



AFRL-AFOSR-VA-TR-2015-0400

Anthropomorphic Interfaces on Automation Trust, Dependence, and Performance in younger and Older Adults

**Chong Pak
CLEMSON UNIVERSITY**

**10/26/2015
Final Report**

DISTRIBUTION A: Distribution approved for public release.

**Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTA2
Arlington, Virginia 22203
Air Force Materiel Command**

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188		
<p>The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to the Department of Defense, Executive Service Directorate (0704-0188). Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.</p> <p>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.</p>					
1. REPORT DATE (DD-MM-YYYY) 10/14/2015		2. REPORT TYPE Final report		3. DATES COVERED (From - To) 15 Jul 12 - 14 Jul 15	
4. TITLE AND SUBTITLE Anthropomorphic Interfaces on Automation Trust, Dependence, and Performance in younger and Older Adults			5a. CONTRACT NUMBER FA9550-12-1-0385		
			5b. GRANT NUMBER		
			5c. PROGRAM ELEMENT NUMBER		
6. AUTHOR(S) Richard Pak			5d. PROJECT NUMBER		
			5e. TASK NUMBER		
			5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) CLEMSON UNIVERSITY OFFICE OF SPONSORED PROGRAMS 201 SIKES HALL CLEMSON SC 29634-0001			8. PERFORMING ORGANIZATION REPORT NUMBER		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Department of the Air Force Air Force Office of Scientific Research 875 North Randolph Street Arlington, VA 22203-1768			10. SPONSOR/MONITOR'S ACRONYM(S) AFOSR		
			11. SPONSOR/MONITOR'S REPORT NUMBER(S)		
12. DISTRIBUTION/AVAILABILITY STATEMENT Distribution A					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT This proposal sought to better understand the psychological component of human-automation interaction with a focus on understanding what makes automation seem "trustable". Specifically, we will investigate the role of anthropomorphic automation on operator's trust, dependence, and performance with automation. Evidence from the literature and our own recently collected data suggests that the design of automation can affect how operators perceive the automation and their likelihood of using it. We seek to investigate the conditions under which anthropomorphized automation, or automation that appears to possess human-like characteristics, affects the calibration of trust between the operator and the system. A secondary goal is to understand how anthropomorphic automation effects are moderated by the age of the operator. Older users have different reactions to automation (some research shows over-trust while other research shows under-trust).					
15. SUBJECT TERMS automation, anthropomorphic, trust, aging					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER (Include area code)

FINAL REPORT

Anthropomorphic Interfaces on Automation Trust, Dependence, and Performance in younger and Older Adults

Richard Pak PhD
Clemson University

Executive Summary

This proposal sought to better understand the psychological component of human-automation interaction with a focus on understanding what makes automation seem “trustable”. Specifically, we will investigate the role of anthropomorphic automation on operator’s trust, dependence, and performance with automation. Evidence from the literature and our own recently collected data suggests that the design of automation can affect how operators perceive the automation and their likelihood of using it. We seek to investigate the conditions under which anthropomorphized automation, or automation that appears to possess human-like characteristics, affects the calibration of trust between the operator and the system. A secondary goal is to understand how anthropomorphic automation effects are moderated by the age of the operator. Older users have different reactions to automation (some research shows over-trust while other research shows under-trust).

The general goal this project was to examine how extensive use of social responses deliberately engendered by anthropomorphic agents could convey to operators the “trustability” of automation and how this is affected by operator characteristics. Given some of the observed effects of minimal anthropomorphism (our study; Parasuraman & Miller, 2004) what are the critical factors that must be manipulated to affect perceptions of trust and dependence? Under what conditions do we observe effects? Ultimately, the goal was to encourage proper human-automation calibration such that the user relies on the automation when he should but does not when he should not.

The project’s three specific aims along with research products or student theses associated with each aim are below (and can be found in the appendix):

Aim 1: Clarify how automation appearance, task type, and operator characteristics affect trust in automation

Publications:

- Pak, R., McLaughlin, A. C., & Bass, B. (2014). A Multi-level Analysis of the Effects of Age and Gender Stereotypes on Trust in Anthropomorphic Technology by Younger and Older Adults. *Ergonomics*.
- Rovira, E., Pak, R., & McLaughlin, A. C. (under review). Low Memory, Mo' Problems: Effects of individual differences on types and levels of automation. *Human Factors*.

Conference Proceedings

- Bass, B. M., Goodwin, M., Brennan, K., Pak, R., & McLaughlin, A. C. (2013). Effects of age and gender stereotypes on trust in an anthropomorphic decision aid. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 57(1), 1575-1579.
- Leidheiser, W., & Pak, R. (2014). The Effects of Age and Working Memory Demands on Automation-Induced Complacency. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58(1), 1919–1923. doi:10.1177/1541931214581401

Student Thesis:

- Leidheiser, W. (in progress). The Effects of Age and Working Memory Demands on Automation-Induced Complacency.

Aim 2: Determine if emotional expression can assist in optimal human-automation calibration

Student Thesis:

- Bass, B. (2014). Faces as Ambient Displays: Assessing the Attention-Demanding Characteristics of Facial Expressions. Unpublished master's thesis. Available at: http://tigerprints.clemson.edu/all_theses/1941/

Conference Proceedings:

- Bass, B. M., & Pak, R. (2012). Faces as Ambient Displays: Assessing the attention-demanding characteristics of facial expressions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56(1), 2142–2146.

Aim 3: Investigate how anthropomorphically designed automation affects automation error attributions

Student Thesis:

- Branyon, J. (in progress). Investigating older adults' trust, causal attributions, and perception of capabilities in robots as a function of robot appearance, task, and reliability.

Conference Poster:

- Branyon, J. J., & Pak, R. (2015). Investigating older adults' trust, attributions, and capability perceptions of robots. Presented at the American Psychological Association 123rd Annual Meeting. Toronto, ON: American Psychological Association

This report is organized around the three aims. In the course of this project, research efforts toward Aim 1 were expanded to include the influence of individual differences (study 2). Aim 3 was modified to examine the research question in the context of human-robot interaction.

Table of Contents

Aim 1: Clarify how automation appearance, task type, and operator characteristics affect trust in automation 4

Study 1: The effect of agent age and gender on trust in anthropomorphic automation in younger and older adults4

 Executive summary4

 Introduction4

 Methods, Procedure, and Results5

 Conclusion.....5

Study 2: The effect of individual differences in working memory on trust and performance with automation of varying degrees8

 Executive summary8

 Introduction8

 Methods, Procedure, and Results10

 Conclusion.....10

Aim 2: Determine if emotional expression can assist in optimal human-automation calibration..... 12

Study 1: Faces as ambient displays12

 Executive summary12

 Introduction12

 Methods, Procedure, and Results18

 Conclusion.....18

Aim 3: Investigate how anthropomorphically designed automation affects automation error attributions 19

Study 1: Investigating Older Adults’ Trust, Causal Attributions, and Perception of Capabilities in Robots as a Function of Robot Appearance, Task, and Reliability19

 Executive Summary19

 Introduction19

Methods, Procedures, Results23

Conclusion23

References..... 24

Appendix..... 33

Aim 1: Clarify how automation appearance, task type, and operator characteristics affect trust in automation

Study 1: The effect of agent age and gender on trust in anthropomorphic automation in younger and older adults

Executive summary

Previous research has shown that gender stereotypes, elicited by the appearance of the anthropomorphic technology, can alter perceptions of system reliability. The current study examined whether stereotypes about the perceived age and gender of anthropomorphic technology interacted with reliability to affect trust in such technology. Participants included a cross-section of younger and older adults. Through a factorial survey, participants responded to health-related vignettes containing anthropomorphic technology with a specific age, gender, and level of past reliability by rating their trust in the system. Trust in the technology was affected by the age and gender of the user as well as its appearance and reliability. Perceptions of anthropomorphic technology can be affected by pre-existing stereotypes about the capability of a specific age or gender.

Introduction

Interactive computer systems that exhibit human-like, or anthropomorphic, traits can lead users to perceive and treat them differently than non-human-like systems (Nass, Steuer, & Tauber, 1994). Thus it is imperative to understand how users' perceptions of the system might be affected by their social reactions to anthropomorphic technology. One way in which a system may elicit social reactions is by eliciting stereotypes (Yee, Bailenson, & Rickerson, 2007).

Stereotypes are preconceptions about the traits, behaviour, or abilities of a group and can set expectations of a stereotyped individual. Stereotypes can have both negative and positive connotations that may be inconsistent with real group attributes but provide adaptive value because they filter and organize incoming information, thereby easing processing and interpretation (Hilton & von Hippel, 1996). Stereotypes can be activated and applied with or without conscious awareness (Greenwald & Banaji, 1995; Banaji, Hardin, & Rothman, 1993). Unfortunately, when the stereotype is highly simplified or inaccurate, it can lead to errors in perceptions and behavior.

Stereotype activation for computerized agents can also interact with individual differences, such as physical characteristics. Qiu and Benbasat (2009) found that an anthropomorphic decision aid significantly increased perceptions of social presence and led to increased trust of the agent. The strength of these effects was influenced by the degree to which the decision aid agent was similar to the user on a visible factor, such as ethnicity. The link between trust and apparent physical characteristics was explained via similarity-attraction theory that predicted that people would be more attracted to those similar to them (Byrne, 1971). The user may have attributed their attraction to a similar ethnicity as trustworthiness of the agent.

In another example of the moderating role of individual differences in susceptibility to anthropomorphic effects, susceptibility to flattery (insincere praise) depended on the level of

computer experience of the user (Johnson, Gardner, & Wiles, 2004). Johnson, Gardner, and Wiles found that susceptibility to flattery from a computer depended on the user's experience level with computers – the judgments of highly experienced users were more affected by flattery than less experienced users. Further, Lee (2010) found that people who exhibited less analytical and more intuitive cognitive style were more susceptible to flattery from a computer.

In sum, stereotypes can affect user perceptions of a computer or automated aid and can be moderated by individual differences. Some of the aids described in the previous studies were forms of automation that functioned in a decision-support capacity; thus some automation bias may be based on stereotypes (Skitka, Mosier, & Burdick, 1999). However, no research has explicitly examined how these factors might interact with machine-related factors of automation, such as reliability of the automation, or how various activated stereotypes might interact (e.g., age and gender).

Using participants in younger and older adult age groups, we collected judgments of trust of a simulated agent embedded within a decision aid that varied in gender, age, and reliability using a factorial survey with concrete health-related vignettes. Following the social cognition literature, we expected that age and gender stereotypes would most affect trust in the decision aid when system performance was ambiguous, but that there would be different effects for different age groups and genders of users. Specific research aims were: 1) Determine the amount of variance in trust due to within-person variation compared to between-person variation, 2) Determine how age of the agent, gender of the agent, and reliability of the decision aid agent affected judgments of trust in the aid, and 3) Determine how individual differences such as age and gender of the participant affected trust ratings of various decision aids. The results informed basic knowledge of how differing age and gender groups responded to stereotypes as well as informing the design of decision aids targeting particular groups of users.

