Published in final edited form as:
*Hum Factors*. 2011 December ; 53(6): 672–686. doi:10.1177/0018720811421909.

# Understanding the Effect of Workload on Automation Use for Younger and Older Adults

**Sara E. McBride**, **Wendy A. Rogers**, and **Arthur D. Fisk**
Georgia Institute of Technology, Atlanta, Georgia

## Abstract

**Objective—**This study examined how individuals, younger and older, interacted with an imperfect automated system. The impact of workload on performance and automation use was also investigated.

**Background—**Automation is used in situations characterized by varying levels of workload. As automated systems spread to domains such as transportation and the home, a diverse population of users will interact with automation. Research is needed to understand how different segments of the population use automation.

**Method—**Workload was systematically manipulated to create three levels (low, moderate, high) in a dual-task scenario in which participants interacted with a 70% reliable automated aid. Two experiments were conducted to assess automation use for younger and older adults.

**Results—**Both younger and older adults relied on the automation more than they complied with it. Among younger adults, high workload led to poorer performance and higher compliance, even when that compliance was detrimental. Older adults' performance was negatively affected by workload, but their compliance and reliance were unaffected.

**Conclusion—**Younger and older adults were both able to use and double-check an imperfect automated system. Workload affected how younger adults complied with automation, particularly with regard to detecting automation false alarms. Older adults tended to comply and rely at fairly high rates overall, and this did not change with increased workload.

**Application—**Training programs for imperfect automated systems should vary workload and provide feedback about error types, and strategies for identifying errors. The ability to identify automation errors varies across individuals, thereby necessitating training.

### Keywords

automation; compliance; reliance; age; workload; error

## Introduction

In the past several decades, complex automated systems have been brought into a wider array of situations and contexts, including transportation, health care, and the home. These systems are being operated by a diverse population of users. With the introduction of systems such as in-vehicle navigation systems or automated blood pressure and blood glucose monitors, this expansion of automation has increased the ease, efficiency, and safety of many everyday tasks. However, the potential benefits these systems offer can be attained only if individuals use the automation appropriately.

Although there are many automated systems that are highly reliable, it is important to consider that automation can and does make errors. Because automation can be imperfect, users may have to monitor the automation to detect when the automation errs and recover from those errors by adjusting their behavior. For example, if a navigation system provides inaccurate instructions, the driver should not follow the automation's instructions to avoid a potentially dangerous situation or getting lost. Indeed, researchers have found that as the reliability of automation moves further away from 100%, dependence on the automation tends to decrease (Madhavan & Wiegmann, 2007; Sanchez, Fisk, & Rogers, 2004; Wilkison, Fisk, & Rogers, 2007).

One factor that may influence how people monitor and use imperfect automation is their workload level. In high workload situations, fewer attentional resources may be available for monitoring imperfect automation, potentially resulting in a failure to detect automation errors. Although the construct of workload has appeared in models of automation use, its role as a predictor variable has often been only hypothesized because of the scarce number of empirical studies examining workload (Parasuraman & Mouloua, 1996; Parasuraman & Riley, 1997). Of the research that has been conducted, Dixon and Wickens (2006) found that the costs of low reliability on dependence on automation were amplified in high workload situations. Furthermore, a study conducted by Dixon, Wickens, and Chang (2005) revealed that detection time for system failures was significantly longer for participants experiencing high workload compared to low workload.

To understand how reliability and workload affect the manner in which humans use automation, it is necessary to consider the constructs of reliance and compliance. Reliance and compliance have often been used in the literature as a means of partitioning the analysis of human behavior based on the state of the automation. Automation either is silent or is providing information to the user, such as an alert or instructions. Research has demonstrated the importance of taking these two different states into consideration when analyzing human responses to automation. There is evidence that these two behavioral constructs may be affected differently as a function of the degree and type of errors exhibited by the automation (Dixon & Wickens, 2006; Dixon, Wickens, & McCarley, 2007; Meyer, 2001).

Compliance refers to a user following alerts or instructions presented by the automation. For example, consider a navigation system that provides directions to a driver's destination. If the system instructs the user to take a left turn onto a particular street and the user follows the instruction and takes the left turn without double-checking any map, then the user has *complied with* the system.

Reliance refers to use of the automation when it is silent, which is any time there is not an alert from the automation. Consider a driver who is navigating an unfamiliar stretch of highway and relies on the GPS to know when to exit. If the automation is silent and does not instruct the driver to exit, and the driver continues on the highway without exiting and without double-checking any map, then the driver has *relied on* the system.

Users of imperfect automation have to engage in a supervisory role, monitoring the automation to detect and overcome possible failures, and this process may be negatively affected by workload. Because monitoring the automation is an additional task demand, interactions with imperfect automation often become more complicated. One population that may be particularly affected by the addition of this monitoring task is older adults, who tend to have more difficulties performing two or more tasks concurrently compared to younger adults (Kramer & Madden, 2008). If monitoring automation is indeed more difficult for older adults, one would expect that older adults would comply with and rely on automation to a greater extent than would younger adults.

However, previous research is mixed concerning whether age-related differences in reliance and compliance exist. Some research has found that older adults display greater reliance on automation (Johnson, 2004; Mayer, Fisk, & Rogers, 2008); this may not be desirable if the automation is prone to misses. In addition, there is evidence that older adults do not detect as many automation malfunctions as younger adults in a multitask environment (Vincenzi & Mouloua, 1998, 1999). However, other researchers have documented that patterns of automation usage appear similar between younger and older adults (Ho, Wheatley, & Scialfa, 2005). Furthermore, the research to date has not examined whether there are different patterns between younger and older adults' use of automation as a function of their workload. Considering the potential automated systems have to benefit older adults, such as in-home medical devices, a comprehensive understanding of how older adults interact with automation, particularly under various workload levels, is needed.

The current research was designed to gain a deeper understanding of the role of workload in human–automation interaction, particularly with regard to task performance and rates of compliance and reliance. Workload was manipulated by varying the search-detection demands within a task condition. Furthermore, this research was motivated by a desire to explore how various user groups, namely, younger and older adults, interact with imperfect automation. The study reveals how workload affects both younger and older adults and allows for comparisons regarding general automation usage patterns and the effect of workload across the two experiments. Both experiments used the same automated system in the context of a dual-task scenario. However, the two age groups were tested separately because the difficulty levels of the task had to be adjusted for each age group.

It was expected that higher workload in one task would require more attentional resources, leaving fewer resources available for monitoring the automation in the other task. This reduction in monitoring would presumably lead to greater compliance with and reliance on the automation among the groups experiencing higher workload. Complying with and relying on this imperfect automated system would thus result in failure to detect when the automation committed a miss or a false alarm, leading the higher workload groups' performance in the task with the automated aid to suffer.

These findings would indicate that in situations where human operators experience a high level of workload and must monitor imperfect automation, the operators will not monitor the automation as effectively as if they were experiencing lower workload. However, if the data do not support these hypotheses and workload does not affect use of the automation as expected, it may suggest that individuals use strategies to ensure that they maintain their rate of compliance and reliance regardless of their workload level. The present results add valuable information to models of human–automation interaction regarding the role of workload- and age-related differences.

