Effects of Decision Type and Aid Accuracy on User Performance

Lori Mahoney
Wright State University - Main Campus, mahoney.32@wright.edu

Joseph W. Houpt
Wright State University - Main Campus, joseph.houpt@wright.edu

EFFECTS OF DECISION TYPE AND AID ACCURACY ON USER PERFORMANCE

Lori Mahoney
Joseph Houpt
Wright State University
Dayton, OH

Automated aids provide users additional information for making decisions. The way the aid presents the information requires the user to either make the same decision as unaided or to agree or disagree with the aid's recommendation. In this study, we measured response times and accuracy without an aid and with an aid where either: 1) the subject makes the same decision as the unaided condition, or 2) the subject agrees or disagrees with the automated aid's decision. Results show subjects were more accurate with direct selection decisions, more accurate aids, and easier tasks, with an interaction between decision type and aid accuracy. Subjects were faster with direct selection decisions and more accurate aids, with an interaction between decision type and aid accuracy. Using a cognitive model we found information accumulation rates and caution varied across conditions.

The addition of an automated aid for speeded choice tasks gives the user additional information to make their decision; a correct aid response leads to a faster and more accurate human response (Rovira, McGarry & Parasuraman, 2007; Wickens, Clegg, Vieane & Sebok, 2015), but a number of different factors, including trust, workload, and automation reliability influence automation use, disuse (e.g. underutilization or neglect) and misuse (e.g. overreliance or complacency) (Parasuraman & Riley, 1997). Multiple studies have shown that decreased aid reliability decreases human performance (Rovira, McGarry & Parasuraman, 2007; Rovira, Cross, Leitch & Bonaceto, 2014; Wickens, Clegg, Vieane & Sebok, 2015). Rovira, et al (2014) showed there was little variability in response accuracy for low task demand; however, for high task demand accuracy improved with reliable automation and did not degrade below manual performance with imperfect automation.

Another factor to consider is that the way an automated aid presents the information requires the user to either make the same decision as they would unaided (i.e. select the correct signal) or to agree or disagree with the aid's recommendation. Parasuraman, Sheridan, & Wickens (2000) proposed a 10-level model for automation of decision and action selection functions, such as the output of an automated target recognition system or an aircraft-ground collision avoidance system. The model distinguishes between level 4 where the automation suggests one recommendation, and level 5, where the automation executes the one recommedation if the human user approves. In both cases the user is presented with one option to make a decision, but how the information is presented differs and requires the user to make a different decision. We hypothesize that the agree/disagree (level 5) decision decreases user performance and that there is an interaction with aid accuracy. In the agree/disagree condition an inaccurate or less accurate aid requires the user to determine if the aid provided a correct recommendation by comparing it to their own decision (i.e. make at least two decisions); when this is done in series the response time (RT) is slower for this condition. For the higher accuracy aid the user can select the aid's recommendation without knowing if it agrees with theirs because

they know the aid is almost always correct; users will not need to compare the aid's recommendation to theirs for every signal, so their perfomance is improved. The direct selection (level 4) decision does not require the user to evaluate the aid against their recommendation so the user is making one less decision in this condition regardless of the aid's accuracy, therefore we expect the RT to be faster for this condition. We expect this difference in decision type to also affect accuracy. We predict the level 5 decision accuracy will be lower, especially for the less accurate aid, since the user is basing their decision on their own recommendation. Additionally, since the agree/disagree decision is asking a different question of the user than the decision to select the correct signal, either with or without an aid, we expect that the mean drift rate, the response boundary, or both should vary across the test conditions for a linear ballistic accumulator (LBA) model, in which results are accumulated linearly and independently for all responses. The response boundary should be larger for conditions with more difficult tasks and/or less accurate aids since the user will have less certainity in the correct decision and be more cautious in making their decision. The mean drift rate should be lower for the agree/disagree decision condition with the less accurate aid since users are making more decisions and are less efficient. The mean drift rate should be higher for the more accurate aid since users are more efficient with an aid that is almost always correct.

