



Effects of individual differences in working memory on performance and trust with various degrees of automation

Ericka Rovira, Richard Pak & Anne McLaughlin

To cite this article: Ericka Rovira, Richard Pak & Anne McLaughlin (2016): Effects of individual differences in working memory on performance and trust with various degrees of automation, Theoretical Issues in Ergonomics Science, DOI: [10.1080/1463922X.2016.1252806](https://doi.org/10.1080/1463922X.2016.1252806)

To link to this article: <http://dx.doi.org/10.1080/1463922X.2016.1252806>



Published online: 21 Nov 2016.



Submit your article to this journal [↗](#)



Article views: 38



View related articles [↗](#)



View Crossmark data [↗](#)



Effects of individual differences in working memory on performance and trust with various degrees of automation

Ericka Rovira ^a, Richard Pak ^b and Anne McLaughlin ^c

^aDepartment of Behavioral Sciences & Leadership, Behavioral Sciences & Leadership, US Military Academy, West Point, NY, United States; ^bDepartment of Psychology, Clemson University, Clemson, SC, United States; ^cDepartment of Psychology, North Carolina State University, Raleigh, NC, United States

ABSTRACT

Previous studies showed performance benefits with correct automation, but performance costs when the automation was incorrect (i.e. provided an incorrect course of action), particularly as degrees of automation increased. Automation researchers have examined individual differences, but have not investigated the relationship between working memory and performance with various degrees of automation that is both correct and incorrect. In the current study, working memory ability interacted with automation reliability and degree of automation. Higher degrees of correct automation helped performance while higher degrees of incorrect automation worsened performance, especially for those with lower working memory. Lower working memory was also associated with more trust in automation. Results illustrate the interaction between degree of automation and individual differences in working memory on performance with automation that is correct and automation that fails.

ARTICLE HISTORY

Received 17 December 2015
Accepted 21 October 2016

KEYWORDS

Human automation interaction; degrees of automation; individual differences; working memory; trust; task load; mental workload

Relevance to ergonomics theory

The current results confirm the important role of working memory ability in the use of automation: individuals with high working memory ability seem most able to perform the task and evaluate the automation by appropriately calibrating their trust, while those lower in working memory ability inappropriately calibrate their trust and rely on automation, even when it is incorrect.

Introduction

Commercial pilots are supported by sophisticated technology in the cockpit, soldiers are supported with automated targeting systems, drivers are supported by blind spot warnings, and consumers with modern mobile phones can get restaurant recommendations tailored to their location. In each of these examples of automation, a system is carrying out a task that once was carried out by the user, thus alleviating some work. For example, blind spot warning systems in vehicles constantly monitor areas that are difficult for the driver to view. Without this system, the driver would have to incur additional work to

maintain a high level of attentional vigilance but also would need to adjust their position to see the blind spot.

Forms of automation can be characterised along two dimensions: type and level. The *type of automation* is characterised by the stage of information processing supported: information acquisition, information analysis, decision-making, or action implementation. The first two stages are often combined and called ‘information automation’ while the latter two stages are often combined and referred to as ‘decision automation’. The *level of automation* denotes the allocation of the task, from a low level allocated to the automation (manual) to a highly autonomous level (Parasuraman, Sheridan, and Wickens 2000; Sheridan and Verplank 1978). Collectively, changes in types and levels of automation can be referred to as *degrees of automation* (Onnasch et al. 2014) with higher degrees of automation supporting later stages of information processing (e.g. decision-making rather than attention) and more of the task allocated to the automation.

A growing body of research has examined how human performance is affected by failures of highly reliable automation (Crocoll and Coury 1990; Endsley and Kaber 1999; Galster, Bolia, and Parasuraman 2002; Lorenz et al. 2002; Onnasch et al. 2014; Rovira, McGarry, and Parasuraman 2007; Sarter and Schroeder 2001; Wickens and Xu 2002; Wickens and Dixon 2005). The interest is motivated by the severe human performance consequences of highly reliable, yet *imperfect* automation that can cause out-of-the-loop unfamiliarity (Wickens 1992), automation complacency (Parasuraman, Molloy, and Singh 1993), loss of situation awareness (Endsley and Kiris 1995), passive monitors versus active controllers (Lee and Moray 1994) and skill degradation (Bainbridge 1983).

In a meta-analysis of 18 studies, Onnasch et al. (2014) found performance benefits for correct automation and performance decrements after an automation failure with higher degrees of automation. Of most interest were the decrements in performance found when automation support moved from information automation to decision automation. Thus, an important goal for designers is to mitigate performance costs associated with failures of higher degrees of automation by facilitating appropriate trust calibration (e.g. Rovira et al. 2014). One approach is to better understand the role of individual differences in cognitive ability on the appropriate use of automation in complex decision-making tasks.

Individual differences

Some early research explored sources of individual differences and performance with automation (e.g. Singh, Molloy, and Parasuraman 1993). However, these early investigations focused on what could be considered personality characteristics (e.g. complacency potential; Singh, Molloy, and Parasuraman 1993) and complacent behaviour due to less monitoring of the automated task (Parasuraman and Manzey 2010). Another source of individual differences in use of automation may be working memory (Baddeley 1986; Engle 2002). Working memory plays a key role in executive control processes which are thought to underlie effective decision-making and situation awareness (Endsley and Kiris 1995). Thus, individuals of higher working memory ability should be better able to generate, remember and evaluate consequences of different courses of action than individuals with low working memory ability. However, to date, examination of working memory’s influence on automation use has been indirect and has not included various degrees of automation. For example, Chen and Terrence (2009) investigated the effects of

automation failures and individual differences in perceived attentional control, a component of working memory (Shipstead et al. 2014), in a military multitask environment. Perceived attentional control was assessed using a subjective self-assessment of individuals' attentional focus and shifting. They found that those with high perceived attentional control were more negatively affected by false alarms, while individuals with low perceived attentional control suffered more with miss-prone automation. In the context of their task (military gunner and robotics operator), perceived attentional control was an important moderator of how operators reacted to automation false alarms and misses.

More direct evidence of the importance of individual differences in working memory comes from a study by de Visser et al. (2010). They investigated the role of working memory in an automated unmanned aerial vehicle task by varying task load (low, high) and automation reliability (manual, 100% reliable automation, and 20% reliable automation). Participants completed the Operation Span working memory test (Engle 2002). Working memory scores significantly correlated with performance on the automated task. For each automation task performance measure, linear models that included working memory accounted for more of the variance in performance as compared to the linear models without the working memory measure. Thus, when individual differences in working memory were accounted for, more variation in performance with automation was explained. Critically, however, this study did not manipulate the *degrees* of automation. Also, as the analysis combined both automation correct and automation failure trials for prediction in linear modelling, making it unclear if working memory influenced performance on only trials where the automation was correct or on trials where the automation failed.