## Experiment 1

### Method

**Participants—**Participants were 42 younger adults (25 males, 17 females) aged 18 to 28 (*M* = 20.39, *SD* = 2.39) who were recruited through a university-based online experiment scheduling program (see Table 1 for demographic details). They received 3 hr of experiment participation credit for their involvement.

Participants were randomly assigned to the three workload conditions described below; the groups did not differ significantly in terms of their age, education, or health ratings. Three tests were administered to determine whether the groups differed in the following abilities: vocabulary (Shipley, 1986), memory span (Reverse Digit Span; Wechsler, 1997), and perceptual speed (Digit Symbol Substitution; Wechsler, 1997). Table 1 shows that there were no significant differences in these abilities across the workload groups. Moreover, all participants scored within three standard deviations on the ability tests; therefore, no participants were excluded.

**Materials—**The participants interacted with an Automated Warehouse Management System adapted from Mayer et al. (2008) that required them to act as a warehouse manager in charge of two tasks: (a) receiving packages into inventory and (b) dispatching trucks once they were filled to capacity. Participants worked to accumulate as many task points as possible. Performance on each task resulted in the addition or deduction of points. The point system was implemented to encourage participants to maximize performance on both tasks.

In the *Receiving Packages* task, participants were presented with a target bar code and a list of bar codes (see Figure 1). Their goal was to find the bar code in the list that matched the target bar code within a 7-s time limit. They navigated through the list of bar codes using the up and down arrow keys on a keyboard and selected a bar code using a key labeled *Receive*. The following feedback was provided: After making a bar code selection, the word *CORRECT* or *INCORRECT* appeared on the screen, indicating whether or not the choice was correct. If a participant failed to make a response within the 7 s, *TIMEOUT* appeared on the screen. The feedback appeared for 1 s, followed immediately by the next Receiving Packages trial.

Concurrent with the Receiving Packages task, participants were responsible for determining when trucks should be dispatched (*Dispatching Trucks* task). For this latter task, assistance was provided from an automated aid that monitored the interior of a truck to determine when that truck had been filled with boxes. When the automation detected that the truck had reached full capacity, it would provide an alert by displaying the message *DISPATCH TRUCK*. Participants pressed a key labeled *Dispatch* if they believed the truck was full.

A truck continued to be filled with boxes until the participant dispatched it. If the participant did not dispatch it within 10 s of it becoming full, the truck would become overloaded (i.e., filled beyond capacity). If participants failed to dispatch the truck before it overloaded, or if they dispatched the truck before it was full, points were deducted from the Dispatching Trucks tasks. After each trial, feedback indicated whether the dispatched truck was not full yet, full, or overloaded, as well as the points gained or lost. The time until a truck reached its full point varied between 12 and 22 s.

**Seeing Inside the Truck—**Participants could view the interior of the truck at any point by pressing the space bar on the keyboard. The inside of the truck remained visible as long as the space bar was depressed, although only a very brief visual inspection was necessary to determine how full the truck was (see Figure 1). Only one task could be viewed; hence, when the space bar was pressed to see the inside of the truck, the Receiving Packages task disappeared, keeping the participants from performing that task; the time progression for the Receiving Packages task was not paused. When the space bar was released, the Receiving Packages task reappeared on the screen and participants could resume that task. In addition, when the space bar was initially pressed, there was a 2-s delay before the interior of the truck would appear, increasing the amount of time taken away from the Receiving Packages task. This setup was intended to simulate tasks wherein double-checking automation requires time and effort away from other tasks.

A truck could be viewed multiple times. If participants viewed a truck before it was full, they had the opportunity to view it again. If they chose not to view it again, they might be able to estimate how long it would take for it to complete filling and dispatch it at that point. However, it is important to note that a truck could be dispatched without ever viewing its interior. The Dispatch Truck key could be pressed while completing the Receiving Packages task. The only way participants could be certain they were dispatching a full truck was if they viewed the truck in the 10-s window when the truck was visibly full, and then dispatched it before that 10-s window closed and the truck became overloaded.

**Types of Automation Error—**The automated system was capable of making two types of errors, a miss and a false alarm. A miss constituted the automation not recognizing that the truck was full and, consequently, failing to alert the participant. A false alarm occurred when the automation notified the participant that the truck was full before it really was full. False alarms occurred 6 to 10 s from the start of a trial. Within a block, the time from the start of the trial to the false alarm, and the time from the false alarm to the truck being full was the same.

**Procedure—**Participants first provided informed consent. Participants' near and far visual acuity were assessed using a handheld and a wall-mounted Snellen eye chart, respectively. The chart was read from a distance of 14 in. for near vision and 20 ft. for far vision. To be included in the study, participants had to demonstrate at least 20/40 vision for both near and far vision. The Shipley Vocabulary test was administered on the first day of testing; the Reverse Digit Span and Digit Symbol Substitution tests were administered on the second day of testing.

Participants received an overview of the Automated Warehouse Management System and how to interact with it. Participants were not told the reliability level of the automation. They were told only that the automation they would be interacting with is very reliable but may make errors. They were also told that the automation might make two types of errors: a miss or a false alarm. There were seven practice blocks giving experience with both tasks and illustrating correct automation as well as misses and false alarms. Participants then completed eight blocks, with a block defined as the filling of 20 trucks and a single truck defining a trial. These were completed over the course of 2 sequential days (4 blocks per day). Automation reliability was 70%; within each block of 20 trucks, three misses and three false alarms occurred in a random sequence.

**Design: Independent Variable—**Workload (low, moderate, high) was a between-participants variable. Workload was manipulated by varying the search load in the Receiving Packages task. This was accomplished by altering the number of characters in each bar code as well as the number of bar codes in the list of possible matches. The low workload group had 3 characters per bar code and 3 bar codes in the list, the moderate workload group had 4 characters per bar code and 6 bar codes in the list, and the high workload group had 6 characters per bar code and 11 in the list (see Figure 2). The specific values for these two parameters were selected because they evoked the desired level of workload as evidenced by scores on the NASA-TLX measure of subjective workload (Hart & Staveland, 1988) in pilot testing. Each participant was randomly assigned to one of the three workload groups.

**Design: Dependent Variables—**Three performance measures were collected for each task. In the Receiving Packages task, performance was measured as the percentage of trials that were (a) correct matches, (b) incorrect matches, or (c) timed out. For the Dispatching Trucks task, performance was measured as the percentage of trucks that were (a) dispatched on time, (b) dispatched not full, or (c) overloaded. One can think of trucks dispatched not full as false alarms and trucks overloaded as misses. In addition, participants' use of the automation (compliance and reliance) was assessed. Compliance was calculated as p(compliance) = p(truck not viewed AND dispatch truck alert present), and reliance was calculated as p(reliance) = p(truck not viewed AND no alert present).