We presented scenarios involving a decision aid (a smartphone “app”) for diabetes management via a factorial survey. The decision aid contained a simulated anthropomorphized agent. Factorial surveys have been widely used to examine how beliefs, judgments, and decision-making are influenced by situational factors (Rossi & Anderson, 1982). Specific factors of the scenario were manipulated (in a factorial manner) and the participant rated all combinations of factors. The agent was a health care provider offering advice on a specific diabetes-related dilemma. Because our dependent variable (trust) was a social judgment about a situation, a factorial survey was an ideal way to measure the influence of manipulated variables (age, gender, reliability of automation) as well as individual differences of the participants (Rossi & Anderson, 1982; Hox, Kreft, & Hermkens, 1991).

Methods, Procedure, and Results

[can be found in Pak, McLaughlin, & Bass (2014) attached in Appendix]

Conclusion

As automation in consumer products and systems embodies human-like traits (e.g., anthropomorphic agents), stereotypes that users hold of age and gender may play an important

role in trust and use of that automation. Prior research established that people apply gender stereotypes to computers but the purpose of this study was to examine if powerful and pervasive age stereotypes, as well as gender stereotypes, would be applied to anthropomorphic agents.

The finding that trust varies with reliability is not surprising; with higher levels of perceived reliability, users, particularly older adults, may become complacent (Mouloua, Smither, Vincenzi, & Smith, 2002; Ho, Wheatley, & Scialfa, 2005). What is surprising is that this relationship between trust and complacency interacts with attributes of technology and individual differences in a way that is roughly consistent with the stereotype literature, specifically, age and gender stereotypes of doctors. However, perceived age group and gender of the agent and its reliability moderated the application of stereotypes. When the agent appeared young, male agents were more trusted than female agents only when reliability was low. This gender difference disappeared at other levels of reliability. This pattern might suggest that unless the reliability of the system is catastrophically low (45%), most participants do not exhibit gender stereotypic thinking; perceptions of trust are primarily driven by reliability. However, when the reliability is very low, participants clearly shift to more stereotypic thinking and seem to attribute low performance to gender.

When the agent appeared older, male agents were more trusted than female agents only at medium levels of reliability. That is, stereotypic judgments appear at more moderate levels of reliability (70% versus 45%) if the aid is older rather than younger. The finding of gender stereotypic effects at 45% reliability when the agent is young but at 70% when the agent is old seems to suggest that older female agents are judged more harshly than younger female agents. Giving this finding one design recommendation is that when it is crucial for users to maintain high levels of trust in imperfect automation, a younger male agent is optimal because it seems less susceptible to large fluctuations in perceptions of trust as a function of gender (i.e., gender stereotypic thinking). More specifically, if it is undesirable to have users exhibit gender differences (or bias) in trust then using younger agents was preferable to older agents. A male agent was recommended over female because trust in female agents appeared more erratic as a function of reliability compared to male agents (e.g., the steep plunge in trust at 45% reliability for young females). However, this design recommendation does not take into account the gender or age group of the user as our results showed that individual differences also seem to interact with the agent characteristics.

Some anthropomorphic aspects of the aid did interact with participant individual differences to affect trust. Younger adults in low reliability conditions tended to trust older agents over younger agents while older adults did not show any significant differences in trust as a function of agent age. Based on Model 3, if the goal is to maintain high levels of trust in imperfect automation in young adult users, older agents (regardless of agent gender) are preferred. For older adult users, there was no significant difference in trust as a function of agent age group. However, there did appear to be a trend toward higher trust of younger agents with increasing reliability so for older users, a young agent may be optimal.

One caveat is that we did not assess a priori the pre-existing stereotypes held by our participants (as such an assessment might have influenced their behavior in the experiment.) However, the stereotype literature is replete with research that shows the pervasiveness of the "warm but not

competent" stereotype of older adults not only in the United States but worldwide (Cuddy, Norton, & Fiske, 2005). Another limitation is the use of a diabetes scenario. Although none of the participants in our study reported having diabetes, older adults may be more aware of diabetes simply because it is more common in their cohort than among younger adults (26.9% versus 11.3% respectively; American Diabetes Association, 2011). Thus, simply being in a cohort that is more affected by diabetes may influence how one perceives diabetes advice. Another limitation was that because we assessed subjective perceptions of the automation (trust) because it is uncertain if trust translates to behavior. However past research has shown that perceptions of trust in automation are strongly correlated with behavior (e.g., Lee & Moray, 1994).

Study 2: The effect of individual differences in working memory on trust and performance with automation of varying degrees

Executive summary

We explored the extent to which individual differences in cognitive ability affected the use of types and levels of automation support in a complex decision-making task. Previous studies show performance benefits with reliable automation but performance costs with imperfect automation, particularly as automation support increases. Cognitive abilities are also critical to decision-making and correlate with automation reliance. We examined decision-making performance with varying types and levels of imperfect automation that supported 86 participants performing a simulated command and control task. Participants also completed a spatial working memory task. Reliable automation with increased automation support resulted in higher accuracies. When automation failed, the reverse was true: increased automation support resulted in lower accuracy, especially for those with lower working memory ability. Those with higher working memory were less susceptible to the detrimental effects when seemingly supportive automation failed. Further, lower working memory was associated with more trust in automation. These results confirm the link between automation performance and individual differences, but also demonstrate the limits of the “conventional wisdom” that higher, reliable automation support unilaterally helps performance while higher, imperfect automation support harms performance (cf. Onnasch, Wickens, Li, & Manzey, 2013).

Introduction

A growing body of research has examined how human performance is differentially affected by various types and levels of highly reliable but imperfect automation (Crocoll & Coury, 1990; Endsley & Kaber, 1999; Galster, Bolia, & Parasuraman, 2002; Lorenz, Di Nocera, Röttger, & Parasuraman, 2002; Sarter & Schroeder, 2001; Wickens & Xu, 2002; Rovira, McGarry, & Parasuraman, 2007; Onnasch, Wickens, Li, & Manzey, 2014). The interest is motivated by the severe human performance consequences of highly reliable, yet imperfect automation such as: out of the loop unfamiliarity (Wickens, 1992), automation complacency (Parasuraman, Molloy, & Singh, 1993), loss of situation awareness (Endsley & Kiris, 1995), and skill degradation (Bainbridge, 1983).

In a meta-analysis of 18 automation studies examining the differential effects of types and levels of automation, Onnasch et al. (2014) found performance benefits for reliable automation and performance decrements after an automation failure with decision automation and increased levels of automation. Of most interest were the decrements in performance found when automation support moved across the critical boundary from information automation to decision automation; a change in type of automation. Thus, an important goal for designers is to mitigate performance costs associated with failures of decision automation and failures at increased levels of automation by facilitating appropriate trust calibration (e.g., Rovira, Cross, Leitch, & Bonaceto, 2014). One approach is to better understand how individual differences in cognitive ability affect the appropriate use of imperfect types and levels of automation in complex decision-making tasks.

In study examining automation performance and individual differences in cognitive abilities, Chen and Terrence (2009) investigated the effects of imperfect automation and individual differences in a military multitask environment. Specifically, they were interested if individual differences in a component of working memory capacity, perceived attentional control (Shipstead, Lindsey, Marshall, & Engle, 2014), impacted how operators interacted with miss versus false alarm prone automation. Attentional control was assessed using a subjective measure of individuals' perceived attentional focus and shifting. They found that individuals 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.

Individual differences in working memory also seem to play a role in mediating operator performance with automation. Parasuraman, de Visser, Lin, and Greenwood (2012) examined whether certain genotypes could predict an individual's susceptibility to automation bias (adhering to imperfect automation). Researchers looked at two specific single nucleotide polymorphisms (SNPs) or variants of the DBH gene that regulate Dopamine (DA) and norepinephrine (NE). DA and NE levels are associated with DBH enzyme activity (low, high) that contributes to neural activity in the prefrontal cortex known to play a critical role in working memory ability. Using a command and control task (Rovira, et al., 2007), Parasuraman et al. (2012) varied the automation support (manual, reliable, and automation failure) that low and high DBH enzyme groups experienced. They found no difference between the low and high DBH enzyme groups with manual and reliable automation, but with automation failures individuals in the low DBH enzyme group performed better compared to individuals in the high DBH enzyme group. Parasuraman et al. (2012) attributed this effect to individual differences in working memory induced by enhanced DA availability in the low DBH enzyme group. However, because they did not measure working memory or other cognitive abilities, it is still unclear if individual differences in working memory interact with automation reliability to affect performance.

The importance of individual differences in working memory was examined in another study (de Visser, Shaw, Mohamed-Ameen, & Parasuraman, 2010). Researchers investigated the role of working memory in an automated UAV task by varying task load (low, high) and automation reliability (manual, reliable, and automation failure). Participants completed both the Operation Span (OPAN) and Spatial Span (SSPAN) working memory tests (Engle, 2002). Researchers found a significant correlation with OPSAN scores and performance on the automated task. For each automation task performance measure, they found that linear models that included working memory accounted for more of the variance in performance as compared to the linear models without the individual differences OPSAN measure. Thus, when individual differences in working memory are accounted for, more variation in performance with automation can be explained. Critically, however, this study did not vary in types or levels of automation.

The current research was aimed at understanding the sources of performance differences underlying human-automation interaction with imperfect automation across different types and levels of automation (for a review see Onnasch et al, 2014) as it specifically relates to individual differences. First, we varied types and levels of imperfect automation and task load. Second, we

measured individual differences in working memory ability by using a performance-based working memory task compared to self-reported measures of abilities, complex proxy tasks (e.g., video game performance), or genetic predictors of cognitive performance. Finally, we systematically varied primary task demand: evidence from a review of 20 automation reliability studies suggested that dependence on imperfect automation would be stronger with increased task demand (because the operator's limited resources are expended; Wickens & Dixon, 2007). We hypothesized that individual differences in working memory would differentially impact reliance on varying types and levels of automation. Specifically:

- 1) First, consistent with previous literature, we hypothesized that:
 - a) operators would perform better with reliable automation compared to manual control.
 - b) there would be no difference between task load conditions when the automation was reliable.
 - c) the differential impact of information versus decision automation would be evident with automation failures, especially when task load was high.
- 2) Second, as suggested by Parasuraman et al. (2012), we expected individuals with higher working memory ability to show less of a decrement when formerly supportive automation failed compared to individuals with lower working memory ability. Specifically, with automation failures, high task load, and increasing automation support it was predicted that the benefits of better spatial working memory ability would be highlighted.
- 3) Third, we expected a relationship between variations in cognitive ability and self-report measures of trust. Specifically, individuals with lower working memory abilities would trust the automation more compared to individuals with higher spatial working memory abilities because individuals with lower working memory abilities would need to rely on the automation more than those with higher working memory abilities.

