visual inspection of the assessment functions, shown in Figure 10, shows clear differences between the threshold conditions for the starting difficult order condition. This pattern is similar to the capacity coefficients in which participants perform more efficiently in terms of both response times and accuracy with the display they do not start with, when they start with the difficult condition. This was confirmed with a Bayesian t-test on the fPCA scores. Participants who started with the threshold line absent performed more efficiently with the threshold line ($M = -0.06$, $SD = 0.15$) than without it ($M = 0.09$, $SD = 0.11$), BF_{10} =

398991. Likewise, Participants who started with the threshold line present performed more efficiently without the threshold line ($M = 0.06$, $SD = 0.17$) than with it ($M = -0.12$, $SD = 0.20$), $BF_{10} = 4015.51$. This was not true for participants who started with the easy version of the task. Participants who started without the threshold line in the easy version of the task performed similarly between display conditions (Absent: $M = -0.002$, $SD = 0.16$; Present $M = -0.03$, $SD = 0.18$), $BF_{10} = 0.26$. Participants who started with the threshold line in the easy condition seem to performed more efficiently with the threshold line ($M = 0.006$, $SD = 0.15$) than without ($M = 0.05$, $SD = 0.10$), however, this was not confirmed with the fPCA analysis, $BF_{10} = 0.95$.

Figure 10

Group-level Assessment Functions, A(t), across Threshold and Order Conditions



Note. Each panel represents a different order condition, labelled by which threshold and difficulty condition the participant viewed first. For example, Absent-Difficult in the upper left panel shows the group level assessment functions for both threshold conditions for participants who started with no threshold in the difficult version of the task. The horizontal line at A(t) = 1 represents the UCIP baseline.

Discussion

The purpose of this study was to examine the effects of display design and target prevalence on human performance with an automated aid. We found no evidence in favor of an effect of display, nor an interaction between display and difficulty or display and prevalence. This contradicts our hypotheses that participants would perform more efficiently with multiple decision supports, especially in situations with more uncertainty. Additionally, we did not find that the threshold improves participant's response times and minimizes response bias. We found strong evidence that participants were slower with the combined support and the automated aid compared to the threshold alone and no support conditions. However, participants were better at discriminating the short and long bars with the combined support than with the no support display. This suggests that participants were slower but more accurate with these decision supports compared to no support. Further, there was no substantial benefit in terms of sensitivity with the combined support display over a display that contained a single form of support (e.g., Automation Alone or Threshold Alone).

With the workload capacity analysis, we found that participants were consistently performing at unlimited to limited capacity, suggesting they are performing with similar if not with less efficiency with the automated aid relative to without. This is consistent with the mean response time analysis that showed participants performing slower with the automated aid and combined support. Interestingly, participants with and without the threshold performed at limited capacity in terms of the assessment functions, which are based on both response time and accuracy. This suggests that the accuracy benefit of the decisional supports shown in the sensitivity results may not outweigh the temporal disadvantages they produce.

Just looking at these interactions, we might conclude that having multiple decision supports do not necessarily improve performance, and that users may only need one form of decision support. However, we did find consistent interactions with order, display, and prevalence across our analyses, which suggests that there may be additional influences that

contribute to the effectiveness of the decision support aids. Generally, it seems that the order that participants viewed the conditions played a major role in their performance with the threshold line and the automated aid. The assessment function results provide the most comprehensive analysis of how participants are performing with the automated aid, taking into account both response times and accuracy. From these results, it seems that participants perform with similar efficiency if not more efficiently with the automated aid when combined with the other decisional support. However, this was not true for one order condition (when participants started with the threshold in the difficult condition), in which participants performed less efficiently with the combined automated aid and threshold line, then when processing the automated aid alone. These results suggest that there may be an added benefit to having multiple decisional support aids under certain circumstances.