Hypotheses

The purpose of the current research was to address two significant gaps in the literature on individual differences and automation: indirect measurement of working memory and the unknown role of degrees of automation on operator performance. To examine the role of individual differences in working memory on performance, we varied degrees of automation and task load and measured individual differences in spatial working memory ability (henceforth referred as working memory). This is contrasted with previous studies that used self-reported proxies for working memory (Chen and Terrence 2009), or indirect genetic markers of cognitive performance (Parasuraman et al. 2012), or have not examined the role of degrees of automation (de Visser et al. 2010). We manipulated task load because evidence from a review of 20 automation reliability studies suggested that dependence on imperfect automation would be stronger under higher task load (because the operator's limited resources are expended; Wickens and Dixon 2007).

We hypothesised that individual differences in working memory would differentially impact performance with various degrees of automation:

- (1) Consistent with previous literature, we hypothesised that:
 - a. decision accuracy would be better with correct automation compared to manual control (a manipulation check)
 - b. there would be no effect of task load on decision accuracy when the automation was correct.

- c. the differential impact of information versus decision automation would be evident with automation failures, especially when task load was high. While participants would not know if a given trial would be a correct trial or an automation failure trial, if participants are relying on the automation, they should make an appropriate decision when the automation was correct, but accuracy should degrade with an automation failure.
- (2) As suggested by Parasuraman et al. (2012), we expected individuals with higher working memory ability to show less of a performance decrement when the automation failed compared to individuals with lower working memory ability. This key relationship is expected because working memory is associated with reduced attentional control (Unsworth and Engle 2007), which may inhibit the ability of individuals with lower spans to monitor for automation failures (Sarter, Mumaw, and Wickens 2007). In addition, when faced with the prospect of carrying out the task manually, low-span individuals' reduced working memory capacity makes it more difficult to update the contents of working memory, and thus reduces their time available to consider and correct incorrect automation. With automation failures, high task load, and increasing degrees of automation, it was predicted that the benefits of better spatial working memory ability would be highlighted.
- (3) We expected a relationship between variations in cognitive ability and self-report measures of trust (Lee and Moray 1994; Jian, Bisantz, and Drury 2000). Individuals with lower working memory ability would trust the automation more compared to individuals with higher working memory ability because individuals with lower working memory ability would need to rely on the automation more than those with higher ability.

Methods

Participants

A total of 86 cadets (18 women) from the U.S. Military Academy participated in this study for extra credit. Ages ranged from 18 to 24 years ($M = 20.27$, $SD = 1.25$). One participant was excluded due to equipment failure, and hence subsequent analysis was of 85 participants.

Stimuli and task procedures

Participants completed this study in two hours including training and breaks. Participants first completed a spatial working memory task followed by a simulated artillery sensor-to-shooter targeting task. Response time and accuracy were collected for all measures though the dependent measure of interest for the sensor-to-shooter task was decision-making accuracy, as response time was restricted for this task.

Working memory measure

A spatial working memory task assessed working memory ability (Figure 1; Greenwood et al. 2005). A fixation cross appeared for 500 ms followed by one, two, or three black dots (1.65° in diameter, each indicating a target location) at random screen locations for

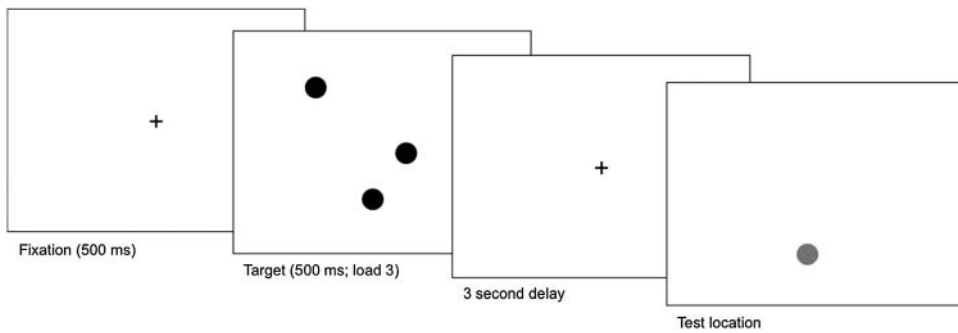


Figure 1. Working memory measure used to indicate spatial working memory capacity at three levels of load.

500 ms. Simultaneously with dot offset, the fixation cross reappeared for 3 s. At the end of the delay, a single red test dot appeared on the screen. This test dot appeared either at the same location as one of the target dots (match condition) or at a different location (non-match condition). On non-match trials, the distance between the correct location and the test dot varied randomly over three levels ($\sim 1.3^\circ$, 2° , or 2.6° of visual angle). Participants indicated whether the test dot location matched one of the target dots using their index fingers to select one of two responses on a keyboard.

A composite working memory score was created consisting of accuracy on trials (collapsed across distances) at three levels of memory load, and in both match and non-match conditions. Z-scores were computed for each of the six conditions (three levels of load, match/non-match) and a mean was taken to form a composite for each individual. Thus, this composite score was not standardised, but reflected the average of the standardised scores.

Artillery sensor-to-shooter targeting task

A low-fidelity software simulation of an artillery sensor-to-shooter targeting system was used with various degrees of automation (Rovira, McGarry, and Parasuraman 2007). The artillery task consisted of three components in separate windows: a terrain view, a task window, and a communications module (Figure 2). A two-dimensional terrain view of a simulated battlefield displayed red enemy units (labelled E1, E2, ... E_x), yellow friendly battalion units (B1, B2, and B3), green friendly artillery units (A1, A2, ... A_x), and one orange friendly headquarter unit (HQ). Participants were required to identify the most dangerous enemy target and to select a corresponding friendly unit to engage in combat with the target. The criteria for enemy unit engagement selection (derived by consulting with military subject matter experts) was based not only on the closest distance between it and friendly units but also the relative distance to the HQ unit, with a red unit that was closer to the HQ than another red unit classified as more dangerous and requiring engagement. Specifically, the following criteria had to be met: (1) only artillery units could engage enemy units in combat; (2) the friendly unit closest in distance to an enemy unit was to be given the highest priority for combat engagement; (3) if two friendly units were equally distant from an enemy unit, or if a friendly unit could engage in combat with two