Methods, Procedure, and Results

[can be found in Rovira, Pak, & McLaughlin, (under review), attached in appendix]

Conclusion

The extent to which automation enhances decision-making depends on individual differences in cognitive ability. 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 reliable automation depended on individual differences in working memory. This finding may help optimize human-automation interaction. Further, our findings that working memory ability is related to trust in automation suggest more work should consider this individual difference. Our study replicated prior research that operators would perform better with reliable automation compared to manual control (Hypothesis 1a). In addition, task load did not differentiate performance when the automation was reliable (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 automation (information and low-decision) and higher automation (medium-decision). It appears that lower automation can mitigate some of the performance penalty of

increased task load when automation fails while performance significantly declines with automation failures and higher types and levels of automation. The drop in decision accuracy with increased task load occurs 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 & Xu, 2002).

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

When automation failed, all participants' accuracies declined as the type and level of automation increased. However, those with lower working memory were more severely impacted by automation failures than those with higher working memory. Taken together, these results confirmed hypothesis 2 regarding the effects of type and level of automation and working memory. These results also added detail to the conventional wisdom that increasing automation type or level benefits performance but can lead to catastrophic performance when automation fails (i.e., the lumberjack effect; Onnasch et al., 2014). When automation support was low but reliable, those with higher working memory outperformed those with lower working memory, and when automation failed, those with lower working memory suffered more than those with higher 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, 2012), but also extends the literature by further specifying the automation conditions (type and level of automation support and trial reliability) under which working memory affects performance.

Finally, hypothesis 3 which predicted a relationship between working memory and trust in automation was supported. We found that working memory was weakly but significantly negatively correlated to measures of trust. Specifically, individuals with higher working memory ability had lower trust, reliance, and lower beliefs that automation would improve their performance.

Aim 2: Determine if emotional expression can assist in optimal human-automation calibration

Study 1: Faces as ambient displays

Executive summary

Ambient displays are used to provide information to users in a non-distracting manner. The purpose of this research was to examine the efficacy of facial expressions as a method of conveying information to users in an unobtrusive way. Specifically, the current study assessed the attention-demanding characteristics of facial expressions using the dual-task experiment paradigm. Results from the experiment suggest that Chernoff facial expressions are decoded with the most accuracy when happy facial expressions are used. There was also an age-effect on decoding accuracy; indicating younger adults had higher facial expression decoding performance compared to older adults. The observed decoding advantages for happy facial expressions and younger adults in the single-task were maintained in the dual-task. The dual-task paradigm revealed that the decoding of Chernoff facial expressions required more attention (i.e., longer response times and more face misses) than hypothesized, and did not evoke attention-free decoding. Chernoff facial expressions do not appear to be good ambient displays due to their attentional demanding nature.

Introduction

Ambient displays can take many forms. For example, the battery meter icon of a computer interface, or a dangling string from the ceiling to represent network traffic on a computer network (Weiser & Brown, 1995). These examples are considered “ambient” because they convey information to the user without being substantially taxing on cognitive faculties (i.e., they are in the background and do not require the user to change focus or switch attention). Several important characteristics have been identified for the design of a good ambient display. Examples of these characteristics include: providing useful and relevant information, having a sufficient information design, using consistent and intuitive mapping, and appropriate matching between the system and the real world (Mankoff et al., 2003). If these characteristics are adequately fulfilled by facial expressions, then facial expressions could be considered a good form of ambient display. The purpose of this study is to determine if face stimuli can serve as ambient indicators of quantitative information.

One situation where ambient displays may be helpful is in human-automation interaction (HAI). In some HAIs, users may become unaware of the hidden decision making processes or outcomes of automation. They may also lose track of the automation’s reliability over time (i.e., forget how reliable or unreliable it has been in the past). Such information (uncertainty of current processes, past reliability) can lead to fluctuations in trust that may not be justified (un-calibrated trust); that is trust that may be unwarranted. Un-calibrated trust can manifest itself as continued use of unreliable automation (misuse) or unwarranted discontinued use of reliable automation (disuse) both of which cause non-optimal HAIs (Parasuraman, 1997).

One way in which an automated system can encourage proper calibration is by presenting as much information about its operation as possible. For example, it could present its own confidence in its recommendation, so called “system confidence”, or it could present a historical picture of its own reliability (both are information that are easily accessible by a system). This concept can be categorized in the ambient display heuristic of useful and relevant information. For example, if the system is working from faulty data, it will weight its advice as potentially unreliable. Presenting critical information, such as system confidence, is a way of diminishing the uncertainty that can exist in HAIs (Bubb-Lewis & Scerbo, 1997). Trust is a malleable variable that can be shaped through interactions with a system (Antifakos, Kern, Schiele, & Schwaninger, 2005).

If a system is presenting the operator with its system confidence level, then the operator will be able to build a more appropriate trust relationship with the automation. However, this presentation needs to be salient and the automation state indicator should not add attentional demands to the user (Parasuraman, 1997). Some previous research has indicated that methods such as tactile output and auditory output may be helpful in conveying system confidence (Wisneski, 1999; Poupyrev, Maruyama, & Rekimoto, 2002; Sawhney & Schmandt, 2000). While these modalities are novel in certain capacities, a less intrusive and less attention demanding modality would be more beneficial to users. Thus, the ideal stimulus display type would be one that provides the user with meaningful information, while not becoming a distraction or a drain on the user’s attention (Antifakos, Kern, Schiele, and Schwaninger, 2005). Coding information as emotional expression in human-like faces may fulfill this role.

Neuroimaging studies have supported the notion that the emotional processing of faces is a more effective pathway than the processing of other stimuli. A previous study compared the automatic processing of emotional facial expressions versus emotional words. Rellecke (2011) hypothesized that facial expressions would be encoded more automatically than words, due to their perceptual features and humans’ natural ability to encode them. This study was novel because it took two theoretically attention-free emotional processing stimuli (i.e., faces and words), and compared their efficiency and effect. The degree of encoding automaticity was being tested for each of these stimuli. Based on the results of the electroencephalogram (EEG), the event-related brain potentials (ERPs) recorded for the facial expression conditions were found to have a prolonged effect on the brain.

This finding alludes to emotional facial expression processing as being automated to a higher extent than emotional word processing. Rellecke (2011) discusses the potential necessity for preconditions for the high automatic processing of emotional words. This was apparent because the two stimuli were tested in the same superficial stimulus analysis task, but only one (i.e., facial expression) led to advanced pre-attentive processing. Facial expression seems to be a stimulus that needs no prompting or preconditions to allow fast, but also meaningful processing (Rellecke, 2011). Data analysis found that happy faces were decoded earlier than other faces (i.e., 50-100 ms).

This supports the theory that happy faces are advantageous in the early stages of emotional processing and may be instrumental in attention-free encoding. Also, data showed that angry faces were advantageous for later decoding (i.e., 150-450 ms). This coincides with previous

research that states angry expressions, or threat-related expressions, have prolonged effects on the brain (Rellecke, 2011). These differences in emotion type on ERPs show that there may be a specific type of emotion that elicits faster decoding for humans.

Chernoff Faces

Chernoff faces were created as a way to represent multivariate data in a way that would allow the viewer to gain information in a quick, yet complete manner. For example, some of the original Chernoff faces were used to represent fossil data. The Chernoff faces displayed information pertinent to the fossils (i.e., inner diameter of embryonic chamber, total number of whorls, maximum height of chambers in last whorl, etc.) through variations including, but not limited to the faces: head shape, eye size, mouth size/shape, and eyebrow size/slant. Chernoff's rationale was that due to the extreme familiarity of faces, people would easily detect differences in the configuration of a face, even if the differences were small ones (Chernoff, 1973). It was expected that people would at least be able to examine faces more quickly than examining a row of numbers. Assuming that this is true, a schematic facial expression should act as a superb source of information output.

Chernoff faces have up to 18 characteristics that can be manipulated (Nelson, 2007). When representing multivariate data (e.g., the fossil data) it is beneficial to have multiple facial elements that can be manipulated and used for representing various data. However, when representing univariate data (i.e., a single percentage score) it seems that having a lower number of manipulated facial features is more beneficial. Therefore, it could be problematic to have several individual facial elements for the human to properly decode. As Montello and Gray (2005) state, it is more beneficial to have a stimulus that communicates information univariately rather than multivariately when the goal is to give the user a single quantity. A pseudo-Chernoff face may be a remedy for this dilemma (Montello & Gray, 2005). This "pseudo-Chernoff" face could be created by systematically manipulating one facial characteristic, while holding all others constant. To properly convey a simple quantitative score the Chernoff face may only need to have one facial characteristic manipulated. Through this manipulation, the human may be more apt to decode the Chernoff face accurately and quickly, while noticing subtle changes (Kabulov, 1992).

The issue of whether interpreting Chernoff faces is a relatively less attention-demanding task is of primary importance to the current study. Previous studies have investigated the effectiveness of Chernoff faces as a pre-attentive stimulus with mixed results. A study concluded that Chernoff faces are not processed pre-attentively, and do not benefit users more than other modes of visual information display (Morris, Ebert, & Rheingans, 2000). The process of identifying the characteristics (eyebrow slant, eye size, nose length) of the Chernoff face was said to be a serial process. Participants' accuracy of target stimuli identification improved when they were given more time and less distracters, indicating that the task was not pre-attentive (Morris, Ebert, & Rheingans, 2000). A similar study investigated data visualization and used Chernoff faces as one of the "glyph stimuli" to discover which data visualizations were the most effective (Lee, Reilly, & Butavicius, 2003). Glyphs are data visualizations that are characterized by their attempt to display multivariate data through the manipulation of features on the glyph that correspond to raw data. It was found that participants had lower accuracy scores and took longer to answer

questions when exposed to the glyph stimuli (Lee, Reilly, & Butavicius, 2003). This indicates a serial processing of information from the Chernoff faces, which is in agreement with the findings of Morris, Ebert, & Rheingans (2000).

Age-Related and Cultural Effects on Decoding

Despite the ease with which humans are able to decode emotional facial expressions, it is still moderated by age. Age can alter a person's ability to correctly perceive and understand the facial expression that is presented to them. Neuropsychological research has shown that age-related issues in facial expression decoding may be a result of problems with the medial temporal lobe (Orgeta & Phillips, 2007). The amygdala is housed here, which corroborates with previous research that suggests the amygdala is necessary for facial expression decoding (Whalen, 1998; Morris, 1998). Despite these age-related issues; a competing theory has been asserted regarding older adults' ability to decode emotional facial expressions. The socioemotional selectivity theory asserts that social behavior is essentially a byproduct of time (Carstensen, Issacowitz, & Charles, 1999). In a sense, time can be thought of as the chronological age of a human. As the human ages, they essentially have less time to live and fulfill goals. This affects the way they view their decisions and weight their goals. The two types of goals that make up the socioemotional selectivity theory are knowledge-based and emotion-based goals (Carstensen, Issacowitz, & Charles, 1999). Younger adults are more likely to pursue knowledge-based goals because they have more time potential. The trade off for knowledge in lieu of emotional goals appears to be a worthy endeavor. Older adults supposedly take the opposite approach and view emotional-based goals as top priority. Older adults' view time as a non-renewable resource, and seek to spend anytime they have left enjoying positive emotional experiences (Carstensen, Issacowitz, & Charles, 1999).