Compliance and reliance were each subdivided into trials in which the automation was correct and trials in which the automation was incorrect (false alarm or miss). Appropriate compliance and reliance depend on the quality of the automation's suggestions. For each trial, (i.e., an individual truck), the automation can (a) provide correct information, (b) commit a miss, or (c) commit a false alarm. In each situation, the participant should behave in a certain way to ensure success. Therefore, compliance was calculated as p(compliance | automation correct) and p(compliance | automation false alarm). Similarly, reliance was calculated as p(reliance | automation correct) and p(reliance | automation miss). If the automation is correct, compliance and reliance are appropriate because the automation is providing accurate information. However, if the automation commits a false alarm, then compliance is not appropriate; participants can detect the false alarm only if they do not comply but see that an alert to dispatch occurred before the truck was full. If the automation commits a miss, then reliance is not appropriate; participants can detect the miss only if they do not rely but instead check to see if the truck is full. For comparison, optimal compliance and reliance were defined as complying or relying 100% of the time on automation correct trials and 0% of the time on automation incorrect trials. Although these optimal levels of compliance and reliance were not expected, they provide a baseline to assist in interpretation of the findings.

The only cue available to participants that may have enabled them to detect false alarms and misses was the length of time the truck spent filling, which varied between 12 s and 22 s. Although this range in filling time was selected to make it difficult to know when the automation erred, it was not impossible to make a judgment about how long it was taking for a truck to fill.

**Design: Point Structure—**A point structure was devised to motivate participants to do well on both of the tasks. The Dispatching Trucks task earned participants 100 points for a correctly dispatched truck, but a truck that was overloaded or was sent before it was full resulted in a loss of 200 points. Furthermore, to ensure participants attended to both tasks, the points attainable on the two tasks were roughly equivalent. The maximum points attainable in the Dispatching Trucks task in a block was 2,000 (100 points per truck if correctly dispatched, multiplied by 20 trucks per block). Therefore, we sought to create a situation wherein participants could score a maximum of 2,000 points in the Receiving Packages task as well. To this end, we conducted pilot testing to determine the maximum number of Receiving Packages trials that a participant could complete. We then divided 2,000 points by that number to determine how many points each trial should be worth. Because the workload varied on the Receiving Packages task, making it possible to complete more or fewer trials, it was necessary to compute how many points a trial was worth separately for each workload group (see Table 2).

## Results

Statistical tests were conducted using either a one-way analysis of variance with planned contrasts between the workload groups or a chi-square test of independence. The alpha level was set to .05 for all statistical tests.

**Receiving Packages Performance—**Depending on what workload group the participants belonged to, the list of possible matches presented to them in the Receiving Packages task varied in length and complexity. Performance on this task (i.e., correct matches, incorrect matches, or timed out) was examined to understand whether individuals experiencing increased workload had a more difficult time successfully completing the task. As expected, increased workload caused performance on the Receiving Packages task to suffer (see Figure 3). However, performance remained between 85% and 95% correct for all three workload groups.

The analysis of the correct matches revealed that the workload manipulation had a significant, negative effect on the percentage of trials that were correctly matched, such that low workload was associated with a greater percentage of trials that were correctly matched, *F*(2, 39) = 58.01, *p* < .01, $^2$ = .74 (see Figure 3). Planned contrasts revealed the low and moderate workload groups had a significantly higher percentage correct than did the high workload group, *t*(39) = 10.11, *p* < .01, *t*(39) = 8.28, *p* < .01, respectively. There was not a statistically significant difference between the low and moderate workload groups, *p* = .08.

A chi-square test of independence was performed to determine whether the pattern of errors differed as a function of workload. Error type differed across workload group, $^2$(2, *N* = 4,527) = 528.17, *p* < .001. To determine which cell or cells produced the statistically significant results, residuals (the difference between the observed frequency and the expected frequency) were converted to *z* scores and compared to a critical value corresponding to an alpha of .05 (i.e., ±1.96). Within the low and moderate workload groups, incorrect errors were above the expected frequency and timeout errors were below the expected frequency. In the high workload group, the pattern switched; incorrect errors were below the expected frequency, and time-out errors were above the expected frequency.

**Dispatching Trucks Performance—**Although the workload manipulation occurred within the Receiving Packages task, it was expected that workload would also influence performance on the Dispatching Trucks task by affecting compliance and reliance behavior. Performance in the Dispatching Trucks task is depicted in Figure 3. Performance remained between 85% to 90% correct, even in the high workload group. Similar to the effect observed in the Receiving Packages task, workload caused Dispatching Trucks performance to suffer.

The data show that the percentage of trucks correctly dispatched was lower in the high workload group, *F*(2, 39) = 3.63, *p* = .04, $^2$= .16. Planned contrasts revealed the high work-load group had a significantly lower percentage correct than the low workload group, *t*(39) = 2.63, *p* = .01. There was not a significant difference between the low and moderate or moderate and high workload groups for percentage correct, all *p*s > .07.

A chi-square test of independence was performed to determine whether the pattern of errors differed as a function of workload. The analysis did not provide support for the notion that error type differed across workload group, $^2$(2, *N* = 699) = 5.53, *p* = .06.

**Compliance With Automation—**Increased workload led to higher levels of compliance with the automation, both when the automation was correct and when it committed a false alarm (see Figure 4). That is, participants were more likely to refrain from viewing the truck when workload was high. Planned contrasts revealed that the high workload group complied more than the low workload group, and this was true for trials where the automation was correct as well as false alarms, *t*(39) = –2.10, *p* = .04; *t*(39) = –2.35, *p* = .02. No significant differences were found between the low and moderate or moderate and high workload groups for automation correct or false alarm trials all *p*s > .21.

Note that when the automation was incorrect, participants in the high workload group were more likely to comply with the automation, meaning that they did not verify its instruction before sending the truck, leading to a situation in which they erroneously dispatched a truck that was not full. Recall that the Dispatching Trucks performance data revealed that participants in the high workload group had significantly more instances of dispatching trucks that were not full. This finding is likely a result of the high workload group's greater compliance with the automation when it was committing a false alarm.

Another pattern to note is that compliance was lower in trials where the automation false alarmed compared to the automation correct trials. Compliance behavior differed as a function of whether the automation was providing accurate or inaccurate information. This may indicate that participants had some ability to recognize when the automation was erring and adjusted their compliance behavior accordingly on a trial-to-trial basis. The only cue available to participants that could be used to identify automation false alarms was the temporal nature of this task. Although the time required for the truck to fill varied between 12 and 22 s, it is feasible that participants used some kind of time-based strategy to determine whether an alert was early or not.

**Reliance on Automation—**Reliance was not significantly affected by workload, and this was true for reliance when the automation was correct as well as during miss trials, all *p*s > .29 (see Figure 4). Planned contrasts did not reveal significant differences between any of the workload groups, all *p*s > .16.

Similar to the compliance data, the reliance data showed that participants adjusted their reliance behavior to rely on the automation less on miss trials. This suggests that misses, although less perceptually salient than false alarms, were detected to some degree and led participants to reduce their reliance in those trials.

## Discussion

Participants who experienced higher workload performed significantly worse on both the Receiving Packages and Dispatching Trucks tasks. Among the groups experiencing higher workload, participants were more likely to comply with the automation, both when the automation was correct and incorrect. Because participants in the high workload group complied with the automation to a greater degree, even when it was incorrect, the high workload group had a significantly higher rate of trucks that were dispatched before they were full. However, reliance behavior was not significantly influenced by workload.