Little research has been done to explore the effect of the order that one receives displays on performance with displays, as typically participants in these experiments only experience one display design (e.g., Meyer, 2001; Yamani & McCarley, 2018). However, there has been extensive research on adoption of new technology and interface designs in real world context (e.g., Andriessen, 2012). Similar to the effects observed in this study, users experience and grow accustom to one display design before adopting a new design which may change how they approach tasks. Andriessen (2012) highlights how the adoption of new technology is dependent on a number of factors beyond the objective utility of the new technology, namely how the user *perceives* the usefulness of this new technology/interface and the ease of which a user can integrate this technology into current work practices. This often requires more attention paid to the introduction and training with new technology, both of which were not considered in this study, which may have contributed to poor adaptation to the new displays.

The variability in performance over the order could be a result of user stickiness to a particular strategy that may complement one display over another. As shown with the

assessment functions, participants generally similarly, if not better, with the threshold line than without, with the exception of starting with the threshold line in the difficult condition. A user who first experiences the threshold line in the difficult condition might be less willing to use the threshold line in their judgments, as it will be less accurate than in the easy condition. As mentioned in the methods section, the accuracy of using just the threshold line is different across difficulty conditions. Therefore, a participant starting in the difficult condition would be less willing to use it, and may benefit more from ignoring it. Meyer (2001) briefly discussed how strategy differences with the aid could change the effectiveness of a display. For example, participants who decide to not use the threshold line may benefit from displays that do not contain the threshold line, because the threshold line is not there to distract them. Likewise, participants who use the automated aid more might benefit from the aid's recommendation being displayed closer to other relevant information. For these participants, they might benefit more from having the Combined Support display. Participants who decided to ignore the automated aid more, might perform better with a more No Support display or the Threshold Alone display, because the automated aid is not available.

These results also provide insights on the relationship between target prevalence and automated aids. For example, the implementation of an automated aid or utilizing an effective display could weaken the effects of low prevalence in the real world (e.g., increase the sensitivity of the participants). 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 for computer aided detection systems that assist radiologists in detecting areas of interest in medical imagery. 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 consequences of low prevalence: the use of decision supports. As demonstrated in this study, the low prevalence effect might weaken when participants have a decision support (e.g., the automation and the threshold) to assist them in their decisions. Though participants' criteria were still shifted in the low prevalence condition, participants had least the same, if not greater sensitivity with the decision supports, which would result in fewer misses. The effectiveness of the threshold line and the combination of the threshold line and the automated aid was strongly dependent on when the participants experienced that display. Further research should explore not only the use of the automated aid, but how it is introduced in determining its effectiveness.

Some research has focused on the use of cues to alleviate the effects of prevalence (e.g., miss rates) 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, may not be feasible or beneficial 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 and a visual decision support is more feasible and beneficial for real-world situations.

It is important to note that the automated aid tested in this study were not necessarily sensitive to indicating an the infrequent target and had similar false alarms and misses across conditions. This is unlike several automated aids used in the real world that are highly sensitive and have more false alarms with low prevalent targets. For example, Parasuraman, Hancock and Olofinboba (1997) found that collision warning systems are highly sensitive to the low probable event of a collision, lessening the effectiveness of an aid for a user to respond. The performance of the automation must be taken into account when applying automated aids in low prevalent conditions. Though automated aids may be imperfect, they can still be useful for users (Wickens & Dixon, 1997). However, it is unclear how display design can help alleviate the effects of highly sensitive aids in those conditions. One next step in this research would be to consider how the impact of highly sensitive aids in low prevalent conditions can be alleviated by display design. Moreover, it should be considered that training with different display designs might have differential effects on performance with a given display, which should be further explored.

Limitations and Future Directions

There are several limitations to this study. First, the automated aid's cue in automated aid only display covers a larger space (the entire bar) than with the Combined Support

display that only covers the smaller threshold line. 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 Combined Support. 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 without the threshold line, then we might determine the cue in the automated aid alone display to be more salient than in the Combined Support display. 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, we hope to replicate these designs with a more complex task that encourages participants to rely more on the automated aid.

With participants adopting a mix of strategies, it becomes difficult to determine the effect of display design on performance. It is possible that participants in some cases (e.g., with the easy task) may be more willing to adopt and stick with a strategy that works for the first display. Because the current study did not evaluate all possible strategies in the tasks, we 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. In terms of display design, these results indicated that integrating the automated aid with a visual form of

decision support might improve performance with the aid but this effect may be strongly contingent on how and when this new display is introduced. Further research is needed to understand the role of condition order on performance with these displays.