According to the socioemotional selectivity theory, older adults may actually be more aware of certain emotional situations and images than non-emotional (Orgeta & Phillips, 2007). Orgeta and Phillips (2007) showed older adults as being more accurate at identifying positive facial expressions, opposed to negative facial expressions. Older adults were found to identify positive emotions as accurately as younger adults. There was no significant difference between the older adults and younger adults in terms of identifying positive facial emotions (i.e., happiness and surprise). However, older adults were significantly worse than younger adults at identifying negative facial emotions (i.e., sadness, anger, and fear). The results of this study indicated that there is an age-related difference for the decoding of negative facial expressions, but not positive facial expressions (Orgeta & Phillips, 2007). The ease of recognition for certain emotional expressions versus others is an area that is pertinent to this research area. As Orgeta and Phillips (2007) showed, older adults may have a positivity bias that allows them to overcome any cognitive decrements that interrupt other emotional decoding, thus decoding positive facial expressions as accurately as younger adults. Other research has supporting data showing that positive expressions (e.g., happiness) are processed more quickly, supported by faster N170 latencies (Batty & Taylor, 2003). Perhaps this quick processing attributes to the robustness of the happy facial expression compared to other expressions.

A previous study manipulated the factors of chronological age and the participant's working self-concept to determine if the positivity effect could in fact be evoked in younger adults, and

likewise the negativity effect in older adults (Lynchard & Radvansky, 2012). During the experiment the participant would complete a possible selves orienting task. The older adults completed the younger possible selves orienting task, while the younger adults completed the older possible selves orienting task. Essentially, this made the participant's working self-concept the opposite of their chronological age. The results showed that there was a reversal of stereotypical age-related emotional information processing. Younger adults displayed a positivity effect, which is thought to be a unique attribute of older adults. Similarly, older adults displayed a negativity effect, which is thought to be unique to younger adults (Lynchard & Radvansky, 2012). This study showed that more than just chronological age plays a role in the socioemotional selectivity theory. Humans are subject to emotional information processing biases based on less concrete variables such as their working self-concept.

Decoding facial expressions is a cross-cultural behavior that is a critical part of human life. There are six basic emotions that transcend culture. These are: anger, happiness, fear, surprise, disgust, and sadness (Ekman & Friesen, 1975). These emotions can be represented with facial expressions (Lee, 2006; Batty, 2003). Because these facial expressions are not confined to specific cultures, it puts no restraints on the ability of different people groups to successfully decode these facial expressions. It appears that increasing age is a factor that may cause differences in aspects of facial expression decoding, while cultural background seems to be of no hindrance. The unique quality that facial expressions have in their prevalence and familiarity in human culture makes them a good candidate for an ambient display. This quality of facial expressions allows the heuristic of matching the system to the real world to be met.

Limitations of Previous Literature

The previous literature has provided a foundation for knowledge about facial expressions, but there are limitations to these studies. The Hess (1997) study presented emotional facial expressions in a single-task format. The participants viewed the image and rated it on the emotionality and intensity that they perceived. This methodology does not clarify whether facial emotion decoding is truly resource/attention-free as neuropsychological studies suggest. A dual-task experiment should be implemented to properly measure attention usage. In order to gain this data; measures of response time, accuracy, and subjective workload should be used. The Hess (1997) study also measured decoding accuracy for each facial expression image through the presentation of several emotion scales at once. The participant was presented with seven emotional labels, which they manipulated to show the intensity of emotion for the previous picture. Instead of presenting seven individual scales, it seems to be less complicated to present one scale or to have a quick input device (e.g., keyboard number keys) after the image is viewed.

The Hess (1997) study presented facial expression intensity in increments of 20 % intensity. This intensity scale may not provide enough precision or a complete spectrum of facial expression decoding data. The Orgeta and Phillips (2007) study also presented only four intensity levels. The number of intensity levels may need to be increased (i.e., create smaller increments of percentage changes between each stimuli) to capture a more accurate representation of participants' ability to decode facial expression. Another limitation in the Orgeta and Phillips (2007) study was the facial images were presented in increasing order as the participant advanced through the experiment. This method may have led to participants forming an anticipation bias that the next facial image was going to be more expressive.

Previous research has also provided evidence that age-related effects may cause differences in the ability for humans to properly decode facial expressions. It has been shown that older adults are worse at identifying negative facial expressions (i.e., sadness, anger, and fear). Older adults struggled significantly versus younger adults in properly recognizing the negative emotions at intensity levels of 50 %, 75 %, and 100 %. It appears that older adults have a higher recognition threshold for certain negative emotions than younger adults. Basically, older adults do not pick up on negative facial stimuli as easily as younger adults and need more intense facial expressions to determine the appropriate emotional state (Orgeta & Phillips, 2007). In order to determine if theories such as the socioemotional selectivity theory pertain to Chernoff face recognition, there needs to be an independent variable of age with levels of younger and older adults.

The variable of gender of the facial expression stimuli could be considered a confounding variable. Hess (1997) used two male and two female actors to create facial expressions for their study. Results of this study showed that the gender of the stimuli (i.e., actors) did influence participant rating accuracy. For the expressions of happy and sad, there was an interaction of the gender of the stimuli x intensity of the expression (Hess, 1997). Because of this reported interaction, it would be beneficial to use non-gender specific stimuli to eliminate this confounding variable.

Previous studies have looked at users' ability to properly decode facial expression type (Ekman & Friesen, 1975), intensity (Tsurusawa, Goto, Mitsudome, Nakashima, & Tobimatsu, 2007; Hess 1997), and the effectiveness of Chernoff faces (Chernoff 1973; Tsurusawa, Goto, Mitsudome, Nakashima, & Tobimatsu, 2007; Morris, Ebert, & Rheingans, 2000). The purpose of the current study is to examine the users' ability to accurately decode a quantitative value from Chernoff facial expressions.

Overview of the Study

In order to determine the attention usage by the participants, a dual-task methodology was used. Our study used the dual-task paradigm to measure the attention-demanding characteristics of facial displays. The Hess (1997) study measured participant's decoding accuracy with several scales after each trial. This method may create confusion for the participant, and not accurately record participant decoding time. The interface should allow for quick and simple input of the facial expression intensity from the participant. The current study used only one measurement scale (direct key entry) after each trial to eliminate any confusion for the participants about what the scales are measuring and give a better approximation about how quickly the participant can decode the facial expression. In the Orgeta and Phillips (2007) study the facial expressions were shown in increasing order. This technique was not replicated in the current study. Instead, a randomized sequence of facial expression stimuli was used to control for any biases that could be formed due to participant expectations. The Chernoff face stimuli were manipulated differently compared to previous research (Chernoff, 1973; Tsurusawa, Goto, Mitsudome, Nakashima, & Tobimatsu, 2007; Morris, Ebert, & Rheingans, 2000). Only the mouth was manipulated in order to gain understanding about the affect of this one variable on decoding. Finally, the current study used a more precise facial expression intensity scale than previous research (Hess, 1997; Orgeta & Phillips, 2007). To accomplish this, a facial expression scale presenting emotions in

increments of 10 % was used. Our assumption was that by making these modifications the current study would be able to address the research question with more accuracy.

Methods, Procedure, and Results

[can be found in Bass, 2014, attached in appendix]

Conclusion

The goal of the study was to investigate whether Chernoff face stimuli could serve as ambient (i.e., relatively resource-free) indicators of quantitative information, using a dual-task paradigm. In general, we hypothesized that sad face emotion decoding would show age-related differences but happy faces would be immune to age-related differences. This was based on the literature indicating positive facial expressions provided a decoding advantage (i.e., are more easily decoded; Bartneck & Reichenbach, 2005; Calvo & Lundqvist, 2008; Rellecke, 2011), and the finding that older adults could decode positive facial expressions as accurately as younger adults (Orgeta & Phillips, 2007). However, we found that the relationship between younger and older adults' decoding accuracy did not significantly change due to facial expression condition. Therefore, there was an age-related difference in decoding accuracy in the happy face condition.

However, when collapsing across age group, participants had higher decoding accuracy when they were presented with happy facial expressions. This finding supports a general "happy face advantage" and suggests that when compared to sad Chernoff facial expressions, happy Chernoff facial expressions are more advantageous for decoding. In terms of using a Chernoff face for the display of quantitative information; the use of happy facial expression was shown to be an overall more decodable stimuli. This finding corroborates previous research that show evidence of more accurate happy face decoding (Hess, 1997).

Aim 3: Investigate how anthropomorphically designed automation affects automation error attributions

Study 1: Investigating Older Adults' Trust, Causal Attributions, and Perception of Capabilities in Robots as a Function of Robot Appearance, Task, and Reliability

Executive Summary

The purpose of the current study is to examine the extent to which the appearance, task, and reliability of a robot is susceptible to stereotypic thinking. Stereotypes can influence the types of causal attributions that people make about the performance of others. Just as causal attributions may affect an individual's perception of other people, it may similarly affect perceptions of technology. Stereotypes can also influence perceived capabilities of others. That is, in situations where stereotypes are activated, an individual's perceived capabilities are typically diminished. The tendency to adjust perceptions of capabilities of others may translate into levels of trust placed in the individual's abilities. A cross-sectional factorial survey using video vignettes will be utilized to assess young adults' and older adults' attitudes toward a robot's behavior and appearance. We hypothesize that a robot's older appearance will result in lower levels of trust, more dispositional attributions, and lower perceptions of capabilities while high reliability should positively impact trust.

Introduction

When interacting with technology, people focus on human-like qualities of the technology more than the asocial nature of the interaction (Reeves & Nass, 1996; Nass & Moon, 2000) attributing human-like qualities such as personality, mindfulness, and social characteristics. The attribution of human-like qualities makes technology susceptible to stereotyping based on appearance and etiquette (e.g., Nass & Lee, 2001; Parasuraman & Miller, 2004; Eyssel & Kuchenbrandt, 2012). For example, when a male or female anthropomorphic computerized aid was included in a trivia task, participants were more likely to trust the male aid's suggestions and ranked the female aid as less competent (Lee, 2008).