Compliance and reliance both decreased for trials where the automation erred (false alarmed or missed, respectively) compared to trials where the automation was correct. This may indicate participants' ability to detect automation errors as they are happening and to modify their behavior to overcome the errors.

As evidenced by the results of Experiment 1, workload can lead to significant differences in the way individuals interact with automated systems. Although it is not surprising that high workload was detrimental to performance on the Receiving Packages task, it was not obvious that the effect of high workload would carry over to the Dispatching Trucks task. However, because high workload led individuals to comply with automation to a greater degree, even when they should not have, their performance on the Dispatching Trucks task suffered as well.

# Experiment 2

## Overview

Another goal of this research was to examine how older adults interact with imperfect automation and how this may be affected by their level of workload. There has been some evidence that older adults may use automation differently, such as depending on automation to a greater extent than their younger counterparts (Johnson, 2004; Mayer et al., 2008). These differences may be exacerbated in situations where workload is high, as monitoring imperfect automation on top of performing the tasks becomes more challenging, particularly for older adults who have more difficulty performing multiple tasks simultaneously (Kramer & Madden, 2008). Therefore, in Experiment 2 older adults interacted with the same automated system as the one used in Experiment 1 to determine the effect of workload on older adults' performance and automation interactions. As explained in the method section, the difficulty level was adjusted to ensure that older adults could perform the task.

## Method

**Participants—**Participants consisted of 42 older adults (19 males, 23 females) aged 65 to 75 (*M* = 70.17, *SD* = 2.8) from the community who received $12 per experimental hour. The study was approximately 3.5 hr long. Table 3 provides the demographic information for the three workload groups; the only significant difference was for education, wherein the high workload group had more formal education than the other groups. However, as also evidenced by Table 3, there were no significant differences for any of the ability measures across the workload groups. In addition, all participants scored within three standard deviations on the ability tests; therefore, no participants were excluded.

**Changes to Experiment 1 Method—**The stimuli, design, and procedure of Experiment 2 were identical to those of Experiment 1 except for the following changes. In the Receiving Packages task, older adults were given 10 s to complete the task rather than 7 s. This modification was made to compensate for decreases in attention-switching ability and slowing of response times for older adults (Craik & Salthouse, 2000).

In addition, the characteristics of the bar code list were altered for the older adults because preliminary testing revealed that the workload levels used in Experiment 1 were too difficult for older adults, leading to much lower levels of performance. Therefore, the low group had 1 character per bar code and 2 bar codes in the list, the moderate group had 3 characters per bar code and 3 bar codes in the list, and the high group had 4 characters per bar code and 6 in the list (see Figure 5).

Finally, the point structure was changed as a result of preliminary testing that revealed older adults completed fewer trials of the Receiving Packages task than younger adults. Because of this, and the fact that we wanted to keep the points attainable across both tasks equal (just as we did in Experiment 1), older adults received more points per Receiving Packages trial so that their maximum total would remain at 2,000 points for that task.

## Results

**Receiving Packages Performance—**Participants' performance on the Receiving Packages task is presented in Figure 6. Overall, performance ranged from 85% to 95% for the three groups. As expected, the performance of participants experiencing high workload was worse than of those experiencing low workload. High workload was associated with fewer correct matches, $F(2, 39) = 10.4$, $p < .01$, $^{2} = .35$. For the correct trials, significant differences were found among all three workload groups, such that the low and moderate workload groups had a higher percentage correct than participants in the high workload group, $t(39) = 4.55$, $p < .01$; $t(39) = 2.55$, $p = .02$, and participants in the low group had a higher percentage correct than those in the moderate workload group, $t(39) = 2.00$, $p = .05$.

A chi-square test of independence was performed to determine whether the pattern of errors differed as a function of workload. Error type differed across workload group, $^{2}(2, N = 5{,}028) = 277.61$, $p < .001$. Residual analysis revealed that within the low workload group, incorrect errors were above the expected frequency and time-out errors were below the expected frequency. In the moderate and high workload groups, the pattern switched; incorrect errors were below the expected frequency, and time-out errors were above the expected frequency.

**Dispatching Trucks Performance—**Dispatching Trucks performance is presented in Figure 6. Percentage correct averaged approximately 80%, which is lower than performance on the Receiving Packages task. Considering the automation's reliability level was 70%, participants were using the automation to their benefit. If they were simply depending on it without ever double-checking it, their performance would have matched the reliability of the automation at 70%. In actuality, their performance data indicate that participants were able to catch and overcome some of the errors committed by the automation.

As can be seen in Figure 6, performance on the Dispatching Trucks task did not follow the same pattern as that of the Receiving Packages performance. There was not a significant effect of workload on trucks dispatched correctly, $p = .47$.

A chi-square test of independence was performed to determine whether the pattern of errors differed as a function of workload. Error type differed across workload group, $^{2}(2, N = 1{,}255) = 13.77$, $p = .001$. To determine which cell or cells produced the statistically significant results, residuals (the difference between the observed frequency and the expected frequency) were converted to $z$ scores and compared to a critical value corresponding to an alpha of .05 (i.e., ±1.96). Within the low workload group, not full trucks were below the expected frequency.

Although the analysis did not reveal significant differences between the workload groups, one pattern that emerged in the Dispatching Trucks performance data was that the moderate workload group had numerically greater performance than the low or high workload groups. The moderate workload group had a higher percentage of trucks correctly dispatched and fewer trucks dispatched early or overloaded. Although these were not statistically significant differences, this pattern reappeared in the compliance and reliance data.

**Compliance With Automation—**Participants' compliance (i.e., the percentage of trucks in which they chose not to view the interior of the truck when the Dispatch Truck alert appeared) was examined (see Figure 7). Participants were noncomplying (i.e., viewing the truck and trying to catch automation false alarms) approximately 40% of the time when the automation was correct and 60% of the time when the automation was false alarming. This may be surprising given that one might expect older adults to have difficulty with checking the truck in addition to performing the Receiving Packages task, based on the dual-tasking literature (Kramer & Madden, 2008).

However, these data demonstrate that participants recognized the benefit of double-checking the truck to catch false alarms and were able to do so, thereby enhancing their dispatching trucks performance. A comparison of compliance when the automation was correct versus when the automation false alarmed showed that compliance was lower during false alarms, which is in line with what would be considered optimal behavior. Participants should not comply with automation that is providing false alarms. Participants were thus displaying some level of error detection.

An examination of the differences between the workload groups revealed that workload did not appear to have an effect on compliance behavior, when the automation either was correct or false alarmed, all *p*s > .44. Planned contrasts showed there were no significant differences between any of the workload groups for compliance during either automation correct or incorrect trials, all *p*s > .21.

Although the effect of workload on compliance behavior was not significant, the pattern noted in the Dispatching Trucks section involving the moderate workload group can be seen here as well. The moderate workload group complied numerically less than the other two workload groups when the automation was incorrect, meaning that the moderate group had fewer instances wherein they dispatched the truck when they should not have because the alert was early. This likely contributed to their higher level of performance in the Dispatching Trucks task, as they would have more correctly dispatched trucks and fewer trucks dispatched early, or before they were full.