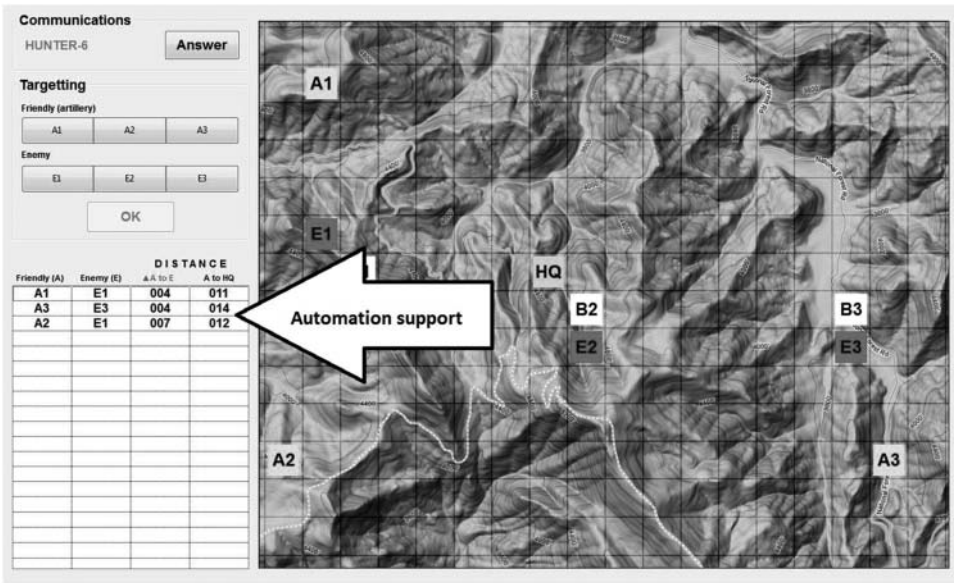


Figure 2. The sensor-to-shooter task interface. The example shows medium-decision automation condition with low task load. Note that with this degree of automation, the target pairings in the automation are ordered in ascending distance by enemy-friendly (third column) and friendly HQ (fourth column).

enemy units that were both an equal distance away from the friendly unit, then it was important to select the unit closest to headquarters.

The bottom left of the task window provided the different degrees of automation support. This window was absent in the manual, unaided conditions. The lowest degree of automation support was *information automation*, which provided an exhaustive list of all possible engagement combinations ordered alphabetically, including the distances between enemy-friendly units, and headquarters. Because no explicit suggestion for decision-selection was provided, this corresponded to information automation in the Parasuraman et al. (2000) taxonomy. This is contrasted with the manual, unaided condition which left the participants to calculate distances on their own. The next higher degree of automation, *low-decision automation*, gave a list of all possible engagement combinations as before but the listings were prioritised by distance with the best selection first and the worst choice last, making this a form of decision automation. The automation determined priority by enemy-friendly distances with the friendly unit closest to HQ getting a higher priority. In the *medium-decision automation* condition, the highest degree of automation in this study, the participant was provided only the top three options for engagement as ordered by distance, including the distances between all enemy targets, friendly units, and headquarters (Figure 2). Our distinction between low- and medium-decision automation is informed by Sheridan and Verplank's (1978) distinction between 'level 2', where the automation shows a complete list of decision alternatives, and 'level 3', where the automation narrows the selection of alternatives to a few.

Participants could either use the assistance of the automation or make their own enemy-friendly unit engagement decisions, but were required to respond within 10 s.

Participants were able to verify the automation by reviewing the terrain view and manually computing distances themselves by counting grid boxes. After they made their selection, or if 10 s had elapsed, the trial ended and the terrain map was replaced with a new grid of enemy, friendly, and HQ units.

To increase the difficulty of completing the sensor-to-shooter task, a call for communications (call sign) appeared every 6 s and remained displayed until the next call sign. Participants were required to click on the ANSWER button every time their personal call sign appeared while they were selecting units. The personal call sign appeared randomly around every 40–50 s into the experiment. This secondary task was always performed; there was no single task condition.

Workload and trust measures

To examine perceived workload of the task, participants completed a computerised version of NASA-Task Load Index (TLX) after each block (Hart and Staveland 1988). Participants rated their trust in automation after each automation-present block using an on-screen visual analogue scale ranging from 0 to 100 (adapted from Lee and Moray 1994) and completed a longer trust questionnaire at the end of the study (adapted from Jian, Bisantz, and Drury 2000). Example questions at the end of each block included: To what extent did you *rely* (i.e. actually use) the automation aid in this scenario? To what extent do you think the automation *improved* your performance in this scenario compared to performance without the automation?

Experimental design

The experiment was a 4 (Degree of automation support: manual, information automation, low-decision automation, and medium-decision automation) \times 2 (Task load: low, high) within-subjects design. *Task load* was manipulated by increasing the number of friendly and enemy units from three to six. As Wickens and Dixon (2005) found that automation reliability below a certain point (70%) was worse than no automation at all, the overall reliability of our automation was set at 80%.

Each of the eight conditions, 4 (Degree of automation support: manual, information automation, low-decision automation, and medium-decision automation) \times 2 (Task load: low, high), constituted a block of trials during the experimental task. Participants were informed that although the automation was highly reliable, it was not 100% reliable, but no further information on reliability was given. During practice, participants completed eight correct trials at both task load levels for each degree of automation support before a new degree of automation support was introduced. The order of the degrees of automation support was counterbalanced between participants (participants either started with manual, information, low-decision, or medium-decision conditions). Within each of the four conditions, task load was counterbalanced so that participants either started with low or high task load. Each of the eight blocks contained 40 trials. Trials were created by an algorithm that placed pieces randomly on a generic topographic map with the only constraint that battalion units (yellow) were always adjacent to enemy units (red). This was based on consultation with subject matter experts to create layouts that were plausible and realistic. Otherwise, enemy, friendly, and HQ positions were randomly placed.

Of the 40 trials per condition when automation was present (information, low-decision, and medium-decision conditions), 32 trials presented correct automation and 8 were incorrect automation. The automation incorrect trials were randomly dispersed throughout the block with the constraint that the first automation failure in any block did not occur until past the eighth trial to allow participants to build trust and not immediately discount the automation (Wickens and Xu 2002). Each participant completed a total of 320 test trials (8 blocks of 40 trials each).

The dependent variable was accuracy of enemy-friendly engagement selections (hereafter referred to as decision accuracy). Decision accuracy was calculated as the percentage of trials in which the participant correctly selected the optimal enemy-friendly pairing. Other measures included NASA-TLX (mental workload), trust after every block (for blocks with automation support), and trust at the end of the study (measured once).

Results

Repeated measures analyses of variance (ANOVAs) were conducted to evaluate the effects of degrees of automation, task load, and automation correctness (correct/incorrect) on decision accuracy, subjective mental workload, and trust. Multilevel linear models (MLMs) were conducted to measure the role of individual differences in cognitive ability on task performance under the various manipulations.

Manual control versus automation

Collapsing across the three degrees of automation (information, low-decision, and medium-decision), we compared decision accuracy without automation (manual) to trials with correct automation and trials with automation failures. A 3 (Automation correctness: manual, correct automation, automation failure) \times 2 (Task load: low, high) repeated measures ANOVA revealed a significant interaction between automation correctness and task load, showing that task load had different effects on decision accuracy depending on automation correctness conditions, $F(1,84) = 51.9, p < .05, \eta_p^2 = .38$ (Figure 3). Follow-up pairwise comparisons, Sidak-adjusted for multiple comparisons, showed the source of the interaction was that higher task decreased decision accuracy in both the manual condition (low task load $M = .75, SD = .14$; high task load $M = .61, SD = .16; p < .05$) and the correct automation condition (low task load $M = .88, SD = .07$; high task load $M = .79, SD = .10; p < .05$), but not the automation failure condition ($p > .05$). The lack of difference between load conditions might have been due to a floor effect when automation failed (i.e. performance was so low when automation was incorrect that task load had no further effect).