The purpose of the current study is to examine the extent to which the appearance, task, and reliability of a robot is susceptible to stereotypic thinking. The theoretical relevance is that the results of this study will inform the limits of stereotypic thinking by investigating whether stereotypes are applied to robots. The practical relevance is that the current study may inform the design of robots to enhance human-robot interaction, particularly for older adults who tend to be less accepting of technological aids than other age groups (Czaja et al., 2006).

Stereotypes and Aging

In order to make efficient social judgments about others, individuals rely on the use of heuristics. One example heuristic involves placing an individual into a pre-determined schema (i.e., a stereotype). Stereotypes are cognitive shortcuts that result in impressions of others (e.g., Ashmore & Del Boca, 1981). Therefore, older adults may be more likely than younger adults to

apply stereotypes when they do not have other sources of information available to them (i.e., under situations of ambiguity).

Stereotypes are more likely to be activated in domains that are inconsistent with prescriptive societal gender or age roles (e.g., Kuchenbrandt, Häring, Eichberg, Eyssel, & André, 2014). For example, individuals perceived a female-voiced computer to be more informative about romantic relationships than the male-voiced computer (Nass, Moon, & Green, 1997). Although gender stereotypes have been studied using anthropomorphic technological aid paradigms, aging stereotypes have been investigated to a lesser degree within this context. Pak, McLaughlin, & Bass (2014) examined whether the physical appearance of an anthropomorphic aid would activate stereotypic thinking and affect individuals' trust in the aid. Using a factorial design, Pak et al. manipulated the technological aid's gender and age (younger, older) as well as participants' perceptions of the reliability of the automation. Participants were told that the automation was either 45%, 70%, or 95% reliable. However, the automation always provided a correct answer during testing. The task in this study was a health behaviors test regarding participants' knowledge about diabetes. Before beginning the task, participants were told that the automated aid was a Smartphone application recommended by a doctor designed to help people make the best decisions about diabetes. As the participants answered each question, the decision aid smart phone app would appear on the screen and the agent would recommend a correct answer. All of the agents were dressed as doctors. Participants rated their subjective trust in the automation and whether they would actually use the advice of the application on a 1-7 Likert scale.

Pak, McLaughlin, & Bass (2014) found that both younger and older adult participants trusted the older anthropomorphic aids more than the younger aids, the male aids more than the female aids, and more reliable applications than less reliable applications. However, stereotypic thinking was activated when perceptions of reliability were low or ambiguous. When the app had low reliability, the younger female aid was trusted less than younger male agents. Also, under medium reliability, the older female aid was trusted less than the older male aid. These results suggest that trust in automation can be influenced by physical appearance (i.e., gender and perceived age) of the technology. These results also further support the notion that technology is, like humans, also susceptible to stereotyping.

Physical appearance is known to play a large role in the activation of aging stereotypes. The link between physical characteristics and stereotypes has been well established in the social cognition literature (Brewer & Lui, 1984; Hummert, 1994; Hummert, Garstka, & Shaner, 1997). Within this context, facial features are considered to be the main source of information used in order to activate stereotypes. Hummert et al. (1997) found that negative age stereotypes were associated with the perception of advanced age through facial photographs. Overall, these findings suggest that physical cues are major indicators within the context of social judgments.

Stereotypes about older adults, although pervasively negative, can be multidimensional in the right context. People hold both positive and negative stereotypes about older adults (Hummert, 1993). When adults of all ages completed a trait card-sorting task where they were asked to generate traits they associated with older adults, Hummert and colleagues (1994) found approximately 10 different aging stereotypes, including positive ones like the "golden ager" who leads an active and engaged lifestyle. Although many stereotypes are held in common by people

of all ages, aging stereotypes tend to become increasingly differentiated as people grow older (Hummert, 1993; Hummert et al., 1994).

Stereotypes and other social beliefs can influence the way in which individuals process information in order to form social judgments, including the types of causal attributions that people make about the performance of others (Fiske & Taylor, 1991). When trying to determine the causality of an event, people tend to use two types of information: internal or dispositional qualities of the individuals involved in an outcome and the influences of the situation itself (Gilbert, 1993; Krull, 1993; Krull & Erikson, 1995). Potential biases in the attribution process can occur as a function of the valence of the situational outcome, the degree of ambiguity of the situation (or of the information given about causal factors), and the controllability of the situation (Blanchard-Fields, 1994). Blanchard-Fields suggested that, in general, older adults are most likely to make dispositional attributions when the outcome of a situation was negative and the actor's role in the outcome was ambiguous. When personal beliefs about another individual or situation are violated, older adults are also more likely to make dispositional attributions of blame rather than situational (Blanchard-Fields, 1996; Blanchard-Fields, Hertzog, & Horhota, 2012). Just as causal attributions, or the extent to which behavior is attributed to situational or dispositional causes, may affect an individual's perception of other people, it may also similarly affect perceptions of technology. For example, blaming technology for unreliable performance is likely to induce less trust (Moray, Hiskes, Lee, and Muir, 1995; Madhavan, Wiegmann, & Lacson, 2006). Attribution of fault has been studied in the automation and has been referred to as automation bias (Mosier & Sitka, 1996). Automation bias has been defined "as a heuristic replacement for vigilant information seeking and processing" (Mosier & Sitka, p. 202) which results in increased omission errors and commission errors.

Expectations of performance outcomes are influenced by stereotypes. Adults of all ages expect memory performance to decline with age (Lineweaver and Hertzog, 1998). Similarly, older adults' abilities are perceived negatively in domains involving memory (Kite & Johnson, 1988; Kite, Stockdale, Whitley & Johnson, 2005) and physical well-being (Davis & Friedrich, 2010). In memory taxing situations, older adults are perceived as being less credible and less accurate (Muller-Johnson, Toglia, Sweeney, & Ceci, 2007). The tendency to adjust perceptions of capabilities of others based on appearance may translate into levels of trust placed in the individual's abilities.

Trust in Automation

Trust in technological agents is important because it affects an individual's willingness to accept robot's input, instructions, or suggestions (Lussier, Gallien, & Guiochet, 2007). For example, Muir and Moray (1996) found a strong positive relationship between adults' level of trust in an automated system and the extent to which they allocated control to the automated system. Interestingly, Muir (1987) suggests that people's trust in technology is affected by factors that are also the basis of interpersonal trust. Trust in automation is thought to develop overtime (Maes, 1994) suggesting that trust is influenced by past experiences with the technology. For example, Merritt and Ilgen (2008) describe dispositional trust as the trust placed in a person or automation during a first encounter before any interaction has been made while history based trust reflects the prior experience a person has with another person or automation.

Performance based factors have a large influence in perceived trust in HRI (Brule, Dotsch, Bijlstra, Wigboldus, & Haselager, 2014). In fact, a recent meta-analysis suggests that a robot's task performance was the most important factor in adults' trust in robots (Hancock et al., 2011). That is, if the robot performs reliably, the human will exhibit greater trust towards the robot. The same meta-analysis found that behavior, proximity, and size of the robot also affect trust to a lesser extent. However, human-automation trust literature suggests that appearance can have reliable effects on trust (Pak Fink, Price, Bass, & Sturre, 2012). Indeed, studies in the social literature have found that people often judge an individual's levels of trustworthiness based on facial appearance (Oosterhof & Todorov, 2008) and that trust judgments can be formed after only a brief exposure (100 ms) to a face (Willis & Todorov, 2006). It is also important for the robot's appearance to be compatible with its function at face value. Goetz, Kiesler, & Powers (2003) found that people are more likely to accept a robot when its appearance matches its perceived capabilities. This is thought to be the case because when there is a high level of compatibility between appearance and functionality, users expectations are confirmed, boosting confidence in the robot's performance. However, when appearance and capabilities are incompatible, user expectations are violated, which can result in lower levels of trust (Duffy, 2003).

Because studies of human robot interaction are a new field, there are many gaps in the literature especially regarding the social influences on HRI. First, although there is evidence to suggest that stereotypes can affect performance and interactions with anthropomorphized technological aids, we do not know how pre-existing age stereotypes will affect HRI. Next, it is unclear how trust might be moderated by task type and reliability. Although the automation literature suggests that reliability can influence trust, to our knowledge the relationship between robot task domain and trust has not yet been investigated. Finally, how does stereotyping technology affect perception of capabilities and the causal attributions made about performance?

The Current Study

The purpose of this study is to better understand the factors that influence older adults' trust in robots. Specifically, we are investigating whether the robots' appearance, task domain, and reliability of the robot's performance influence trust in the automation. A cross-sectional factorial survey study will be utilized using video vignettes to assess participants' attitudes towards the robots' behavior and appearance. Each vignette will include manipulations of the age of the robot, the domain of the collaborative task, and the reliability of the robot's performance. Dependent measures will include the level of trust participants exhibit toward the robot, causal attributions regarding the robot's performance, and perceived capabilities of the robot.

It is hypothesized that manipulating a robot's appearance, level of reliability, and the task type will have an effect on the level of trust that an older adult exhibits toward a robot, the causal attributions that the individual makes about the robot's performance, and people's perceptions of the capabilities of the robot. Specifically, trust in the robot should be highest when the task is stereotypically congruent with the robot's appearance (e.g., a younger adult performing a cognitive task instead of an older adult performing a cognitive task) and its performance is reliable. This is hypothesized because appearance influences people's trust in automation (Pak,

Fink, Price, Bass, & Sturre, 2012) and aging stereotypes will less likely be activated while interacting with the younger robot. The attributions about the robot's performance may be more dispositional when reliability is low and the task is incongruent with the robot's appearance. This is because older adults are more likely to make dispositional (i.e., internal) attributions of blame when an outcome of an event is perceived as negative (the unreliable condition) and when their beliefs are violated (i.e., when an older looking robot performs the cognitive and physical tasks; Blanchard-Fields, Hertzog, & Horhota, 2012). Perceived capabilities of the robot are hypothesized to depend on the robot's appearance. That is, capability ratings are expected to be higher when the younger looking robot performs the tasks, and rankings are expected to be lower when an older looking robot performs the tasks. This is expected because adults' capabilities in cognitive and physical domains are expected to decline with age (Kite, Stockdale, Whitley, & Johnson, 2005; Davis & Friedrich, 2010). Task domain will be treated as an exploratory variable. However, based on automation trust literature suggesting that trust in robot's capabilities might depend on the domain in which they are placed (e.g., industry, entertainment, social; Schaefer, Sanders, Yordon, Billings, & Hancock, 2012), it is hypothesized that there will be a main effect of task domain such that participants will have more trust in the robot and have higher ratings of perceived capabilities when the robot performs physical tasks.