**Reliance on Automation—**Participants' reliance behavior is shown in Figure 7. Reliance is defined as the act of not viewing the interior of the truck when the automation is silent. There was not a main effect of workload on reliance for correct or miss trials (all *p*s > .18), and planned contrasts did not reveal any significant differences, all *p*s > .10.

The difference between reliance behavior in automation correct trials and miss trials was smaller than the difference observed in the compliance data, suggesting that knowing when not to rely may be particularly difficult. It may be that reliance does not differ as much because, at least for the present task, choosing not to rely is primarily driven by an internally generated cue rather than an externally generated cue, as is the case with the Dispatch Truck alert in the compliance state. Because there is no salient perceptual event associated with the reliance state, these internally generated cues may be less likely to occur, especially for older adults.

### Discussion

Workload had a significant, negative effect on performance in the Receiving Packages task but did not reliably affect performance in the Dispatching Trucks task. Participants' compliance with and reliance on the automation also did not reveal any statistically significant effects of workload. It may be the case that there was greater variability in participants' ability to perform both tasks simultaneously that diluted the effect of the workload manipulation. For example, in the low workload group one would have expected participants to check the truck with greater ease than participants in the high workload group. However, some participants in the low workload group may have found it more difficult to check the truck than participants in the high workload group because of the variability in certain cognitive abilities such as processing speed and task switching abilities among older adults.

Participants tended to rely on the automation more than they complied with it, even when this reliance was to their detriment. Participants did comply less on automation false alarm trials, demonstrating a level of error detection for false alarms. Reliance, however, did not show as much of a difference between automation correct and miss trials, revealing that misses might have been harder to detect.

## General Discussion

This study was designed to assess how younger and older adults interact with imperfect automation and to assess the role of workload in this interaction. These findings add to current models of human automation interaction regarding the role of workload, automation reliability, and age.

High workload was associated with poorer performance in the Receiving Packages task for both age groups, but the Dispatching Trucks performance suffered from increased workload only for the younger adults. However, the performance level of the older adults was lower for all three workload conditions on the Dispatching Trucks task compared to the younger adults. Furthermore, older adults' Dispatching Trucks performance was lower than their Receiving Packages performance, whereas younger adults' performance was more comparable across the two tasks. Older adults' reliance data showed they did not appear to detect automation misses, which may explain this deficit in Dispatching Trucks performance. However, the Dispatching Trucks data also showed that both groups used the automation to their benefit. That is, their performance exceeded the reliability of the automation, demonstrating that participants were able to successfully use the automation and catch some of the automation errors.

In terms of the types of errors participants made, both age groups demonstrated an effect of workload on the pattern of errors they committed in the Receiving Packages task. In the low workload groups, there were more incorrect errors and fewer time-out errors than expected. As workload increased, this pattern switched, resulting in fewer incorrect errors and more time-out errors than expected. This suggests that in high workload scenarios, a user may not make any response rather than making an incorrect response. However, in the Dispatching Trucks task, only older adults' pattern of errors was influenced by workload, and this was true only in the low workload group. Perhaps because the automation assisted in the Dispatching Trucks task and it committed an equivalent number of misses and false alarms, the distribution of errors participants made in response were also roughly equal across the two error types.

To understand how participants' use of the automation led to the observed patterns of performance, participants' compliance with and reliance on the automation were examined. Among younger adults, higher workload led to higher compliance. Because the high workload group complied with the automation when it was providing a false alarm, that group dispatched significantly more trucks that were not yet full. Reliance, on the other hand, was not significantly affected by workload among the younger adults. Among older adults, workload did not have a significant effect on compliance or reliance. This may be the result of a greater amount of variability among older adults in terms of their willingness or ability to double-check the automation while engaged in the Receiving Packages task. It may also be related to the fact that older adults were complying and relying at a relatively high rate in all three workload conditions.

Both younger and older adults tended to rely on the automation more than they complied with it, perhaps because of the nature of false alarms and misses. Indeed, Dixon et al. (2007) found that false alarms tend to suppress both compliance and reliance behavior, whereas misses only decrease reliance, resulting in a situation where compliance should be higher than reliance because it is affected by only one error type. However, in the current study with an equal number of misses and false alarms, reliance was in fact *higher* than compliance.

These results may differ because automation error type was a between-subjects variable in the Dixon et al. study, whereas participants in the present study experienced both misses and false alarms. When presented with two types of automation errors to monitor, an individual may adopt a strategy of focusing primarily on one error type to detect (perhaps the most salient one). However, because many automated systems may present both types of errors to operators, this point warrants further investigation. Indeed, if this study had examined only dependence, many of the important patterns discovered in reliance and compliance behavior would likely have been missed. These findings emphasize the importance of examining automation use in its two forms of compliance and reliance as suggested by Meyer (2001). The nature of the automation may have to vary depending on how it is supporting performance.

The data support previous findings associated with age-related differences in compliance and reliance behavior (Johnson, 2004; Mayer et al., 2008). Although direct comparisons could not be made because of design differences across the two present experiments, the pattern suggested that older adults complied with and relied on the automation to a greater extent than younger adults. Further research is necessary to understand why these patterns of compliance and reliance were observed across the two age groups. In addition, there was also evidence that older adults may not be as sensitive to automation errors as are younger adults, as postulated by Ho et al. (2005).

The findings generated by this study may contribute to the design of automation training regimens. Training for users of imperfect automation should encompass the various levels of workload that users might be expected to experience. If certain workload scenarios affect operators' compliance and reliance behavior to the point of severely disrupting performance, designers may need to incorporate adaptive automation systems that take control when workload exceeds certain boundaries. However, as noted in this study, even the high workload group used the automation to their advantage; therefore, testing will be required to determine when adaptive automation may be necessary.

Another aspect of training that may be informed by the findings of this study is training users to detect errors and modify their behavior accordingly. Depending on the automation in question, this may be relatively easy or difficult to do. However, understanding how to increase a user's ability to identify errors should be a training priority. Supporting users in their identification of automation errors may be especially critical for older adults, who may be less sensitive to automation errors. Users will be better prepared to interact with imperfect automation if their training includes exposure to the type of errors they may encounter, useful feedback that identifies these errors, and information about what steps to take to overcome the errors.

## Acknowledgments

This research was supported in part by Grant P01 AG17211 from the National Institutes of Health (National Institute on Aging) under the auspices of the Center for Research and Education on Aging and Technology Enhancement (CREATE, www.create-center.org). This research was also supported in part by contributions from Deere & Company. We thank Jerry Duncan and Josh Hoffman for their support and advice on this research. Some data and analyses presented here were part of the first author's master's thesis (McBride, 2010). Portions of these data were presented at the 2010 Cognitive Aging Conference (Atlanta, April, 2010) as well as the Human Factors and Ergonomics Society 54th Annual Meeting (San Francisco, September, 2010) and appear in the proceedings of that conference (McBride, Rogers, & Fisk, 2010).