Method

Forty-seven students (32 females, 15 males) from Wright State University participated in this study. All subjects gave informed consent to participate and were given course credit as compensation for their time. Ages ranged from 18 to 36 (M = 19.9). Data from 3 participants were removed due an overall accuracy of less than 70%. For the remaining 44 participants, trials were removed that had response times of zero, where the subject responded faster than the time to present the stimulus, and that had response times slower than 4.16 seconds (99^{th} percentile).

Subjects were presented with long and short vertical rectangular bars while signal uncertainty, automation accuracy, and decision type are manipulated in the context of a manufacturing quality assurance task. Subjects were instructed that short bars were desired and should be selected whereas a long bar should be rejected. We measured RT and accuracy over three decision type conditions for all subjects: the subject decided without automation, with automation (level 4 decision type), or chose to either agree or disagree with the automation's recommendation (level 5 decision type). There were two signal uncertainty conditions determined by the standard deviation of the bar lengths: easy (.15 SD) and difficult (.3 SD), and two automation accuracy conditions: high (95%) and low (80%). The automation accuracy conditions did not apply for the decision type condition of deciding without automation. The order of these ten conditions were counterbalanced across subjects.

Results

Table 1 summarizes the median RT and mean accuracies across all test conditions for all responses (correct and incorrect). For the unaided baseline and level 4 decision type, the hard conditions have longer RT than the easy conditions, for each level of aid accuracy. Additionally, the low aid accuracy conditions for the level 5 decision type have longer RT than all the other conditions. Figure 1 shows the RT distributions by test condition. The level 5 decision type

condition appears to have a bimodal distribution for the correct responses in the high accuracy aid condition. This did not occur for the level 4 decisions. By looking at the distribution of RT for individual subjects, and separately for agree and disagree responses (Figures 2a and 2b), it is evident that some subjects respond very quickly to agree with the automated aid, regardless of its recommendation, while other subjects respond with times similar to when directly selecting the signal (i.e. level 4 decision). In trials with the high accuracy aid this will result in correct responses 95% of the time, which creates two distributions of RT – one quicker for those that always quickly agree and one slower for those that evaluate the stimulus and then decide. The trend is not visually obvious for the low accuracy aid; further analysis is needed to determine if the trend occurs for the low accuracy aid as well.

Table 1. *Summary of Response Times and Accuracies Across Test Conditions*

Condition (# trials)	Median Human RT, ms (95% CI)	Mean Accuracy, % (95% CI)
1 = select signal w/o aid, easy (5251)	639 (610 – 669)	84.8 (81.7 – 87.7)
2 = select signal w/o aid, hard (5255)	652 (617 – 688)	75.8 (73.9 – 77.6)
3 = select signal w/high accuracy aid, easy (5266)	669 (641 – 698)	91.4 (89.7 – 93.1)
4 = select signal w/high accuracy aid, hard (5260)	693 (658 – 727)	86.5 (84.6 – 88.4)
5 = select signal w/low accuracy aid, easy (5253)	657 (621 – 694)	85.3 (83.1 – 87.4)
6 = select signal w/low accuracy aid, hard (5243)	692 (647 – 738)	78.7 (76.7 – 80.6)
7 = agree/disagree w/high accuracy aid, easy (5183)	657 (584 – 730)	93.0 (91.2 – 94.9)
8 = agree/disagree w/high accuracy aid, hard (5125)	657 (576 – 738)	89.3 (86.7 – 92.0)
9 = agree/disagree w/low accuracy aid, easy (5193)	811 (741 – 880)	82.7 (79.3 – 86.0)
10 = agree/disagree w/low accuracy aid, hard (5164)	834 (744 – 924)	77.5 (74.7 – 80.3)