Methods, Procedures, Results

[methods can be found in Branyon (2015), preliminary results in Branyon & Pak (2015) attached in appendix]

Conclusion

This study offers a unique contribution by investigating a well-researched paradigm from the social cognition and aging literatures, stereotypes, and applying it to a novel field, HRI. Preliminary analyses show that although there were no main effects of robot age on the dependent variables, age moderated the effect of task on the robot's perceived capabilities as well as the types of causal attributions individuals made about the robot's performance. In general, the robot was perceived more positively when completing a fine motor task or light cognitive tasks than when it performed a gross motor task (i.e., moving boxes). Reliable cognitive task performance yielded the highest dispositional attribution ratings regardless of robot appearance. This finding suggests that people might attribute outcomes differently in the context of human-robot interaction than in human-human interaction. These findings emphasize the importance of task type on older adults' perceptions of robots. In this context, users trust robots that perform cognitive and light motor tasks more than ones that perform gross motor tasks. It is also important to select the appropriate age appearance for robots based on the tasks they are to perform. Tentatively, the results suggest selecting a younger appearance for a robot that will perform cognitive tasks.

References

- Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdale. *Nature*, ProQuest Nursing & Allied Health Source, 669.
- Amadeo, R. (2014, June 15). Hands-on with Baxter, the factory robot of the future. Retrieved from <http://arstechnica.com/gadgets/2014/06/hands-on-with-baxter-the-factory-robot-of-the-future/1/>
- Antifakos, S., Kern, N., Schiele, B., & Schwaninger, A. (2005). Towards improving trust in context-aware systems by displaying system confidence. Proceedings from Mobile HCI '05: *The 7th International Conference on Human Computer Interaction with Mobile Devices & Services*. Salzburg, Austria. Approach (pp. 15–67). Beverly Hills, CA: SAGE Publications.
- Ashmore, R. D., & Del Boca, F. K. (1981). Conceptual approaches to stereotypes and stereotyping. In *Cognitive processes in stereotyping and intergroup behavior*. Edited by D. L. Hamilton, 1–35. Hillsdale, NJ: Erlbaum.
- attention: Selective deficits in healthy adult carriers of the E4 allele of the apolipoprotein E
- Baddeley, A. (1986). *Working memory*. New York: Oxford University Press.
- Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19, 775–779.
- Banaji, M. R., Hardin, C., & Rothman, A. J. (1993). Implicit stereotyping in person judgment. *Journal of Personality and Social Psychology*, 65(2), 272-281.
- Bartneck, C., & Reichenbach, J. (2005). Subtle emotional expressions of synthetic characters. *International Journal Human-Computer Studies*, 62, 179-192.
- Batty, M., & Taylor, M. J. (2003). Early processing of the six basic facial emotional expressions. *Cognitive Brain Research*, 17, 613-620.
- Bickmore, T. (2011). Etiquette in motivational agents. In C. C. Hayes & C. A. Miller (Eds.), *Human-Computer Etiquette: Cultural Expectations and the Design Implications They Place on Computers and Technology* (206-226). Boca Raton: Auerbach.
- Blanchard-Fields, F. (1994). Age differences in causal attributions from an adult developmental perspective. *Journal of Gerontology*, 49, 43-51. doi:10.1093/geronj/49.2.P43
- Blanchard-Fields, F. (1996). Causal attributions across the adult life span: The influence of social schemas, life context, and domain specificity. *Applied Cognitive Psychology*, 10, 137-146.
- Blanchard-Fields, F., Chen, Y., Schocke, M., Hertzog, C. (1998). Evidence for content-specificity of causal attributions across the adult life span. *Aging, Neuropsychology, & Cognition*, 5, 241-263.
- Blanchard-Fields, F., Hertzog, C., & Horhota, M. (2012). Violate my beliefs? Then you're to blame! Belief content as an explanation for causal attribution biases. *Psychology and Aging*, 27, 324-337. doi:10.1037/a0024423
- Brewer, M. B., & Lui, L. (1984). Categorization of the elderly by the elderly. *Personality and Social Psychology Bulletin*, 10, 585-595.
- Brule, R., Dotsch, R., Bijlstra, G., Wigboldus, D. J., & Haselager, P. (2014). Do robot performance and behavioral style affect human trust?: A multi-method approach. *International Journal Of Social Robotics*, doi:10.1007/s12369-014-0231-5
- Bubb-Lewis, C., & Scerbo, M. (1997). Getting to know you: Human computer communication in adaptive automation. In M. Mouloua & J. M. Koonce (Eds.), *Human-automation interaction: Research and practice* (pp. 92–99). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

- Byrne, D. E. (1971). *The attraction paradigm* (Vol. 11). New York: Academic Press.
- Calvo, M.G., & Lundqvist, D. (2008). Facial expressions of emotion (KDEF): Identification under different display-duration conditions. *Behavior Research Methods*, 40(1), 109-115.
- Carstensen, L.L., Issacowitz, D.M., & Charles, S.T. (1999). Taking time seriously: A theory of socioemotional selectivity. *American Psychologist*, 54(3), 165-181.
- Casper, C., Rothermund, K., & Wentura, D. (2011). The activation of specific facets of age stereotypes depends on individuating information. *Social Cognition*, 29(4), 393-414.
- Chen, J. Y. C. & Terrence, P. I. (2008). Effects of tactile cueing on concurrent performance of military and robotics tasks in a simulated multitasking environment. *Ergonomics*, 51, 1137-1152.
- Chernoff, H. (1973). The use of faces to represent points in k dimensional space graphically. *Journal of the American Statistical Association*, 68(342), 361-368.
- Cleveland, J. N., & Hollmann, G. (1990). The effects of the age-type of tasks and incumbent age compositions on job perceptions. *Journal of Vocational Behaviour*, 36(2), 181-194.
- Crocoll, W. M., & Coury, B. G. (1990). Status or recommendation: Selecting the type of information for decision aiding. *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 1524-1528). Santa Monica, CA: Human Factors and Ergonomics Society.
- Cuddy, A. J., Norton, M. I., & Fiske, S. T. (2005). This old stereotype: The pervasiveness and persistence of the elderly stereotype. *Journal of Social Issues*, 61(2), 267-285.
- Czaja, S. J., & Sharit, J. (1998). Age differences in attitudes toward computers. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 53(5), 329-340.
- Czaja, S. J., Charness, N., Fisk, A. D., Hertzog, C., Nair, S. N., Rogers, W. A., & Sharit, J. (2006). Factors predicting the use of technology: Findings from the center for research and education on aging and technology enhancement (CREATE). *Psychology and Aging*, 21, 333-352. doi:10.1037/0882-7974.21.2.333
- Davis, N. C., & Friedrich, D. (2010). Age stereotypes in middle-aged through old-old adults. *The International Journal Of Aging & Human Development*, 70(3), 199-212. doi:10.2190/AG.70.3.b
- de Visser, E., , Shaw, T., Mohamed-Ameen, A., & Parasuraman, R. (2010). Modeling human-automation team performance in networked systems: Individual differences in working memory count. *Proceedings of the Human Factors and Ergonomics Society 54th Annual Meeting*, 1087-1091.
- DeArmond, S., Tye, M., Chen, P. Y., Krauss, A., Apryl Rogers, D., & Sintek, E. (2006). Age and gender stereotypes: New challenges in a changing workplace and workforce. *Journal of Applied Social Psychology*, 36(9), 2184-2214.
- Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, 56(1), 5-18.
- Drucquer, M. H., & McNally, P.G. (1998). *Diabetes management: Step by step*. Osney Mead, Oxford: Blackwell Science Ltd.
- Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42, 177-190.
- Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human Computer Studies*, 58(6), 697-718.
- Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L.A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors* 44(1), 79-94.

- Ekman, P. & Friesen, W.V. (1975). *Unmasking the face: A guide to recognizing emotions from facial cues*. Oxford, England: Prentice-Hall.
- Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, *42*, 462–492.
- Endsley, M. R., & Kiris, E. O. (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.
- Experimental test of recognition, similarity-attraction, and consistence-attraction. *Journal*
- Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal Of Social Psychology*, *51*(4), 724-731. doi:10.1111/j.2044-8309.2011.02082.x
- Finkelstein, L. M., & Burke, M. J. (1998). Age stereotyping at work: The role of rater and contextual factors on evaluations of job applicants. *The Journal of General Psychology*, *125*(4), 317-345.
- Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, and G. Linclzey (Eds.), *The Handbook of Social Psychology* (4th ed.). New York: McGraw-Hill.
- Fiske, S., & Taylor, S. (1991). *Social Cognition*. New York: McGraw-Hill.
- Fitzgerald, J.T., Anderson, R.M., Funnell, M.M., Hiss, R.G., Hess, G.E., Davis, W.K., & Barr, P.A. (1998). The reliability and validity of a brief diabetes knowledge test. *Diabetes Care*, *21*(5), 706–710.
- Galster, S. M., Bolia, R. S., & Parasuraman, R. (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* (pp. 438–442). Santa Monica, CA: Human Factors and Ergonomics Society.
- gene. *Proceedings of the National Academy of Sciences, USA*, *97*, 11661–11666.
- Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, *117*, 21-38.
- Goetz, J., Kiesler, S., & Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. *The 12th IEEE International Workshop on Robot and Human Interactive Communication, Vol., IXX*, 55-60
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4-27.
- Greenwood, P. M., Lambert, C., Sunderland, T., & Parasuraman, R. (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
- Greenwood, P. M., Sunderland, T., Friz, J. L., & Parasuraman, R. (2000). Genetics and visual
- Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517-527. doi:10.1177/0018720811417254
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam: Elsevier Science.