## Biographies

Sara E. McBride is a graduate student in the Engineering Psychology program at the Georgia Institute of Technology in Atlanta, Georgia, where she received her MS in psychology in 2010.

Wendy A. Rogers is a professor in the School of Psychology at the Georgia Institute of Technology in Atlanta, Georgia, where she received her PhD in psychology in 1991.

Arthur D. Fisk is a professor in the School of Psychology at the Georgia Institute of Technology in Atlanta, Georgia. He received his PhD in psychology from the University of Illinois in 1982.

## References

Craik, FIM., Salthouse, TA., editors. The handbook of aging and cognition. 2nd. Mahwah, NJ: Lawrence Erlbaum; 2000.

Dixon SR, Wickens CD. Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. Human Factors. 2006; 48:474–486. [PubMed: 17063963]

Dixon SR, Wickens CD, Chang D. Mission control of multiple unmanned aerial vehicles: A workload analysis. Human Factors. 2005; 47:479–487. [PubMed: 16435690]

Dixon SR, Wickens CD, McCarley JS. On the independence of compliance and reliance: Are automation false alarms worse than misses? Human Factors. 2007; 49:564–572. [PubMed: 17702209]

Hart, SG., Staveland, LE. Development of NASATLX (task workload index): Results of empirical and theoretical research. In: Hancock, PA., Meshkati, N., editors. Human mental workload. Amsterdam, Netherlands: Elsevier; 1988. p. 139-183.

Ho G, Wheatley D, Scialfa CT. Age differences in trust and reliance of a medication management system. Interacting with Computers. 2005; 17:690–710.

Johnson, JD. Unpublished master's thesis. Georgia Institute of Technology; Atlanta: 2004. Type of automation failure: The effects on trust and reliance in automation.

Kramer, A., Madden, D. Attention. In: Craik, FIM., Salthouse, TA., editors. The handbook of aging and cognition. 3rd. New York, NY: Psychology Press; 2008. p. 189-249.

Madhavan P, Wiegmann DA. Effects on information source, pedigree, and reliability on operator interaction with decision support systems. Human Factors. 2007; 49:773–785. [PubMed: 17915596]

Mayer, AK., Fisk, AD., Rogers, WA. Age differences, expectancy, and resultant reliance and compliance behavior using an automated system; 2008, September; Paper presented at the 52nd annual meeting of the Human Factors and Ergonomics Society; Santa Monica, CA.

Meyer J. Effects of warning validity and proximity on responses to warnings. Human Factors. 2001; 43:563–572. [PubMed: 12002005]

Parasuraman, R., Mouloua, M., editors. Automation and human performance: Theory and applications. Hillsdale, NJ: Lawrence Erlbaum; 1996.

Parasuraman R, Riley V. Humans and automation: Use, misuse, disuse, abuse. Human Factors. 1997; 39:230–253.

Sanchez, J., Fisk, AD., Rogers, WA. Reliability and age-related effects on trust and reliance of a decision support aid; 2004, September; Paper presented at the 48th annual meeting of the Human Factors and Ergonomics Society; Santa Monica, CA.

Shipley, WC. Shipley Institute of Living Scale. Los Angeles, CA: Western Psychological Services; 1986.

Vincenzi, D., Mouloua, M. Monitoring automation failures: Effects of age on performance and subjective workload; 1998, October; Paper presented at the 42nd annual meeting of the Human Factors and Ergonomics Society; Santa Monica, CA.

Vincenzi, D., Mouloua, M. Monitoring automation failures: Effects of age on performance and subjective workload. In: Scerbo, M., Mouloua, M., editors. Automation technology and human performance: Current research and future trends. Mahwah, NJ: Lawrence Erlbaum; 1999. p. 253-257.

Wechsler, D. Wechsler Adult Intelligence Scale III. 3rd. San Antonio, TX: Psychological Corporation; 1997.

Wilkison, BD., Fisk, AD., Rogers, WA. Effects of mental model quality on collaborative system performance; 2007, October; Paper presented at the 51st annual meeting of the Human Factors and Ergonomics Society; Santa Monica, CA.

### Key Points

- Automated systems may be used by a range of individuals, including older adults, and may be used in situations that impose varying levels of workload on users.
- This study investigated how younger and older adults use an imperfect automated aid under conditions of low, moderate, and high workload.
- High workload was associated with lower task performance in both experiments and was also associated with higher compliance among younger adults in Experiment 1, even when this compliance was detrimental to task performance.
- Designers of imperfect automated aids and their training programs will need to consider how a user's workload level and age-related characteristics may affect his or her ability to use the automation effectively.

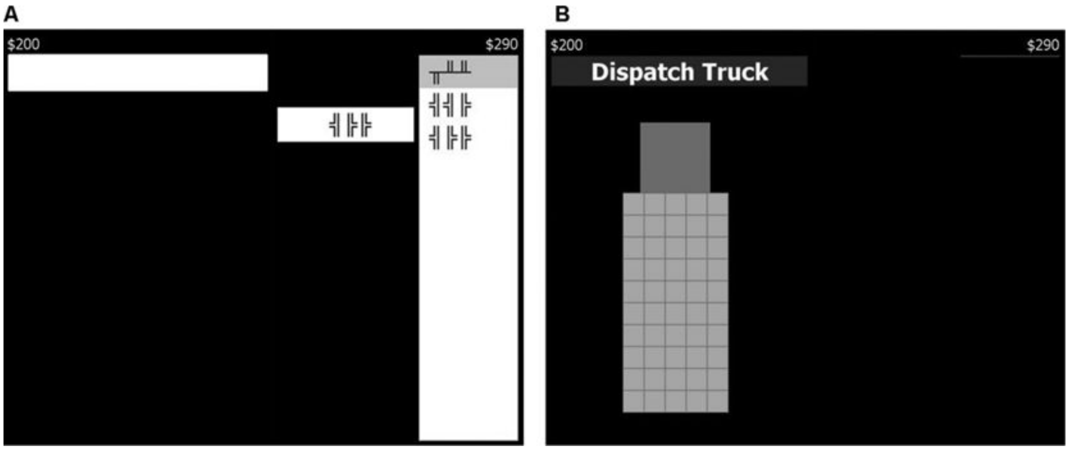

**Figure 1.**

Screenshots of the components of the Automated Warehouse Management System: (a) the Receiving Packages task was presented on the right side of the computer screen and (b) the Dispatching Trucks task was presented on the left side of the computer screen. If participants held down the space bar, they would see the interior of the truck, as depicted in Panel B. However, when the space bar was not depressed, participants saw only what is depicted in Panel A.