Conclusion

We sought to investigate the effects of prevalence and display type on human performance with an automated aid. The results indicated that having decision supports may be costly in terms of response times, but may also be beneficial for increasing user sensitivity. However, this benefit in sensitivity may not outweigh the cost of the slower response times, as indicated by the workload capacity analysis. Based on these results, we suspect that there may be a benefit to having more than one decisional support, but this may be dependent on when and how a participant uses these decisional supports. 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.

Acknowledgements

Declarations

Funding: The authors did not receive support from any organization for this work.

Conflicts of interest: The authors have no conflicts of interest to declare that are relevant to the content of this article.

Ethics Approval: This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Institutional Review Board of Wright State University (Approval Number: 06267)

Consent to participate: Informed consent was obtained from all individual participants in the study.

Availability of data and material (data transparency): Data and analysis code is available on the Open Science Framework.

Authors' contributions: C.K. conceived the ideas and designed the experiments for the presented work. J.H. supervised the project and recommended the data analysis approach for this project. K.B. conceived the ecological display design for the experiments and assisted in analyzing the data. C.K. took lead in writing the manuscript. All authors discussed the results and contributed to the final manuscript.

References

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

Blaha, L.M., Townsend, J.T., Kneeland, C.M., & Houpt J.W. (Under Review). Capacity Coefficient Analysis for Single-Target Self-Terminating Processes. *Journal of Mathematical Psychology.*

Botzer, A., Meyer, J., Bak, P., & Parmet, Y. (2010). User settings of cue thresholds for binary categorization decisions. *Journal of Experimental Psychology: Applied, 16*(1), 1-15

Burns, D. M., Houpt, J. W., Townsend, J. T., & Endres, M. J. (2013). Functional principal components analysis of workload capacity functions. *Behavior Research Methods,45*, 1048-1057.

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.

Donkin, C., Little, D. R., & Houpt, J. W. (2014). Assessing the speed-accuracy trade-off effect on the capacity of information processing. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1183.

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., Mendoça, 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).

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

Getty, D. J., Swets, J. A., Pickett, R. M., & Gonthier, D. (1995). System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. *Journal of experimental psychology: applied*, *1*(1), 19.

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.

Houpt, J. W., & Townsend, J. T. (2012). Statistical measures for workload capacity analysis. *Journal of Mathematical Psychology*, *56*(5), 341-355.

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.

Little, D., Altieri, N., Fific, M., & Yang, C. T. (Eds.). (2017). *Systems factorial technology: A theory driven methodology for the identification of perceptual and cognitive mechanisms*. Academic Press.

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

Parasuraman, R., Hancock, P. A., & Olofinboba, O. (1997). Alarm effectiveness in driver-centred collision-warning systems. *Ergonomics, 40*(3), 390-399.

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.

Rich, A. N., Kunar, M. A., Van Wert, M. J., Hidalgo-Sotelo, B., Horowitz, T. S., & Wolfe, J. M. (2008). Why do we miss rare targets? Exploring the boundaries of the low prevalence effect. *Journal of vision, 8*(15), 15-15.

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. 356–374

Russell, N. C. C., & Kunar, M. A. (2012). Colour and spatial cueing in low-prevalence visual search. *Quarterly Journal of Experimental Psychology, 65*, 1327–1344.

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

Townsend, J. T., & Altieri, N. (2012). An accuracy–response time capacity assessment function that measures performance against standard parallel predictions. *Psychological review*, *119*(3), 500.