- Hayes, C.C., & Miller, C. A. (2011). *Human-Computer Etiquette: Cultural Expectations and the Design Implications They Place on Computers and Technology*. Boca Raton: Auerbach.
- Hess, U., Blairy, S., & Kleck, R. E. (1997). The intensity of emotional facial expressions and decoding accuracy. *Journal of Nonverbal Behavior*, 21(4), 241-257.
- Hilton, J. L., & Von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology*, 47(1), 237-271.
- Hippel, von, W., Silver, L. A., & Lynch, M. E. (2000). Stereotyping against your will: The role of inhibitory ability in stereotyping and prejudice among the elderly. *Personality and Social Psychology Bulletin*, 26(5), 523-532.
- Ho, G., Wheatley, D., & Scialfa, C. T. (2005). Age differences in trust and reliance of a medication management system. *Interacting with Computers*, 17(6), 690-710.
- Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behaviour Research Methods*, 39(1), 101-117.
- Hox, J. J., & Bechger, T. M. (1998). An introduction to structural equation modeling. *Family Science Review*, 11, 354-373.
- Hox, J. J., Kreft, I. G. G., & Hermkens, P. L. J. (1991). The analysis of factorial surveys. *Sociological Methods & Research*, 19, 493-510.
- Hummert, M. L. (1993). Age and typicality judgments of stereotypes of the elderly: Perceptions of elderly vs. young adults. *International Journal of Aging and Human Development*, 37, 217-227.
- Hummert, M. L. (1994). Physiognomic cues and the activation of stereotypes of the elderly in interaction. *International Journal of Aging and Human Development*, 39, 5-20.
- Hummert, M. L., Garstka, T. A., Shaner, J. L., & Strahm, S. (1994). Stereotypes of the elderly held by young, middle-aged, and elderly adults. *Journal of Gerontology*, 49(5), 240-249. In P. H. Rossi & S. L. Nock (Eds.), *Measuring Social Judgments. The Factorial Survey*
- Jian, J., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1), 53-71.
- Johnson, D., Gardner, J., & Wiles, J. (2004). Experience as a moderator of the media equation: The impact of flattery and praise. *International Journal of Human-Computer Studies*, 61(3), 237-258.
- Kabulov, B.T. (1992). A method for constructing Chernoff faces oriented toward interval estimates of the parameters. *Soviet Journal of Computers and System Sciences*, 30(3), 94-97.
- Kite, M. E., & Johnson, B. T. (1988). Attitudes toward older and younger adults: A meta-analysis. *Psychology and Aging*, 3, 233-244.
- Kite, M. E., Deaux, K., & Miele, M. (1991). Stereotypes of young and old: Does age outweigh gender? *Psychology and Aging*, 6(1), 19-27.
- Kite, M. E., Stockdale, G. D., Whitley, B. E., & Johnson, B. T. (2005). Attitudes toward younger and older adults: An updated meta-analytic review. *Journal of Social Issues*, 61(2), 241-266.
- Krull, D. S. (1993). Does the grist change the mill? The effect of perceiver's goals on the process of social inference. *Personality and Social Psychology Bulletin*, 19, 340-348.
- Krull, D. S., & Erikson, D. J. (1995). Judging situations: On the effortful process of taking dispositional information into account. *Social Cognition*, 13, 417-438.

- Kuchenbrandt, D., Häring, M., Eichberg, J., Eyssel, F., & André, E. (2014). Keep an eye on the task! How gender typicality of tasks influence human–robot interactions. *International Journal Of Social Robotics*, 6(3), 417-427. doi:10.1007/s12369-014-0244-0
- Kunda, Z., & Sherman-Williams, B. (1993). Stereotypes and the construal of individuating information. *Personality and Social Psychology Bulletin*, 19(1), 90-99.
- Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviours: A parallel constraint satisfaction theory. *Psychological Review*, 103(2), 284-308.
- Lee, E. J. (2003). Effects of “gender” of the computer on informational social influence: The moderating role of task type. *International Journal of Human-Computer Studies*, 58(4), 347-362.
- Lee, E.-J. (2010). The more humanlike, the better? How speech type and users’ cognitive style affect social responses to computers. *Computers in Human Behaviour*, 26(4), 665–672.
- Lee, J. D. (2006). Affect, attention, and automation. In A. Kramer, D. Wiegmann & A. Kirlik (Eds.), *Attention: From theory to practice*. New York: Oxford University Press.
- Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human Computer Studies*, 40(1), 153-184.
- Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors*, 46(1), 50-80.
- Lee, M.D., Reilly, R.E., & Butavicius, M.A. (2003). An empirical evaluation of Chernoff faces, star glyphs, and spatial visualizations for binary data. *Proceedings from APVis '03: Asia Pacific Symposium on Information Visualization*, 24, 1-10.
- Lineweaver, T. T., & Hertzog, C. (1998). Adults' Efficacy and Control Beliefs Regarding Memory and Aging: Separating General from Personal Beliefs. *Aging, Neuropsychology, and Cognition*, 5, 264-296.
- Lohse, G. (1997). The role of working memory on graphical information processing. *Behaviour & Information Technology*, 16(6), 297-308.
- Lorenz, B., Di Nocera, F., Röttger, S., & Parasuraman, R. (2002). Automated fault management in a simulated space flight micro- world. *Aviation, Space, and Environmental Medicine*, 73, 886–897.
- Lynchard, N.A., & Radvansky, G.A. (2012). Age-related perspectives and emotion processing. *Psychology and Aging*. Advance online publication. doi: 10.1037/a0027368.
- Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors* 48(2), 241-256.
- Madhavan, P., Wiegmann, D., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors* 48(2), 241-256.
- Maes, P. (1994). Agents that reduce work and information overload. *Communications of the ACM*, 37, 30–40.
- Mankoff, J., Dey, A. K., Hsieh, G., Kientz, J., Lederer, S., Ames, M. (2003). Heuristic evaluation of ambient displays. Proceedings from SIGCHI '03: *Conference on Human Factors in Computing Systems*. Ft. Lauderdale, FL, USA.
- Marakas, G. M., Johnson, R. D., & Palmer, J. W. (2000). A theoretical model of differential social attributions toward computing technology: When the metaphor becomes the model. *International Journal of Human-Computer Studies*, 52(4), 719-750.
- McKinstry, B., & Yang, S. Y. (1994). Do patients care about the age of their general practitioner? A questionnaire survey in five practices. *The British Journal of General Practice*, 44(385), 349-351.

- Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human automation interactions. *Human Factors* 50(2), 194-210.
- Merritt, S. M., Heimbaugh, H., & LaChapell, J. (2012). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors*, 55(3), 520-534.
- Montello, D.R., & Gray, M.V. (2005). Miscommunicating with isolines: Design principles for thematic maps. *Cartographic Perspectives*, 10-19.
- Moray, N., Hiskes, D., Lee, J., & Muir, B. M. (1995) Trust and human intervention in automated
- Morris, C. J., Ebert, D. S. & Rheingans, P. (2000). An experimental analysis of the effectiveness of features in chernoff faces. Proceedings from SPIE '00: *The 28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making*. Washington, D. C., USA.
- Morris, J. S., Friston, K. J., Buchel, C., Firth, C. D., Young, A.W., Calder, A. J., & Dolan, R. J. (1998). A neuromodulatory role for the human amygdala in processing emotional facial expressions. *Brain*, 121, 47-57.
- Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other?. In R. Parasuraman, M. Mouloua (Eds.) , *Automation and human performance: Theory and applications* (pp. 201-220). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
- Mosier, K. L., Skitka, L. J., Burdick, M. D., & Heers, S. T. (1996). Automation bias, accountability, and verification behaviours. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 40(4), 204-208.
- Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits. *The International Journal of Aviation Psychology*, 8(1), 47-63.
- Mouloua, M., Smither, J. A.-A., Vincenzi, D. A., & Smith, L. (2002). Automation and aging: Issues and considerations. *Advances in Human Performance and Cognitive Engineering Research*, 2, 213-237.
- Muir, B. M. (1987). Trust between human and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27, 527-539.
- Muir, B. M., & Moray, N. (1996). Trust in automation: Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39, 429-460
- Muller-Johnson, K, Togli, M.P., Sweeney, C.D., Ceci, S.J. (2007) The perceived credibility of older adults as witnesses and its relation to ageism. *Behavioral Sciences and the Law*, 25, 355-375.
- Nass, C., & Lee, K. M. (2001). Does computer-generated speech manifest personality?
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56, 81-103.
- Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender stereotypic responses to computers with voices. *Journal of Applied Social Psychology*, 27, 864-876.
- Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence*, 72-78.
- Nelson, E.S. (2007). The face symbol: Research issues and cartographic potential. *Cartographica*, 42(1), 53-64.
- new media like real people and places. New York: Cambridge University Press.
- of Experimental Psychology: Applied*, 7, 171-181.

- Ohman, A., Lundqvist, D., & Esteves, F. (2001). The face in the crowd revisited: A threat advantage with schematic stimuli. *Journal of Personality and Social Psychology*, *80*(3), 381-396.
- Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. *Human Factors*, *56*(3), 476-488.
- Oosterhof, N.N., Todorov, A. (2008). The functional basis of face evaluation.
- Orgeta, V., & Phillips, L. H. (2007). Effects of age and emotional intensity on the recognition of facial emotion. *Experimental Aging Research*, *34*(1), 63-79.
- Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 201-220). Mahwah, NJ: Erlbaum
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381-410.
- Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, *47*(4), 51-55.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230-253.
- Parasuraman, R., de Visser, E., Lin, M.-K., & Greenwood, P. M. (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., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced "complacency." *International Journal of Aviation Psychology*, *3*, 1-23.
- Perdue, C. W., & Gurtman, M. B. (1990). Evidence for the automaticity of ageism. *Journal of Experimental Social Psychology*, *26*(3), 199-216.
- Posthuma, R. A., & Campion, M. A. (2009). Age stereotypes in the workplace: Common stereotypes, moderators, and future research directions? *Journal of Management*, *35*(1), 158-188.
- Poupyrev, I., Maruyama, S., & Rekimoto, J. (2002). Ambient touch: Designing tactile interfaces for handheld devices. Proceedings from UIST '02: *The 15th Annual ACM Symposium on User Interface Software and Technology*. Paris, France. *Proceedings of the National Academy of Sciences*, *105*, 11087-11092
- Qiu, L., & Benbasat, I. (2009). Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of Management Information Systems*, *25*(4), 145-182.
- Radvansky, G. A., Copeland, D. E., & Hippel, W. V. (2010). Stereotype activation, inhibition, and aging. *Journal of Experimental Social Psychology*, *46*(1), 51-60.
- Real Studio (2012). [Computer software]. Austin, TX.
- Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and Rellecke, J., Palazova, M., Sommer, W., & Schacht, A. (2011). On the automaticity of emotion processing in words and faces: Event-related brain potentials evidence from a superficial task. *Brain and Cognition*, *77*, 23-32.
- Rosen, B., & Jerdee, T. H. (1976). The nature of job-related age stereotypes. *Journal of Applied Psychology*, *61*(2), 180-183.
- Rossi, P. H., & Anderson, A. B. (1982). The factorial survey approach: An introduction. In P. H. Rossi, S. L. Nock (Eds.), *Measuring Social Judgments*, Sage Publications, Beverly Hills.