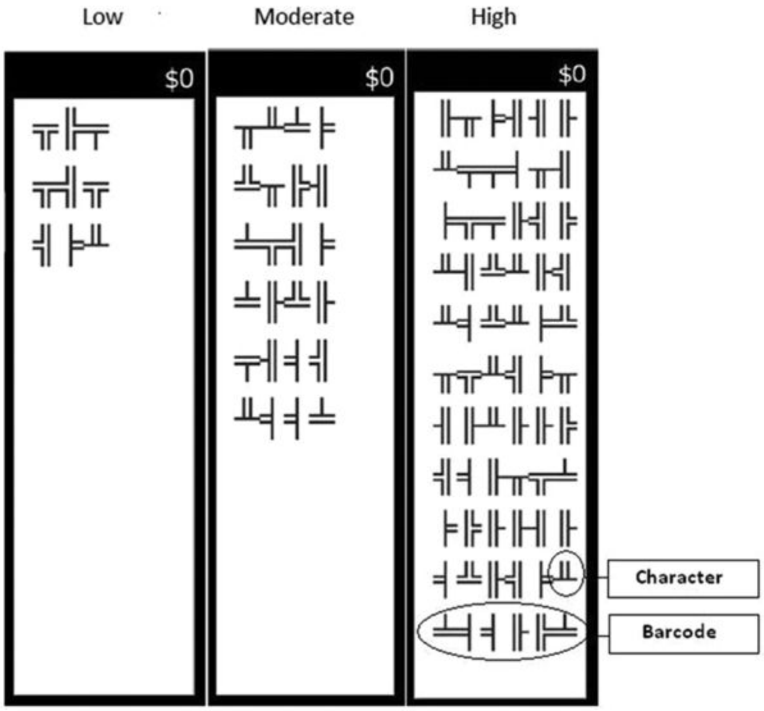

**Figure 2.**
The Receiving Packages search load for the low, moderate, and high workload groups in Experiment 1.

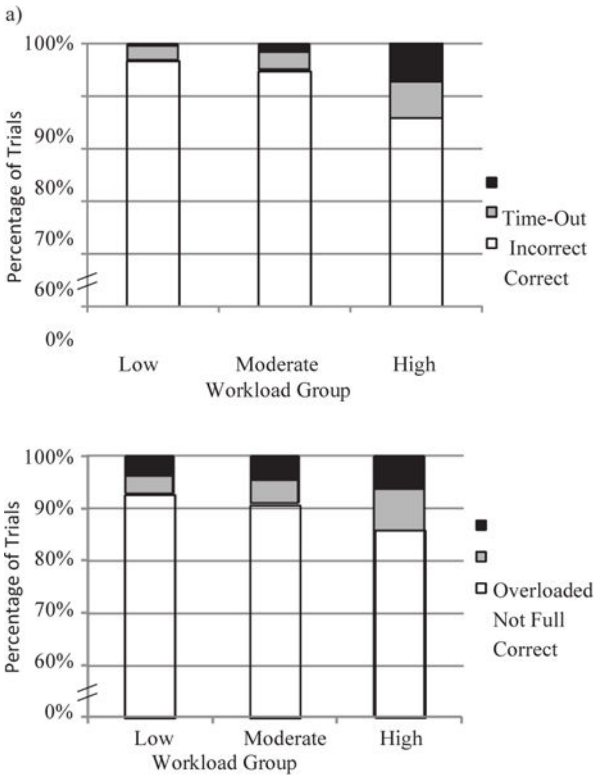

**Figure 3.**
Younger adults' (a) percentage correct, incorrect, and time-out on the Receiving Packages task and (b) percentage correct, not full, and overloaded in the Dispatching Trucks task for each workload condition.

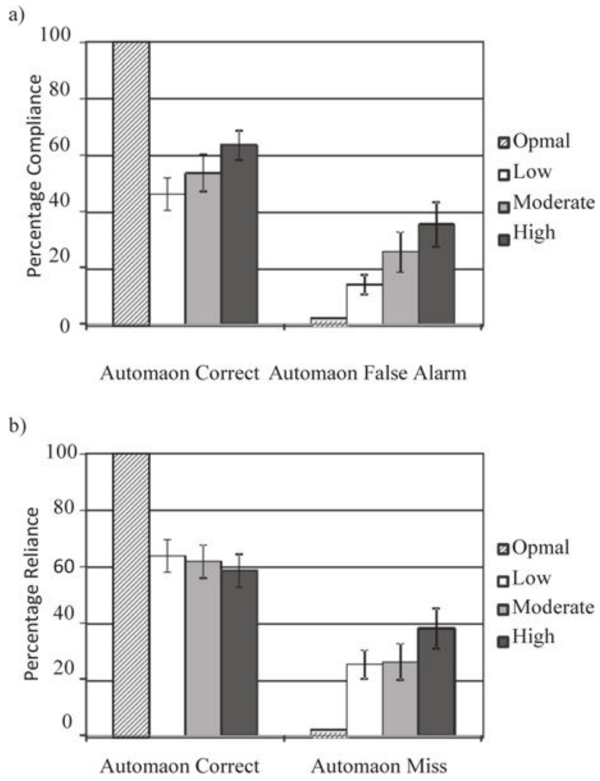

**Figure 4.**

Younger adults’ (a) compliance with and (b) reliance on the automation for trials in which the automation was correct as well as trials in which the automation was incorrect (false alarm or miss, respectively). Optimal compliance or reliance is approximately 100% for correct trials and 0% for incorrect trials. Error bars depict the standard error of the mean.

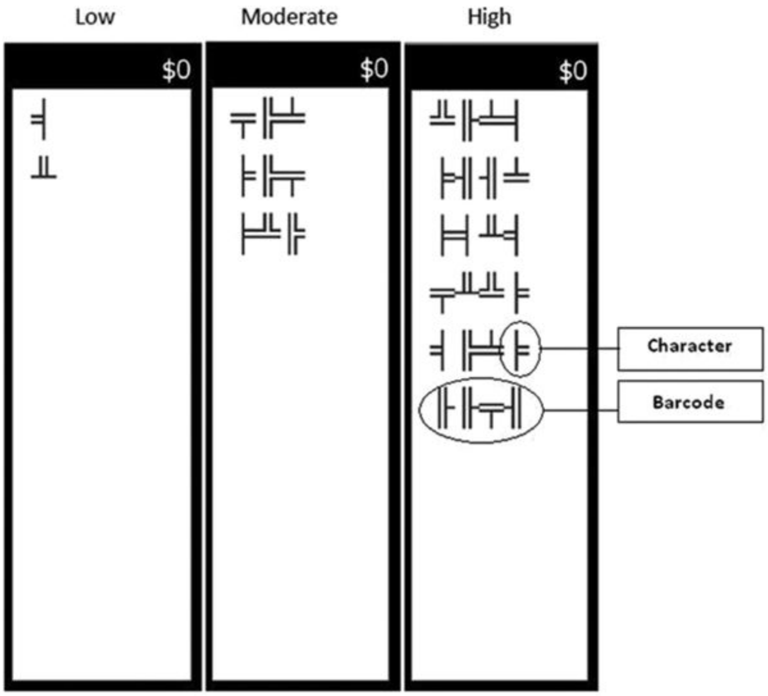

**Figure 5.**

The Receiving Packages search load for the low, moderate, and high workload groups in Experiment 2.