Multilevel models

The manual control condition was not included in the next series of analyses because we examined the effects of automation failures in each degree of automation, and there could be no automation failure in the manual condition.¹ We next used multilevel modelling to examine the influence of individual differences in working memory and decision accuracy. MLMs are regression-based and match some of nomenclature of regression, namely

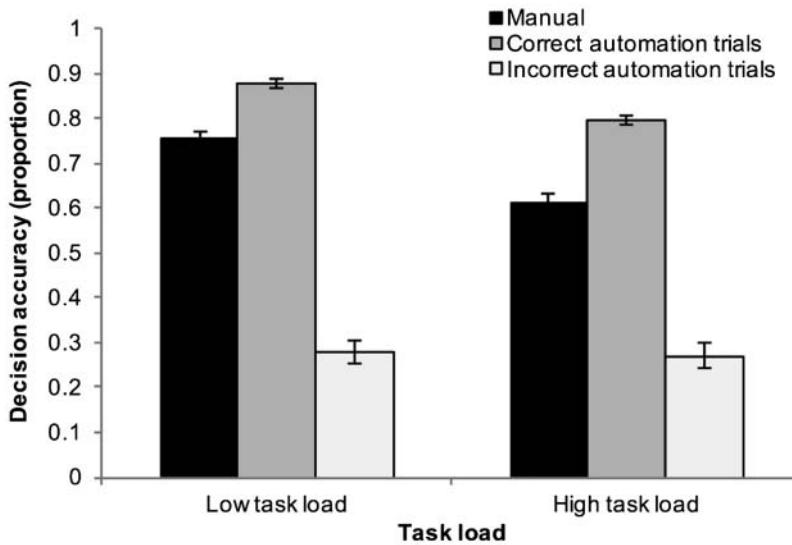


Figure 3. Decision accuracy as a function of task load and automation correctness. Bars indicate standard error.

fixed effects. However, MLMs can also produce random effects. For example, a regression predicts behaviour under the conditions in the study, a fixed effect. A random effects model extends that prediction beyond the conditions, for example, predicting performance under very high automation when the study only included a high-automation condition. MLMs are a preferred form of analysis for nested data structures, where multiple observations are collected from each participant, as they allow the simultaneous estimation of intra-individual and inter-individual differences and compute residuals at each level in the hierarchy (Raudenbush and Bryk 2002). Finally, similar to regression, MLM allows for the inclusion of dichotomous and continuous predictors; however, regression treats nested data as independent observations and is more likely to produce Type I error (Hox and Bechger 1998; Tabachnick and Fidell 2007). Multilevel models recognise the nested structure of the data and do not underestimate the standard errors of the regression coefficients, thus reducing the chance of Type I error (Raudenbush and Bryk 2002). Hoffman and Rovine (2007) provided an accessible discussion of the usefulness of MLMs in human factors research that support the choice of this analysis for the current data. Multilevel modelling was implemented using PROC MIXED through SAS, version 9.4.

A two-level hierarchical model assessed the effects of the within-person variables of degrees of automation, task load, automation correctness, the between-person predictor of working memory score, and their cross-level interactions on decision accuracy in the sensor-to-shooter task. Multiple responses were nested within the 85 participants as each participant performed the sensor-to-shooter task under various degrees of automation support and task load on trials where the automation was either correct or failed, meaning that the accuracy of their responses was nested within those variables. In turn, the within-person manipulations were nested within the attributes of the participant (i.e. their working memory ability). We used a model-building approach where we first ensured there was significant variability at both levels to allow predictors to be entered at those levels (Model 1),

Table 1. Unstandardised coefficients of multilevel models of the within- and between-person effects of predictors on accuracy in a sensor-to-shooter task.

	Model 1		Model 2		Model 3	
	Unconditional model		Random coefficients regression		Slopes and intercepts	
	Estimate	SE	Estimate	SE	Estimate	SE
<i>Fixed effects</i>						
Intercept	0.554***	0.014	0.346***	0.038	0.346***	0.038
<i>Between-person</i>						
Working memory composite score (WM)					0.114	0.060
<i>Within-person</i>						
Automation Support (AutoSupp)			-0.029	0.017	-0.032	0.017
Task load			0.146**	0.052	0.149**	0.051
Reliability			0.276***	0.051	0.274***	0.051
Task load × AutoSupp			-0.081***	0.024	-0.080***	0.024
Task load × Reliability			-0.501***	0.073	-0.502***	0.072
AutoSupp × Reliability			0.154***	0.024	0.159***	0.024
Task load × AutoSupp × Reliability			0.210***	0.034	0.209***	0.033
<i>Cross-level</i>						
Task load × WM					-0.036	0.065
AutoSupp × WM					0.011	0.026
Reliability × WM					0.080	0.065
Task load × AutoSupp × WM					0.013	0.030
Reliability × AutoSupp × WM					-0.089**	0.030
<i>Random effects</i>						
σ^2	0.149	0.007	0.049	0.002	0.047	0.002
τ_{00}	0.005	0.003	0.013	0.003	0.011	0.002
<i>Model fit statistic</i>						
AIC	972.2		-12.6		-28.6	

Working memory composite score was grand-mean centred. SE indicates standard error.

** $p < .01$, *** $p < .001$.

then added the within-participant predictors manipulated in the task (Model 2), and finally added the between-participant predictor of working memory score (Model 3). Each model was compared to the last using Akaike's information criterion (AIC) values to ascertain if the added predictors increased the quality of the model.

Model 1: No predictors. The first step was to run a fully unconditional model, one without any predictors (Table 1: Model 1), to discover the amount of within- and between-person variance in accuracy. This unconditional model served two purposes: first, to determine if there was significant variance at both the within-person (σ^2) and between-person (τ_{00}) levels in accuracy and, second, to provide a baseline to assess the fit of subsequent multivariate multi-level models (Models 2 and 3). The unconditional model revealed significant variance at both levels, with 97% of the variance at the within-person level ($\sigma^2 = 0.149$, $z = 21.48$, $p < .001$) and 3% of the variance at the between-person level ($\tau_{00} = 0.005$, $z = 1.81$, $p = .034$). Thus, subsequent models were run to explain the variance in accuracy using the within-person predictors of task load, automation success/failure, degree of automation support, and the between-person predictor of working memory ability.

Model 2: Within-person variables. Model 2 examined the effects of the within-person manipulations on accuracy (Table 1) and found that task load, automation correctness, and degree of automation accounted for 67% of the 97% within-subject variance. Model fit using the AIC improved from 976.2 to -12.6 (lower values indicate better fit).