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

Table 1

Descriptive statistics (Mean and Standard Deviations) across conditions

Condition	Mean RT	Criterion	Sensitivity	Hits	Miss	FA	CR
Low Prevalence							
No Support Display							
Easy	0.43 (0.40)	0.57 (0.33)	1.88 (0.95)	6.86 (5.09)	13.00 (6.01)	3.41 (3.50)	176.72 (8.15)
Difficult	0.43 (0.35)	0.50 (0.32)	1.96 (0.92)	7.34 (4.76)	13.66 (5.47)	3.24 (3.59)	175.76 (5.82)
Threshold Alone							
Easy	0.48 (0.62)	0.50 (0.36)	2.26 (0.95)	8.79 (5.49)	10.41 (5.47)	3.10 (2.97)	177.69 (6.03)
Difficult	0.47 (0.58)	0.47 (0.35)	2.27 (1.03)	9.41 (5.78)	11.20 (5.90)	3.72 (4.01)	175.66 (6.27)
Automation Alone							
Easy	0.51 (0.49)	0.51 (0.31)	2.09 (0.88)	9.38 (6.49)	12.03 (6.01)	3.59 (3.52)	175.00 (7.83)
Difficult	0.48 (0.41)	0.52 (0.33)	2.16 (0.99)	9.34 (6.73)	10.90 (5.57)	3.79 (4.48)	175.97 (6.18)
Combined Support							
Easy	0.47 (0.53)	0.58 (0.27)	2.39 (0.80)	9.24 (4.93)	9.48 (4.87)	2.76 (2.79)	178.52 (4.47)
Difficult	0.50 (0.65)	0.51 (0.32)	2.27 (1.10)	8.97 (6.18)	10.83 (6.57)	3.14 (3.11)	177.07 (6.08)
Equal Prevalence							
No Support Display							
Easy	0.74 (1.42)	-0.02 (0.16)	1.77 (0.68)	78.34 (13.45)	22.16 (12.64)	18.23 (8.48)	81.27 (11.24)
Difficult	0.69 (0.72)	-0.5 (0.13)	1.71 (0.61)	80.61 (12.07)	22.15 (10.09)	19.19 (8.77)	78.04 (12.38)
Threshold Alone							
Easy	0.69 (0.96)	0.06 (0.18)	1.86 (0.90)	81.34 (12.90)	17.35 (11.39)	22.77 (15.95)	78.54 (16.07)
Difficult	0.69 (0.62)	-0.003 (0.12)	1.74 (0.93)	79.08 (15.79)	21.04 (13.65)	20.50 (12.85)	79.38 (12.45)
Automation Alone							
Easy	0.81 (1.47)	-0.03 (0.13)	2.02 (0.66)	82.03 (11.55)	17.85 (9.10)	15.85 (9.07)	84.27 (12.09)
Difficult	0.80 (0.91)	0.01 (0.10)	1.94 (0.57)	77.19 (13.57)	18.50 (9.31)	16.54 (8.45)	87.77 (12.01)
Combined Support							
Easy	0.75 (0.76)	0.05 (0.14)	2.15 (0.80)	84.77 (11.98)	14.50 (9.92)	17.62 (12.50)	83.12 (14.36)
Difficult	0.78 (0.76)	0.06 (0.14)	2.17 (0.90)	84.31 (15.15)	14.58 (13.30)	17.46 (12.08)	83.65 (13.14)

Table 2

Descriptive statistics (Mean and Standard Deviations) across conditions

Condition	CC	Component 1	Component 2
Low Prevalence			
Threshold Present			
Easy	0.16 (2.25)	0.04 (0.08)	-0.04 (0.18)
Difficult	-0.52 (2.10)	0.02 (0.11)	-0.03 (0.15)
Threshold Absent			
Easy	-1.42 (3.16)	-0.01 (0.13)	0.03 (0.16)
Difficult	-1.14 (2.90)	0.02 (0.13)	0.002 (0.18)
Equal Prevalence			
Threshold Present			
Easy	-1.70 (3.15)	0.01 (0.22)	0.02 (0.09)
Difficult	-2.03 (3.18)	-0.06 (0.24)	0.01 (0.11)
Threshold Absent			
Easy	1.22 (2.86)	-0.01 (0.20)	0.01 (0.08)
Difficult	1.79 (3.10)	-0.01 (0.21)	0.01 (0.07)

Appendix B

Table 1

Bayes factor for each effect from a 4 way Bayesian ANOVAs on Mean RT, Criterion, and Sensitivity