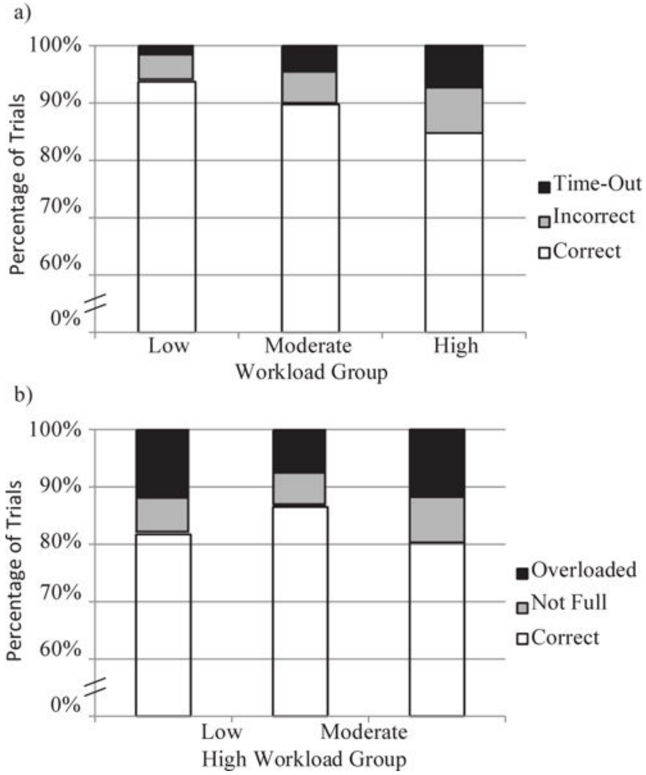

**Figure 6.**

Older adults' (a) percentage correct, incorrect, and time-out on the Receiving Packages task and (b) percentage correct, not full, and overloaded in the Dispatching Trucks task for each workload condition.

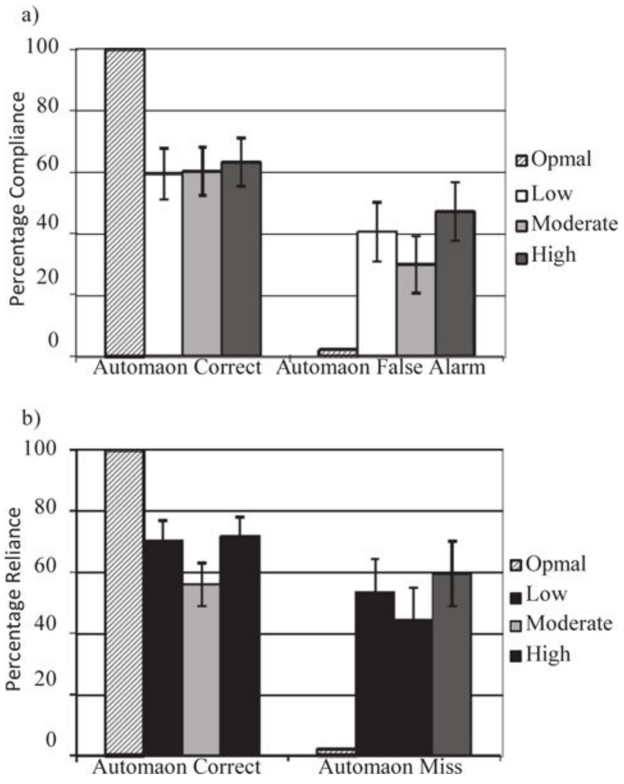

**Figure 7.**

Older adults' (a) compliance with and (b) reliance on the automation for trials in which the automation was correct as well as trials in which the automation was incorrect (false alarm or miss, respectively). Optimal compliance or reliance is approximately 100% for correct trials and 0% for incorrect trials. Error bars depict the standard error of the mean.

## Table 1

Experiment 1: Younger Adult Means and Standard Deviations for the Demographic and Ability Data

| | Low Workload | | Moderate Workload | | High Workload | | | |
|---|---|---|---|---|---|---|---|---|
| | ***M*** | ***SD*** | ***M*** | ***SD*** | ***M*** | ***SD*** | ***F* Value** | **Sig.** |
| Gender (*n*) | | | | | | | | |
| Male | 7 | | | 10 | | 8 | | — Female |
| | | 7 | | 4 | | 6 | | — |
| Age | 19.69 | 1.44 | 19.86 | 2.11 | 21.57 | 2.98 | 2.87 | .07 |
| Education (years) | 13.43 | 1.22 | 13.14 | 1.03 | 13.86 | 1.23 | 1.34 | .28 |
| Health[a] | 4.14 | 0.66 | 4.07 | 0.62 | 4.00 | 0.55 | 0.19 | .83 |
| Digit Symbol Substitution[b] | 78.00 | 13.91 | 70.64 | 8.05 | 74.36 | 10.85 | 1.51 | .23 |
| Reverse Digit Span[b] | 11.29 | 2.13 | 10.57 | 1.79 | 9.86 | 2.32 | 1.64 | .21 |
| Shipley Vocabulary[c] | 30.71 | 4.95 | 32.14 | 2.88 | 30.93 | 2.70 | 0.62 | .54 |

[a] 1 = *poor*, 2 = *fair*, 3 = *good*, 4 = *very good*, 5 = *excellent*.

[b] Wechsler (1997); score is the number correct.

[c] Shipley (1986); score is the number correct.

## Table 2

Receiving Packages Task Point Structure: Points Gained or Loss by Workload Group

| | Low | Moderate | High |
|---|---|---|---|
| Experiment 1:<br>Younger adults | 10 | 15 | 20 |
| Experiment 2:<br>Older adults | 15 | 20 | 25 |

## Table 3

Experiment 2: Older Adult Means and Standard Deviations for the Demographic and Ability Data

| | Low Workload | | Moderate Workload | | High Workload | | | |
|---|---|---|---|---|---|---|---|---|
| | ***M*** | ***SD*** | ***M*** | ***SD*** | ***M*** | ***SD*** | ***F* Value** | **Sig.** |
| Gender (*n*) | | | | | | | | |
| Male | 2 | | 8 | | | 9 | | — Female |
| | | 12 | | 6 | | 5 | | — |
| Age | 70.71 | 2.70 | 69.86 | 2.80 | 69.93 | 3.02 | 0.77 | .47 |
| Education (years) | 14.07 | 2.43 | 15.07 | 2.09 | 17.36 | 2.37 | 7.48 | .01 |
| Health[a] | 3.54 | 0.88 | 3.92 | 0.76 | 3.71 | 0.73 | 0.76 | .47 |
| Digit Symbol Substitution[b] | 51.64 | 13.61 | 47.57 | 9.96 | 52.71 | 10.20 | 0.80 | .46 |
| Reverse Digit Span[b] | 8.14 | 3.74 | 10.21 | 2.67 | 9.21 | 2.15 | 1.75 | .19 |
| Shipley Vocabulary[c] | 33.36 | 6.08 | 33.00 | 6.34 | 36.93 | 2.37 | 2.39 | .11 |

[a] 1 = *poor*, 2 = *fair*, 3 = *good*, 4 = *very good*, 5 = *excellent*.

[b] Wechsler (1997); score is the number correct.

[c] Shipley (1986); score is the number correct.