There was a three-way interaction of degrees of automation, task load, and automation correctness, $F(1,915) = 39.00$, $p < .0001$. A simple-effects analysis showed that on *correct*

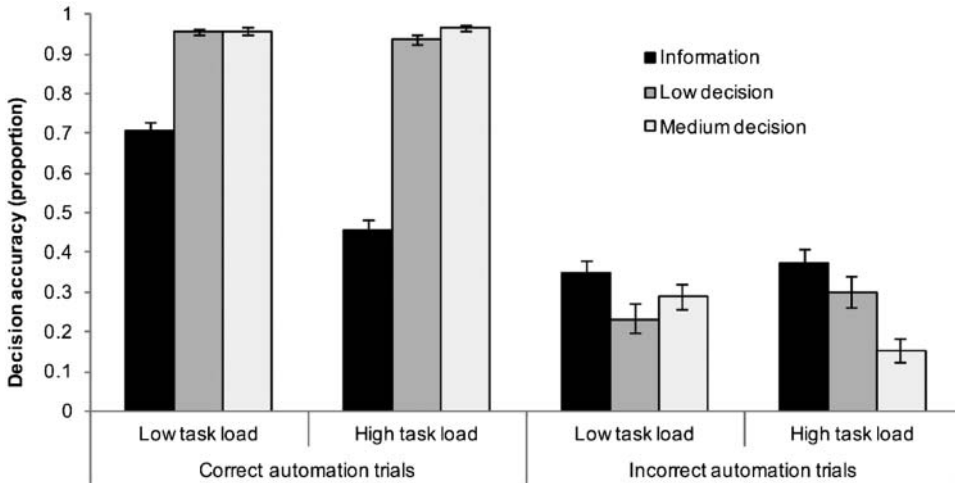


Figure 4. Decision accuracy as a function of automation correctness, task load, and degree of automation. Bars indicate standard error.

automation trials, a higher task load decreased accuracy only under information automation ($p < .05$). This can be seen on the left panel of Figure 4 where accuracy in the information automation condition declined as task load increased, while low and medium automation accuracy were unaffected. For trials with *automation failures*, pairwise comparisons showed that increasing task load significantly decreased accuracy only with medium-decision automation ($p < .05$, Figure 4, right panel). In sum, the interaction can be summarised as increased degrees of automation helped to buffer the degrading effects of task load when automation was correct, but amplified the negative effects of task load when it was incorrect. Note that this lower accuracy at high load on automation failure trials was due to the medium decision support condition scoring poorly, a distinction that could not be made in the earlier repeated-measures ANOVA analysis where automation conditions were combined. Said another way, increased task load hurt the lowest degrees of automation when it was correct but particularly hurt the highest degrees of automation when in error.

Model 3: Cross-level interactions. We expected individuals with higher working memory ability to show less of a decrement with higher degrees of automation when the automation failed as compared to individuals with less working memory ability. With automation failures or high task load, it was predicted that the benefits of having better working memory ability would be amplified (equation available in the Appendix). A third model was conducted to include working memory ability to examine these hypothesised cross-level interactions; this added slope as a random effect to the model.

As in Model 2, Model 3 revealed a three-way cross-level interaction of automation correctness, degrees of automation, and working memory ability (Table 1; $F(1,894) = 8.11$, $p = .005$). Model fit using AIC improved from -12.6 to -28.6 , indicating the benefit of considering individual differences in working memory on accuracy with automation (Figure 5). Task load was controlled for in this model and included as a factor in hypothesised interactions. When automation was correct, a benefit for higher working memory occurred solely with information automation. Simple-effects analyses showed that when

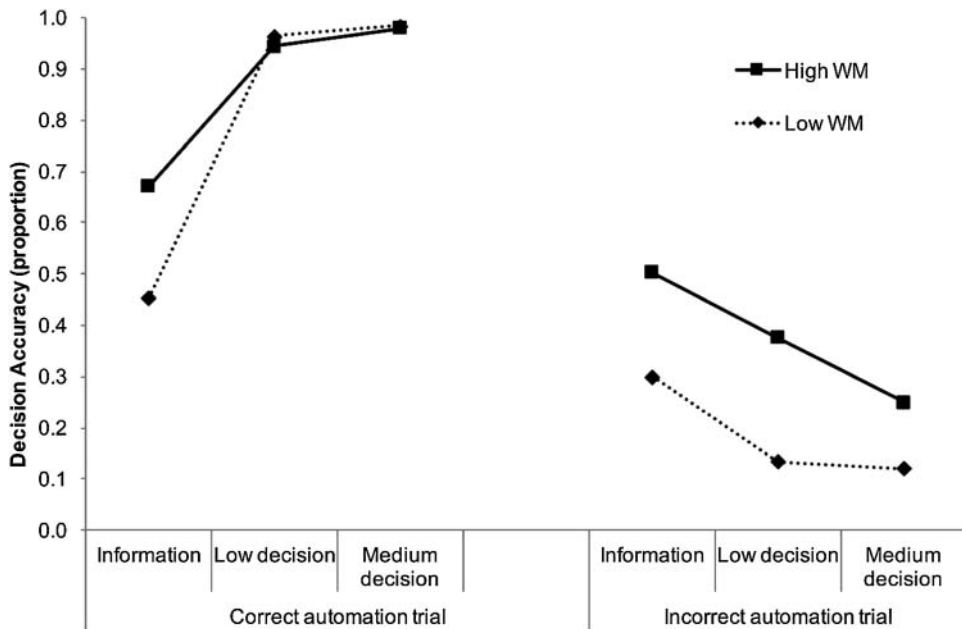


Figure 5. Decision accuracy as a function of automation correctness and degree of automation. Low WM was 1 SD below the mean and high WM was 1 SD above the mean.

the automation was incorrect, across all degrees of automation, higher working memory participants outperformed lower working memory participants $t(1,410) = -8.01, p < .001$ and $t(1,410) = -12.87, p < .001$.

Trust

Trust after every block. Figure 6 shows participants' subjective ratings of trust and self-confidence (at the end of every automation-present block (six measurements)). The interaction of degrees of automation and task load was significant, Wilks' lambda = .64, $F(8,66) = 4.56, p < .05, \eta_p^2 = .36$. Follow-up pairwise tests showed that the source of the interaction was a significant decrease in self-reported reliance (question 2) and decrease in the belief that automation improved performance (question 4) when task load increased but only in the *information automation* and *low-decision automation* conditions.

Trust at the end of the study. Trust was additionally measured once at the end of the study using a questionnaire adapted from Jian, Bisantz, and Drury (2000). Through this trust measure, we produced two values: positive and negative perceptions of automation. Lower working memory scores were associated with more agreement with positive perceptions of automation, $r = -.22, p < .05$, while higher working memory scores were associated with more agreement with negative perceptions of automation; $r = .24, p < .05$. These correlational findings support the role of working memory ability in use of automation: those able to perform the task and evaluate the automation appropriately calibrate their trust; while those lower in working memory inappropriately calibrate their trust in automation and rely on it, even when it fails.