	Mean RT	Criterion	Sensitivity
Display	**171.07**	0.01	**10.23**
Difficulty	0.11	0.12	0.12
Prevalence	**2.01×10^{6}**	**3.03×10^{22}**	0.26
Order	0.42	0.17	0.05
Display x Difficulty	0.79	0.0001	0.03
Display x Prevalence	**4.87×10^{6}**	**1.01×10^{20}**	0.17
Display x Order	**1.25×10^{8}**	**3.43×10^{5}**	**3.04×10^{88}**
Difficulty x Prevalence	**3100.15**	**6.58×10^{20}**	0.01
Difficulty x Order	**309.92**	0.002	>0.001
Prevalence x Order	**3.19×10^{4}**	**9.57×10^{19}**	0.01
Display x Difficulty x Prevalence	**200.15**	**9.41×10^{15}**	**4.78×10^{6}**
Display x Prevalence x Order	**3.19×10^{12}**	**1.27×10^{43}**	**9.86×10^{87}**
Display x Difficulty x Order	**1.66×10^{10}**	**8.30**	**3.89×10^{83}**
Difficulty x Prevalence x Order	**6.09×10^{5}**	**1.03×10^{16}**	> 0.001
Display x Difficulty x Prevalence x Order	**9.74×10^{8}**	**6.47×10^{35}**	**3.15×10^{81}**

Note. Bolded values indicate evidence in favor of the alternative hypothesis.

Table 2

Bayes factor for each effect from the Bayesian ANOVAs on C_z scores, and the components from the fPCA analysis on the Assessment Functions.

	CC	Component 1	Component 2
Threshold	0.26	0.12	0.65
Difficulty	0.19	0.20	0.16
Prevalence	0.60	0.37	0.32
Order	0.43	0.15	0.12
Threshold x Difficulty	0.01	0.01	0.02
Threshold x Prevalence	0.31	0.02	0.11
Threshold x Order	**32.17**	**3.22 x 10^5**	**2.57**
Difficulty x Prevalence	0.03	0.02	0.01
Difficulty x Order	0.003	0.004	0.002
Prevalence x Order	0.03	0.008	0.008
Threshold x Difficulty x Prevalence	0.001	>0.001	>0.001
Threshold x Prevalence x Order	**57.83**	**1.45 x 10^5**	0.01
Threshold x Difficulty x Order	0.11	**1.05 x 10^4**	0.002
Difficulty x Prevalence x Order	> 0.001	> 0.001	>0.001
Threshold x Difficulty x Prevalence x Order	0.001	**24.68**	>0.001

Note. Component 1 is on deviations of the assessment functions around the mean response time, and Component 2 is on deviations of the assessment functions before the mean response times.

Decision Support:

No Support (NS) -- bar graph;

Threshold Alone (TA) -- bar graph and threshold line;

Automation Alone (AA) -- bar graph and automated aiding

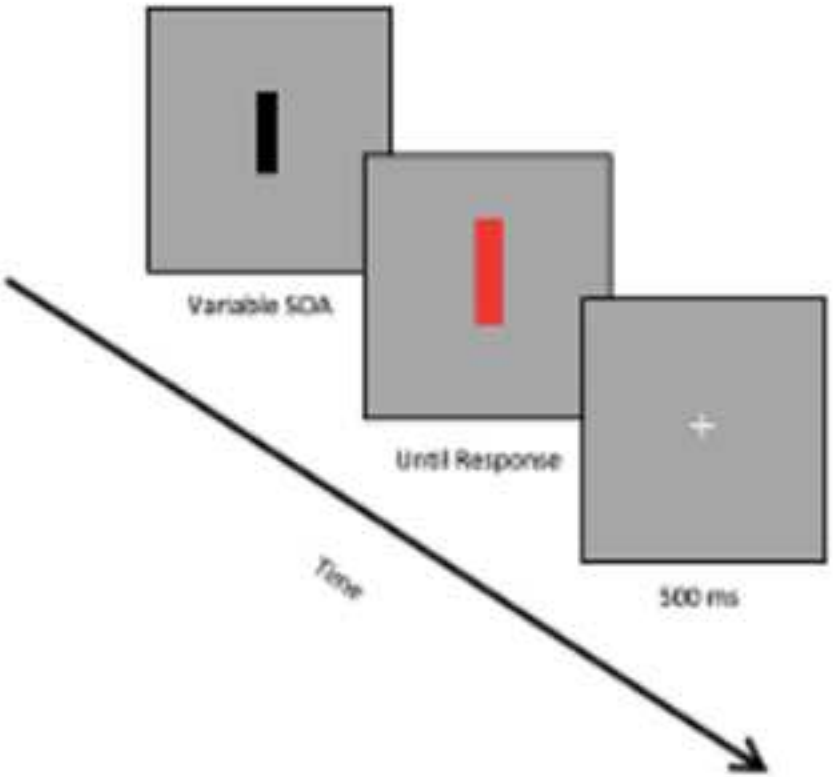
Combined Support (CS) -- bar graph, threshold line and automated aiding