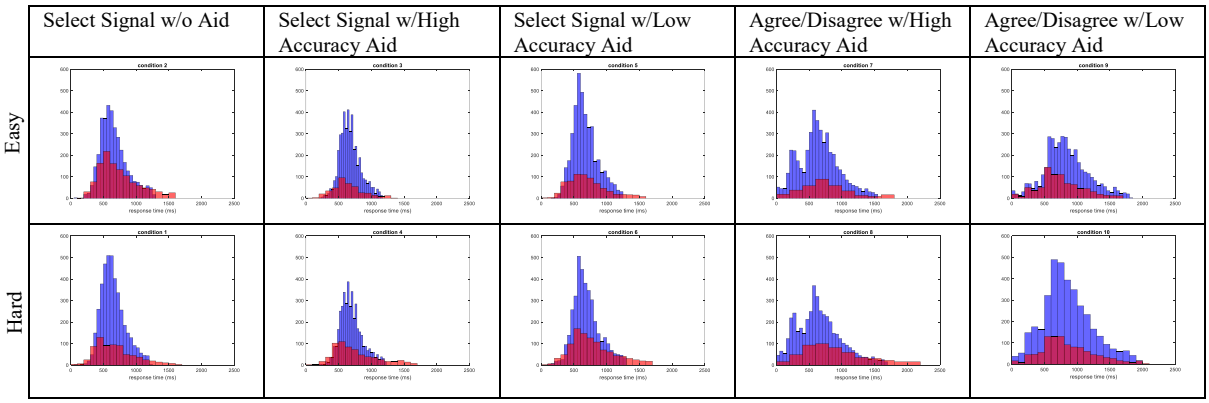

Figure 1. Distribution of correct (blue) and incorrect (red) response times (in milliseconds) by condition. The response times (x-axis) range from 0 to 2500 milliseconds for all plots. The frequency (y-axis) ranges from 0 to 600 for all plots.

A repeated measures ANOVA on median correct RTs in aided conditions indicated decision type, aid accuracy, and the interaction between decision type and aid accuracy were significant ($F(1,301) = 5.6$, $p = .02$, $\eta = .09$; $F(1,301) = 31.7$, $p < .001$, $\eta = .22$; $F(1,301) = 31.7$, $p < .001$, $\eta = .22$). A linear mixed-effects regression model on RT indicated decision type and the interaction between decision type and aid accuracy were significant predictors ($B = 1.26$, $t = 24.1$, $p < .001$; $B = -1.37$, $t = -23.1$, $p < .001$) of RT. Subjects were faster in the select

decision type and decision type as a moderator of aid accuracy on RT is stronger for agree/disagree decisions than for direct selection decisions. A logistic mixed-effects regression analysis of accuracy in aided conditions, indicated that decision type, aid accuracy, difficulty, and the interaction between decision type and aid accuracy were significant predictors (B = -0.13, z = -3.6, p < .001; B = 0.58, z = 14.4, p < .001; B = -0.43, z = 15.1, p < .001; B = 0.38, z = 6.4, p < .001) of accuracy. Subjects were more accurate in the select decision type, more accurate with more accurate aids, and more accurate with easier tasks.