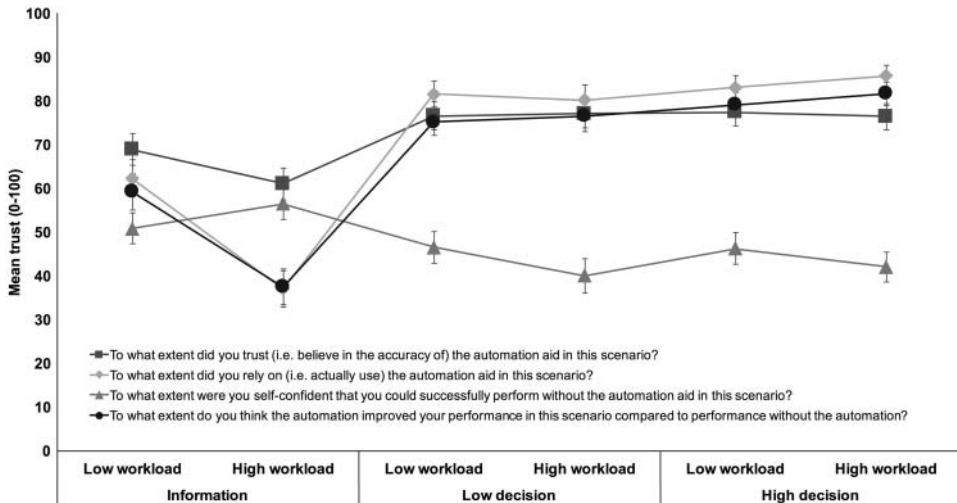


Figure 6. Trust was assessed at the end of every block. Bars represent standard error.

Subjective ratings of mental workload

Lower degrees of automation resulted in higher mental workload and increased mental workload at high task load, but there were no differences with the highest degree of automation (Figure 7). A 4 (Degrees of automation: manual, information, low-decision, medium-decision) \times 2 (Task load: low, high) repeated measures ANOVA revealed the main effects of degrees of automation, $F(3,219) = 61.7, p < .05, \eta_p^2 = .46$, task load, $F(1,73) = 44.1, p < .05, \eta_p^2 = .38$, and the interaction between degrees of automation and

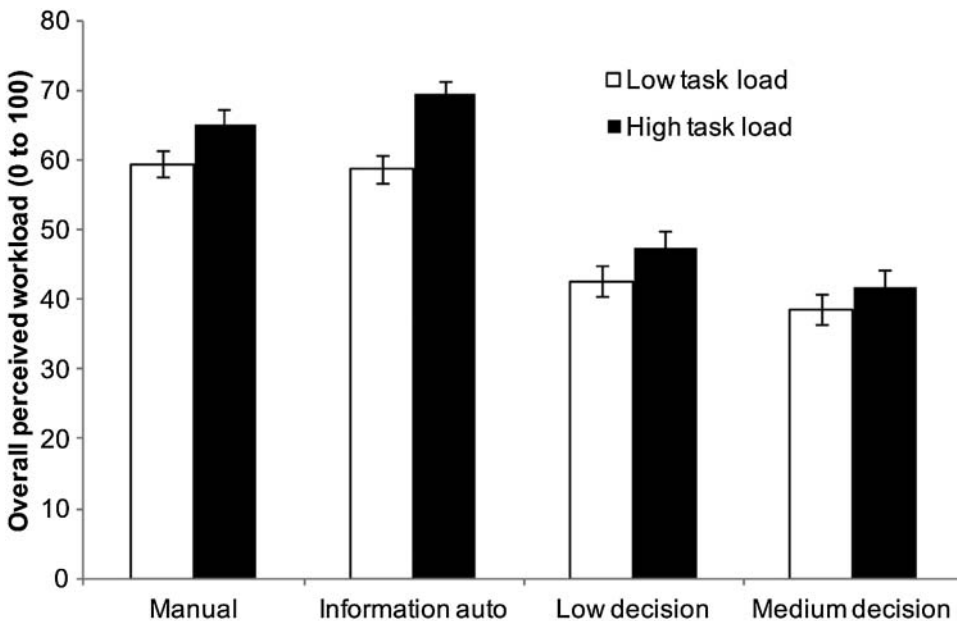


Figure 7. Perceived workload as a function of degree of automation and task load.

task load, $F(3,219) = 6.7, p < .05, \eta_p^2 = .08$. Pairwise comparisons showed that the source of the interaction was an effect of task load on perceived workload (higher task load resulted in higher perceived workload) for manual, $F(1,73) = 25.7, p < .05, \eta_p^2 = .26$, information automation, $F(1,73) = 33.7, p < .05, \eta_p^2 = .32$, and low-decision automation, $F(1,73) = 4.3, p < .05, \eta_p^2 = .06$, but not with medium-decision automation.

Discussion

Using a simulated automated targeting task, we showed that the extent to which an operator experienced both the costs of automation failures and the benefits of correct automation depended on individual differences in working memory. Our findings that working memory ability is related to trust in automation suggest more work should consider this individual difference.

First, our study verified that operators would perform better with correct automation compared to manual control. Second, while task load did not differentiate performance when the automation was correct for low- and medium-decision automation, we did see degraded performance with information automation and high task load compared to information automation and low task load (Hypothesis 1b). Finally, our study showed that with automation failures, there was no difference in accuracy with information automation and low-decision automation between low and high task load but accuracy declined at high task load with medium automation (Hypothesis 1c). These results demonstrate an interesting difference between lower degrees of automation (information and low-decision) and higher degrees of automation (medium-decision). The distinction between lower and higher degrees of automation stems from what is being automated (automation that moves to later stages within the information processing model) and how much is being automated (levels) (Onnasch et al. 2014).

It appears that lower degrees of automation can mitigate some of the performance penalty of increased task load when automation is incorrect, while performance significantly declines with automation failures and higher degrees of automation. Lower decision accuracy with increased task load may occur because the further along the information-processing continuum that automation supports the operator (e.g. cognitive versus perceptual), the more detrimental automation failures are because operators will not have generated their own courses of action (Wickens and Xu 2002).

A critical hypothesis regarded the role of individual differences in working memory and automation performance (Hypothesis 2). The MLM showed cross-level interaction between working memory, automation support and automation correctness. Performance was generally positively affected by increasing, correct degrees of automation but especially for those with lower working memory. Working memory did not differentiate accuracy with correct automation support above that was above information automation. Low- and medium-decision automation may have reduced the working memory demands of the task. Correct and increased automation support was especially beneficial for those with lower working memory (with maximal differences by working memory for information automation).