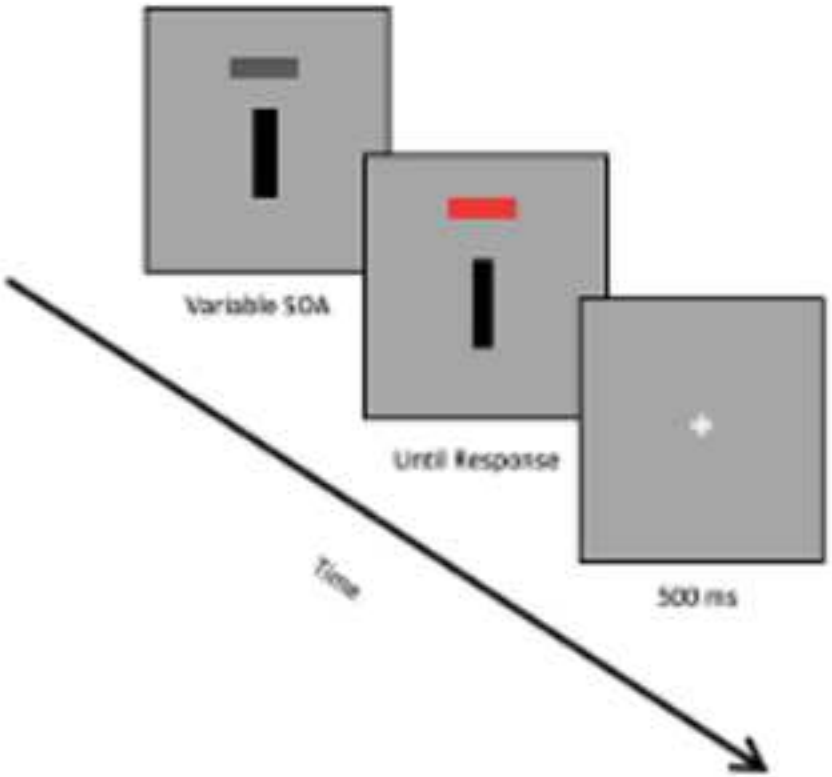
Task Difficulty: Easy, Difficult

Prevalence		Block 1-8	Block 9-16	Block 17-24	Block 25-32
Low		TA / CS Easy	TA / CS Difficult	NS / AA Easy	NS / AA Difficult
		NS / AA Easy	NS / AA Difficult	TA / CS Easy	TA / CS Difficult
		TA / CS Difficult	TA / CS Easy	NS / AA Difficult	NS / AA Easy
		NS / AA Difficult	NS / AA Easy	TA / CS Difficult	TA / CS Easy
Equal		TA / CS Easy	TA / CS Difficult	NS / AA Easy	NS / AA Difficult
		NS / AA Easy	NS / AA Difficult	TA / CS Easy	TA / CS Difficult
		TA / CS Difficult	TA / CS Easy	NS / AA Difficult	NS / AA Easy
		NS / AA Difficult	NS / AA Easy	TA / CS Difficult	TA / CS Easy

Figure

Click here to access/download;Figure;Figure1.png





Figure