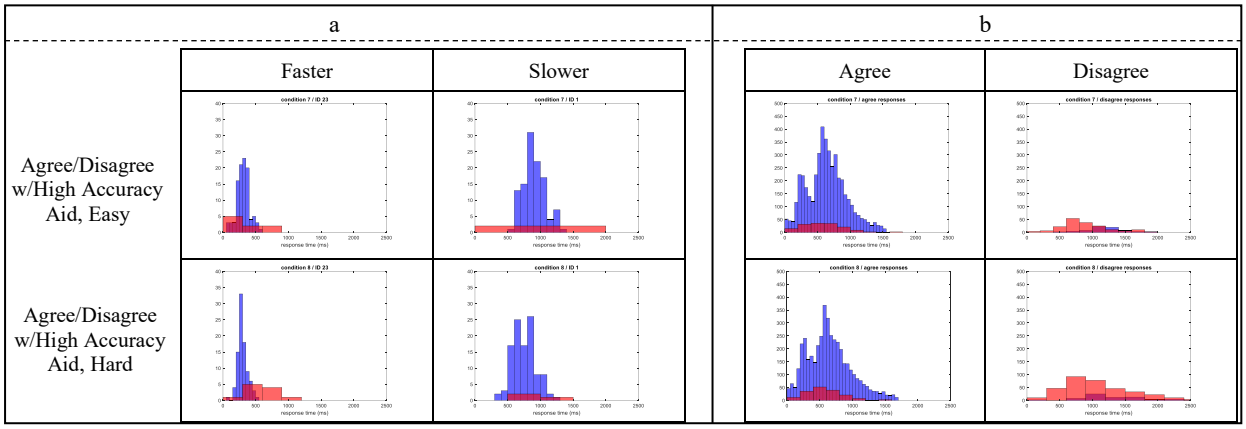

Figure 2. (a). Representative distributions of individual subjects with faster and slower correct (blue) and incorrect (red) response times (in milliseconds) for level 5 decision type in the high accuracy aid condition. The response times (x-axis) range from 0 to 2500 milliseconds for all plots. The frequency (y-axis) ranges from 0 to 40 for all plots.
(b) Distribution of correct (blue) and incorrect (red) response times (in milliseconds) for level 5 decision type in the high accuracy aid condition separated by agree and disagree decision. The response times (x-axis) range from 0 to 2500 milliseconds for all plots. The frequency (y-axis) ranges from 0 to 500 for all plots.

Modeling

The linear ballistic accumulator (LBA) model of decision making consists of five fitted parameters: A, the range of uniform distribution U[0,A] from which starting point k is drawn; b, the response boundary; v, the mean drift rate; s_v, the standard deviation of drift rate; and t_0, the non-decision time. The LBA model uses response times for both correct and incorrect responses and assumes a different drift rate for each (v for correct response and 1-v for incorrect), both of which are heading in parallel toward a common response boundary, b. Evidence accumulates linearly at the mean drift rate, v, for both responses until one reaches the response boundary; this is the model's response and once reached the evidence for the alternative response is discarded.

We compared the fit of four different LBA models using the data from this study: Model 1 fixes all parameters between conditions; Model 2 allows mean drift rate to vary between the 10 test conditions and fixes all other parameters; Model 3 allows response boundary to vary between test conditions and fixes all other parameters; and Model 4 allows response boundary and mean drift rate to vary and fixes all other parameters. The optimal set of fitted parameters for each model was found by maximizing the log-likelihood with a modified version of Steve Fleming's MATLAB code for fitting the LBA model (available at

https://github.com/smfleming/LBA). All participants were assumed to be equal for determining the model parameters.

Initially, each model with one parameter varied (i.e. Model 2 and Model 3) was compared to the most restricted, specific model (i.e. Model 1); each had a log-likelihood value that is larger (i.e. less negative) than the log-likelihood for Model 1, indicating a better goodness-of-fit. A χ^2 comparison of each general model to the specific model results in a χ^2 that exceeds the critical χ^2 (16.9) for 9 degrees of freedom given α=.05 for both models (χ^2_{1v2} = 2164; χ^2_{1v3} = 2768), indicating that the increased log-likelihood for the more complex models is merited by the additional parameters in each. Next the more complex, general model that allowed both *v* and *b* to vary across all test conditions (i.e. Model 4) was compared to Model 2 and Model 3 separately. The additional complexity in Model 4 provides a higher log-likelihood and that additional complexity results in a χ^2 that exceeds the critical χ^2 (28.9) for 18 degrees of freedom given α=.05 (χ^2_{2v4} = 2090; χ^2_{3v4} = 1487). Table 2 lists the optimal parameters for Model 4.

Table 2. *Optimal parameters for Model 4*

Constants	*Parameters that vary by condition*									
A = 488	*Unaided*		*Level 4 Decision*				*Level 5 Decision*			
t_0 = 1.02 x 10^{-3}	*Easy*	*Hard*	*Easy*	*Hard*	*Easy*	*Hard*	*Easy*	*Hard*	*Easy*	*Hard*
s_v = 0.225			*+ Accuracy*		*– Accuracy*		*+ Accuracy*		*– Accuracy*	
v	0.70	0.64	0.79	0.72	0.71	0.66	0.76	0.73	0.66	0.63
b	739	726	803	784	755	756	689	676	799	786

Note: "+ Accuracy" is the more accurate aid. "– Accuracy" is the less accurate aid.