These results extend the literature by showing that there are individual differences in the degree to which automation benefits and hurts performance. When automation failed, accuracy declined as the degree of automation increased and those with lower working

memory were more severely impacted by automation failures than those with higher working memory. The source of this low working memory penalty is likely to be related to the new and higher task requirements with incorrect, high-degree automation. When presented with incorrect automation, and a realisation that it is incorrect, low-span individuals must now deploy critically constrained resources to suppress either competing or intrusive responses (Unsworth and Engle 2007; Conway, Cowan, and Bunting 2001), and then manually calculate a correct course of action. Our correlation (presented in footnote) confirmed that manual calculation in this task to be working memory-intensive. In a neuro-genetic study using the same task as used in this study, Parasuraman et al. (2012) similarly theorised that the link between performance in this task and working memory abilities came specifically from the need to, when automation was incorrect, update the contents of working memory with new information.

When the degree of automation is low and correct, those with higher working memory outperformed those with lower working memory. Taken together, these results supported Hypothesis 2 regarding the effects of degree of automation and working memory. Our results are the first empirical confirmation of the link between automation performance and individual differences in working memory as suggested by previous researchers (de Visser et al. 2010; Parasuraman et al. 2012), but also extend the literature by further specifying the automation conditions (degree of automation and automation correctness) under which working memory affects performance.

Finally, Hypothesis 3 predicting a relationship between working memory and trust in automation was supported. Working memory was significantly negatively correlated with measures of trust: individuals with higher working memory ability had fewer positive perceptions of automation and more negative ones.

Conclusion

Knowing how operators will perform with highly reliable, but imperfect degrees of automation at different task loads is enhanced if we understand the impacts of individual differences in working memory on human automation interaction. Our results add detail to the conventional wisdom that higher degrees of automation help performance and automation failures at higher degrees of automation harm performance: working memory is a crucial differentiator of how people behave with automation. This knowledge may be useful in the design of automated systems that alter the degree of automation based on an awareness of operators' working memory ability.

Based on our findings, one concrete design suggestion for automation that accommodates individual differences in working memory is to provide a mechanism that allows the user to select the degree of automation desired or to have the designer flexibly adjust the degree of automation based on a user's working memory ability. For example, based on the relationship between trust and working memory, users with higher working memory may prefer less automation than users with lower working memory. Alternatively, when automation is not perfect and the user is under high task load, designers of decision support tools may need to provide the option of lower degrees of automation for individuals with low working memory ability and higher degrees of automation support for individuals with higher working memory ability.

Note

1. The correlation between decision accuracy and working memory in the manual condition was significantly positive, $r = 0.23$, $p < .05$, reflecting that the unaided task was moderately working memory-intensive.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Air Force Office of Scientific Research [FA9550-12-1-0385].


Notes on contributors


Ericka Rovira is an associate professor in the Department of Behavioral Sciences & Leadership at the U.S. Military Academy, West Point. She received her PhD in applied experimental psychology from the Catholic University of America in 2006.


Richard Pak is an associate professor in the Department of Psychology at Clemson University. He received his PhD in psychology from the Georgia Institute of Technology in 2005.

Anne McLaughlin is an associate professor in the Department of Psychology at North Carolina State University. She received her PhD in psychology from the Georgia Institute of Technology in 2007.

ORCID

Ericka Rovira  <http://orcid.org/0000-0002-4820-5828>

Richard Pak  <http://orcid.org/0000-0001-9145-6991>

Anne McLaughlin  <http://orcid.org/0000-0002-1744-085X>

References

- Baddeley, A. 1986. *Working Memory*. New York: Oxford University Press.
- Bainbridge, L. 1983. "Ironies of Automation." *Automatica* 19: 775–779.
- Chen, J.Y.C., and P.I. Terrence. 2009. "Effect of Imperfect Automation and Individual Differences Concurrent Performance of Military Robotics Tasks in a Simulated Multitasking Environment." *Ergonomics* 52: 907–920.
- Conway, A.R., N. Cowan, and M.F. Bunting. 2001. "The Cocktail Party Phenomenon Revisited: The Importance of Working Memory Capacity." *Psychonomic Bulletin & Review* 8 (2): 331–335.
- Crocoll, W.M., and B.G. Coury. 1990. "Status or Recommendation: Selecting the Type of Information for Decision Aiding." In *Proceedings of the Human Factors Society 34th Annual Meeting*, 1524–1528. Santa Monica, CA: Human Factors and Ergonomics Society.
- de Visser, E., T. Shaw, A. Mohamed-Ameen, and R. Parasuraman. 2010. "Modeling Human-Automation Team Performance in Networked Systems: Individual Differences in Working Memory Count." In *Proceedings of the Human Factors and Ergonomics Society 54th Annual Meeting*, 1087–1091. Santa Monica, CA: Human Factors and Ergonomics Society.

- Endsley, M.R., and D.B. Kaber. 1999. "Level of Automation Effects on Performance, Situation Awareness and Workload in a Dynamic Control Task." *Ergonomics* 42: 462–492.
- Endsley, M.R., and E.O. Kiris. 1995. "The Out-of-the-Loop Performance Problem and Level of Control in Automation." *Human Factors* 37: 387–394.
- Engle, R.W. 2002. "Working Memory as Executive Attention." *Current Directions in Psychological Science* 11: 19–23.
- Galster, S.M., R.S. Bolia, and R. Parasuraman. 2002. "Effects of Information and Decision-Aiding Cueing on Action Implementation in a Visual Search Task." In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*, 438–442. Santa Monica, CA: Human Factors and Ergonomics Society.
- Greenwood, P.M., C. Lambert, T. Sunderland, and R. Parasuraman. 2005. "Effects of Apolipoprotein E Genotype on Spatial Attention, Working Memory, and Their Interaction in Healthy, Middle-Aged Adults: Results from the National Institute of Mental Health's BIOCARD Study." *Neuropsychology* 19 (2): 199–211. doi:10.1037/0894-4105.19.2.199
- Hart, S.G., and L.E. Staveland. 1988. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." In *Human Mental Workload*, edited by P.A. Hancock and N. Meshkati, 139–183. Amsterdam: Elsevier Science.
- Hoffman, L., and M.J. Rovine. 2007. "Multilevel Models for the Experimental Psychologist: Foundations and Illustrative Examples." *Behaviour Research Methods* 39 (1): 101–117.
- Hox, J.J., and T.M. Bechger. 1998. "An Introduction to Structural Equation Modelling." *Family Science Review* 11: 354–373.
- Jian, J., A.M. Bisantz, and C.G. Drury. 2000. "Foundations for an Empirically Determined Scale of Trust in Automated Systems." *International Journal of Cognitive Ergonomics* 4 (1): 53–71.
- Lee, J.D., and N. Moray. 1994. "Trust, Self-Confidence, and Operators' Adaptation to Automation." *International Journal of Human-Computer Studies* 40: 153–184.
- Lorenz, B., F. Di Nocera, S. Röttger, and R. Parasuraman. 2002. "Automated Fault Management in a Simulated Space Flight Micro-World." *Aviation, Space, and Environmental Medicine* 73: 886–897.
- Onnasch, L., C.D. Wickens, H. Li, and D. Manzey. 2014. "Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis." *Human Factors* 56 (3): 476–488.
- Parasuraman, R., E. de Visser, M.-K. Lin, and P.M. Greenwood. 2012. "Dopamine Beta Hydroxylase Genotype Identifies Individuals Less Susceptible to Bias in Computer-Assisted Decision Making." *PLoS ONE* 7 (6): e39675. doi:10.1371/journal.pone.0039675
- Parasuraman, R., and D. Manzey. 2010. "Complacency and Bias in Human Use of Automation: A Review and Attentional Synthesis." *Human Factors* 52: 381–410.
- Parasuraman, R., R. Molloy, and I.L. Singh. 1993. "Performance Consequences of Automation-Induced 'Complacency'." *International Journal of Aviation Psychology* 3: 1–23.
- Parasuraman, R., T.B. Sheridan, and C.D. Wickens. 2000. "A Model of Types and Levels of Human Interaction with Automation." *IEEE Transactions on Systems, Man, and Cybernetics – Part A* 30: 286–297.
- Raudenbush, S.W., and A.S. Bryk. 2002. *Hierarchical Linear Models*. 2nd ed. Thousand Oaks, CA: Sage Publications.
- Rovira, E., A. Cross, E. Leitch, and C. Bonaceto. 2014. "Displaying Contextual Information Reduces the Costs of Imperfect Decision Automation in Rapid Retasking of ISR Assets." *Human Factors* 56 (6): 1036–1049.
- Rovira, E., K. McGarry, and R. Parasuraman. 2007. "Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task." *Human Factors* 49: 76–87.
- Sarter, N.B., R.J. Mumaw, and C.D. Wickens. 2007. "Pilots' Monitoring Strategies and Performance on Automated Flight Decks: An Empirical Study Combining Behavioral and Eye-tracking Data." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 49 (3): 347–357.
- Sarter, N., and B. Schroeder. 2001. "Supporting Decision Making and Action Selection Under Time Pressure and Uncertainty: The Case of In-Flight Icing." *Human Factors* 43: 573–583.