Discussion

The results from this study show that the type of decision made by a user has an effect on their RT and accuracy and the effect interacts with the accuracy of the automated aid. Additionally, aid accuracy and task difficulty also have an effect on accuracy. Looking only at user accuracy, the results agree with Rovira (2014) for difficult (i.e. high demand) tasks where accuracy improved with reliable automation and did not degrade below unaided performance with less accurate automation. For the easy (i.e. low demand) tasks, we differ from their results in that we find response accuracy improved with more accurate automation and actually degraded below the unaided performance with less accurate automation for the agree/disagree condition. The tasks presented in this study were more abstract than the task in Rovira's study so it is possible that users rely on the aid more for more abstract tasks; this needs further research.

The variations in mean drift rate and response boundary in the LBA model describe varying behavior in subjects across the test conditions. We expected lower response boundaries for easier tasks and/or more accurate aids, but the results were the opposite. The mean drift rate (*v*) and response boundary (*b*) are higher for the easy conditions, across all aid accuracies and decision types, indicating that subjects are more efficient, but also more cautious for the easy conditions; the opposite is true for difficult tasks. For the more accurate aid, across all difficulties for level 4 decisions, the *v* and *b* are higher, again indicating that subjects are more efficient in their choice, but also more cautious. For level 4 decisions with the less accurate aid, the mean drift rates approximately equal the values for the baseline unaided condition, while the *b* are

higher, indicating that subjects are equally efficient in both conditions, but more cautious with a less accurate aid. The level 5 decision values are interesting because they behave oppositely; *v* increases and *b* decreases for the more accurate aid and *v* decreases and *b* increases for the less accurate aid. Subjects are more cautious and least efficient with the less accurate aid and least cautious and more efficient with the more accurate aid for agree/disagree decisions.

We found that when a lower accuracy aid is presented, users respond faster and more accurately making a direct selection, but when a higher accuracy aid (e.g. ≥ 92% for the aid used in this study) is available, users respond faster and more accurately by agreeing or disagreeing with the aid's recommendation. Users are least cautious and more efficient when agreeing or disagreeing with the more accurate aid. When using a less accurate aid users are less cautious and more efficient making a direct selection. This is the difference between level 4 and level 5 interaction in the model proposed by Parasuraman, Sheridan, & Wickens. When designing and implementing an automated aid the goal is generally to create one that has a high accuracy, but if that is not possible then there is a benefit to implementing the automation where it suggests one alternative, but the human still has authority to execute that alternative or choose a different one.

Acknowledgements

The authors appreciate the contributions of Jane Hwang in assisting with conducting the study. The tasks used in this study were based on previous studies by Zinn, Yamani, & McCarley.

References

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253. Retrieved from https://journals-sagepub-com.ezproxy.libraries.wright.edu/doi/

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

Rovira, E., Cross, A., Leitch, E., Bonaceto, C. (2014). Displaying Contextual Information Reduces the Costs of Imperfect Decision Automation in Rapid Retasking of ISR Assets. *Human Factors, 56*(6), 1036-1049. Retrieved from https://journals-sagepub-com.ezproxy.libraries.wright.edu/doi/full/

Rovira, E., McGarry, K., Parasuraman, R. (2007). Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task. *Human Factors, 49*(1), 76-87. Retrieved from https://journals-sagepub-com.ezproxy.libraries.wright.edu/doi/pdf/

Wickens, C.D., Clegg, B.A., Vieane, A.Z., & Sebok, A.L. (2015). Complacency and automation bias in the use of imperfect automation. *Human Factors, 57*(5), 728-739. Retrieved from https://journals-sagepub-com.ezproxy.libraries.wright.edu/doi/full/