- Sheridan, T.B., and W.L. Verplank. 1978. *Human and Computer Control of Undersea Teleoperators* (Technical report). Cambridge, MA: MIT, Man Machine Systems Laboratory.
- Shipstead, Z., D.R.B. Lindsey, R.L. Marshall, and R.W. Engle. 2014. "The Mechanisms of Working Memory Capacity: Primary Memory, Secondary Memory, and Attention Control." *Journal of Memory and Language* 72: 116–141.
- Singh, I.L., R. Molloy, and R. Parasuraman. 1993. "Automation-Induced "Complacency": Development of the Complacency Potential Rating Scale." *International Journal of Aviation Psychology* 3: 111–121.
- Tabachnick, B.G., and L.S. Fidell. 2007. "Multi-Level Linear Modeling." In *Using Multivariate Statistics*. 5th ed., 781–857. Boston, MA: Pearson.
- Unsworth, N., and R.W. Engle. 2007. "The Nature of Individual Differences in Working Memory Capacity: Active Maintenance in Primary Memory and Controlled Search from Secondary Memory." *Psychological Review* 114 (1): 104.
- Wickens, C.D. 1992. *Engineering Psychology and Human performance* (2nd ed.). Scranton, PA: Harper Collins.
- Wickens, C.D., and S.R. Dixon. 2005. *Is There a Magic Number 7 (to the Minus 1)? The Benefits of Imperfect Diagnostic Automation: A Synthesis of the Literature* (Tech. Rep. AHFD-05-01/MAAD-05-1). Savoy: University of Illinois, Aviation Research Lab.
- Wickens, C.D., and S.R. Dixon. 2007. "The Benefits of Imperfect Diagnostic Automation: A Synthesis of the Literature." *Theoretical Issues in Ergonomics Science* 8: 201–212.
- Wickens, C.D., and X. Xu. 2002. *Automation Trust, Reliability and Attention* (Tech. Rep. AHFD-02-14/MAAD-02-2). Savoy: University of Illinois, Aviation Research Lab.

Appendix

Equation for Model 1

Level 1:

$$Accuracy_{it} = \beta_{0it} + r_{it}.$$

Level 2:

$$\beta_{0i} = \gamma_{00} + u_{0i}.$$

Equation for Model 2

Level 1:

$$\begin{aligned} Accuracy_{it} = & \beta_{0it} + \beta_{1it}(\text{Task load}) + \beta_{2it}(\text{AutoSupport}) + \beta_{3it}(\text{Reliab}) \\ & + \beta_{4it}(\text{Task load} * \text{AutoSupport}) + \beta_{5it}(\text{AutoSupport} * \text{Reliab}) \\ & + \beta_{6it}(\text{Reliab} * \text{Task load}) + \beta_{7it}(\text{AutoSupport} \times \text{Reliab} \times \text{Task load}) + r_{it}. \end{aligned}$$

Level 2:

$$\begin{aligned}\beta_{0i} &= \gamma_{00} + u_{0i}, \\ \beta_{1i} &= \gamma_{10}, \\ \beta_{2i} &= \gamma_{20}, \\ \beta_{3i} &= \gamma_{30}, \\ \beta_{4i} &= \gamma_{40}, \\ \beta_{5i} &= \gamma_{50}, \\ \beta_{6i} &= \gamma_{60}, \\ \beta_{7i} &= \gamma_{70}.\end{aligned}$$

Equation for Model 3

Level 1:

$$\begin{aligned}\text{Accuracy}_{it} &= \beta_{0it} + \beta_{1it}(\text{Task load}) + \beta_{2it}(\text{AutoSupport}) + \beta_{3it}(\text{Reliab}) \\ &+ \beta_{4it}(\text{Task load} * \text{AutoSupport}) + \beta_{5it}(\text{AutoSupport} * \text{Reliab}) \\ &+ \beta_{6it}(\text{Reliab} * \text{Task load}) + \beta_{7it}(\text{AutoSupport} \times \text{Reliab} \times \text{Task load}) + r_{it}.\end{aligned}$$

Level 2:

$$\begin{aligned}\beta_{0i} &= \gamma_{00} + \gamma_{01}(\text{WM}) + u_{0i}, \\ \beta_{1i} &= \gamma_{10} + \gamma_{11}(\text{WM}), \\ \beta_{2i} &= \gamma_{20} + \gamma_{21}(\text{WM}), \\ \beta_{3i} &= \gamma_{30} + \gamma_{31}(\text{WM}), \\ \beta_{4i} &= \gamma_{40} + \gamma_{41}(\text{WM}), \\ \beta_{5i} &= \gamma_{50} + \gamma_{51}(\text{WM}), \\ \beta_{6i} &= \gamma_{60}, \\ \beta_{7i} &= \gamma_{70}.\end{aligned}$$