- Röttger, S., Bali, K., & Manzey, D. (2009). Impact of automated decision aids on performance, operator behaviour and workload in a simulated supervisory control task. *Ergonomics*, 52(5), 512-523.
- Rovira, E., Cross, A., Leitch, E., & Bonaceto, C. (2014). Displaying Contextual Information Reduces the Costs of Imperfect Decision Automation in Rapid Retasking of ISR Assets. *Human Factors*.
- Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors* 49(1), 76-87.
- Sarter, N., & Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human Factors*, 43, 573–583.
- Sawhney, N., & Schmandt, C. (2000). Nomadic radio: Speech and audio interaction for contextual messaging in nomadic environments. *ACM Transactions on Computer-Human Interaction*, 7(3), 353-383.
- Shaefer, K. E., Sanders, T. L., Yordon, R. E., Billings, D. R., & Hancock, P. A. (2012). Classification of robot form: Factors predicting perceived trustworthiness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56: 1548. doi: 10.1177/1071181312561308
- Shah, R., & Ogden, J. (2006). What's in a face? The role of doctor ethnicity, age and gender in the formation of patients' judgements: An experimental study. *Patient Education and Counseling*, 60(2), 136-141.
- Sharit, J., Czaja, S. J., & Nair, S. (2003). Effects of age, speech rate, and environmental support in using telephone voice menu systems. *Human Factors*, 45(2), 234-251.
- Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators* (Technical report). Cambridge, MA: MIT, Man Machine Systems Laboratory.
- Singh, I. L., Molloy, R., & Parasuraman, R. (1993). Automation- induced “complacency”: Development of the complacency potential rating scale. *International Journal of Aviation Psychology*, 3, 111–121.
- systems. Hillsdale, NJ: Erlbaum
- Tabachnick, B. G., & Fidell, L. S. (2007). Multi-level linear modeling. In *Using Multivariate Statistics* (5 ed., pp. 781–857).
- Tsurusawa, R., Goto, Y., Mitsudome, A., Nakashima, T., Tobimatsu, S. (2008). Different perceptual sensitivities for Chernoff's face between children and adults. *Neuroscience Research*, 60(2), 176-183.
- Vicente, K. J., & Williges, R. C. (1988). Accommodating individual differences in searching a hierarchical file system. *International Journal of Man-Machine Studies*, 29(6), 647–668. doi:10.1016/S0020-7373(88)80072-5
- Wechsler, D. (1997). Wechsler Memory Scale III. (3rd Ed.). San Antonio, TX: The Psychological Corporation.
- Weiser, M., & Brown, J. S. (1995). Designing calm technology. Retrieved from <http://www.ubiq.com/weiser/calmtech/calmtech.htm>.
- Whalen, P. J., Rauch, S. L., Etcoff, N. L., McInerney, S. C., Lee, M. B., & Jenike, M. A. (1998). Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *The Journal of Neuroscience*, 18(1), 411-418.
- Wickens, C. D., & Xu, X. (2002). *Automation trust, reliability and attention* (Tech. Rep. AHFD-02-14/MAAD-02-2). Savoy: University of Illinois, Aviation Research Lab

- Willis, J., Todorov, A. (2006). First impressions: making up your mind after a 100-Ms exposure to a face. *Psychological Science, 17*, 592–598.
- Wisneski, C. (1999). The design of personal ambient displays. (Unpublished master's thesis). MIT, Boston, MA.
- Zhang, T., Zhu, B., & Kaber, D. B. (2011). Anthropomorphism and social robots. In C.C. Hayes & C.A. Miller (Eds.), *Human-Computer Etiquette: Cultural Expectations and the Design Implications They Place on Computers and Technology* (231-255). Boca Raton: Auerbach.

Appendix

Publications:

- Pak, R., McLaughlin, A. C., & Bass, B. (2014). A Multi-level Analysis of the Effects of Age and Gender Stereotypes on Trust in Anthropomorphic Technology by Younger and Older Adults. *Ergonomics*.
- Rovira, E., Pak, R., & McLaughlin, A. C. (under review). Low Memory, Mo' Problems: Effects of individual differences on types and levels of automation. *Human Factors*.

Conference Proceedings

- Branyon, J. J., & Pak, R. (2015). Investigating older adults' trust, attributions, and capability perceptions of robots. *Presented at the American Psychological Association 123rd Annual Meeting*. Toronto, ON: American Psychological Association
- Leidheiser, W., & Pak, R. (2014). The Effects of Age and Working Memory Demands on Automation-Induced Complacency. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58(1), 1919–1923. doi:10.1177/1541931214581401
- Bass, B. M., Goodwin, M., Brennan, K., Pak, R., & McLaughlin, A. C. (2013). Effects of age and gender stereotypes on trust in an anthropomorphic decision aid. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 57(1), 1575-1579.
- Bass, B. M., & Pak, R. (2012). Faces as Ambient Displays: Assessing the attention-demanding characteristics of facial expressions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56(1), 2142–2146.

Student Thesis:

- Bass, B. (2014). Faces as Ambient Displays: Assessing the Attention-Demanding Characteristics of Facial Expressions. Unpublished master's thesis. Available at: http://tigerprints.clemson.edu/all_theses/1941/
- Branyon, J. (in progress). Investigating older adults' trust, causal attributions, and perception of capabilities in robots as a function of robot appearance, task, and reliability.
- Leidheiser, W. (in progress). The Effects of Age and Working Memory Demands on Automation-Induced Complacency.

This article was downloaded by: [Clemson University]

On: 20 June 2014, At: 06:30

Publisher: Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Ergonomics

Publication details, including instructions for authors and subscription information:

<http://www.tandfonline.com/loi/terg20>

A multi-level analysis of the effects of age and gender stereotypes on trust in anthropomorphic technology by younger and older adults

Richard Pak^a, Anne Collins McLaughlin^b & Brock Bass^a

^a Department of Psychology, Clemson University, Clemson, USA

^b Department of Psychology, North Carolina State University, Raleigh, USA

Published online: 17 Jun 2014.

To cite this article: Richard Pak, Anne Collins McLaughlin & Brock Bass (2014): A multi-level analysis of the effects of age and gender stereotypes on trust in anthropomorphic technology by younger and older adults, *Ergonomics*, DOI: [10.1080/00140139.2014.928750](https://doi.org/10.1080/00140139.2014.928750)

To link to this article: <http://dx.doi.org/10.1080/00140139.2014.928750>

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at <http://www.tandfonline.com/page/terms-and-conditions>

APPENDIX 1: Pak, R., McLaughlin, A. C., & Bass, B. (2014). A Multi-level Analysis of the Effects of Age and Gender Stereotypes on Trust in Anthropomorphic Technology by Younger and Older Adults. *Ergonomics*.

A multi-level analysis of the effects of age and gender stereotypes on trust in anthropomorphic technology by younger and older adults

Richard Pak^{a*}, Anne Collins McLaughlin^b and Brock Bass^a

^aDepartment of Psychology, Clemson University, Clemson, USA; ^bDepartment of Psychology, North Carolina State University, Raleigh, USA

(Received 17 December 2013; accepted 19 May 2014)

Previous research has shown that gender stereotypes, elicited by the appearance of the anthropomorphic technology, can alter perceptions of system reliability. The current study examined whether stereotypes about the perceived age and gender of anthropomorphic technology interacted with reliability to affect trust in such technology. Participants included a cross-section of younger and older adults. Through a factorial survey, participants responded to health-related vignettes containing anthropomorphic technology with a specific age, gender, and level of past reliability by rating their trust in the system. Trust in the technology was affected by the age and gender of the user as well as its appearance and reliability. Perceptions of anthropomorphic technology can be affected by pre-existing stereotypes about the capability of a specific age or gender.

Practitioner Summary: The perceived age and gender of automation can alter perceptions of the anthropomorphic technology such as trust. Thus, designers of automation should design anthropomorphic interfaces with an awareness that the perceived age and gender will interact with the user's age and gender.

Keywords: automation; trust; aging; stereotypes; mobile; health

1. Anthropomorphic technology can elicit stereotypes

Interactive computer systems that exhibit human-like, or anthropomorphic, traits can lead users to perceive and treat them differently than non-human-like systems (Nass, Steuer, and Tauber 1994). Thus, it is imperative to understand how users' perceptions of the system might be affected by their social reactions to anthropomorphic technology. One way in which a system may elicit social reactions is by eliciting stereotypes (Yee, Bailenson, and Rickerson 2007).

Stereotypes are preconceptions about the traits, behaviour, or abilities of a group and can set expectations of a stereotyped individual. Stereotypes can have both negative and positive connotations that may be inconsistent with real group attributes but provide adaptive value because they filter and organise incoming information, thereby easing processing and interpretation (Hilton and von Hippel 1996). Stereotypes can be activated and applied with or without conscious awareness (Banaji, Hardin, and Rothman 1993; Greenwald and Banaji 1995). Unfortunately, when the stereotype is highly simplified or inaccurate, it can lead to errors in perceptions and behaviour.

Nass, Steuer, and Tauber (1994) tested whether users would apply gender-related stereotypes when interacting with a computer that exhibited a gender. Their participants were first tutored by a computer on a specific topic. Tutored topics were either stereotypically female (love and relationships) or stereotypically male (computers and technology). They then moved to a non-gendered computer for testing and to a gendered computer for evaluation of their test responses. When gender of the tutor matched the stereotypical topic, participants rated it as a better teacher. This finding was echoed by Lee (2003) in a study where participants answered difficult trivia questions that were either stereotypically feminine or masculine. After answering the trivia question, participants viewed a female or male computerised agent that presented its own answer and then were allowed to change their answer. More participants changed their answers to agree with the agent when the gender of the agent matched the stereotypical topic.

Stereotype activation for computerised agents can also interact with individual differences, such as physical characteristics. Qiu and Benbasat (2009) found that an anthropomorphic decision aid significantly increased perceptions of social presence and led to increased trust of the agent. The strength of these effects was influenced by the degree to which the decision aid agent was similar to the user on a visible factor, such as ethnicity. The link between trust and apparent physical characteristics was explained via similarity-attraction theory that predicted that people would be more attracted to those similar to them (Byrne 1971). The user may have attributed their attraction to a similar ethnicity as trustworthiness of the agent.

*Corresponding author. Email: richpak@clemson.edu

In another example of the moderating role of individual differences in susceptibility to anthropomorphic effects, susceptibility to flattery (insincere praise) depended on the level of computer experience of the user (Johnson, Gardner, and Wiles 2004). Johnson, Gardner, and Wiles found that susceptibility to flattery from a computer depended on the user's experience level with computers – the judgments of highly experienced users were more affected by flattery than less experienced users. Furthermore, Lee (2010) found that people who exhibited less analytical and more intuitive cognitive style were more susceptible to flattery from a computer.

In summary, stereotypes can affect user perceptions of a computer or automated aid and can be moderated by individual differences. Some of the aids described in the previous studies were forms of automation that functioned in a decision-support capacity; thus, some automation bias may be based on stereotypes (Skitka, Mosier, and Burdick 1999). However, no research has explicitly examined how these factors might interact with machine-related factors of automation, such as reliability of the automation or how various activated stereotypes might interact (e.g. age and gender).

2. Age stereotypes in technology?

Age is one of the first and most salient attributes noticed of a person (Fiske 1998) which suggests it may also be true with anthropomorphic agents. Furthermore, stereotypes about age are stronger (Kite, Deaux, and Miele 1991) and more complex than gender stereotypes (Kite et al. 2005). In Kite, Deaux, and Miele's study assessing age and gender stereotypes using free response, participants viewed a younger (35-year-old) male or female and older (65-year-old) male or female and provided characteristics of the target person. Analysis showed that when negative stereotypes were generated, they were much more likely to be due to the age of the target than the gender. Finally, according to the similarity-attraction hypothesis (Qui and Benbasat 2009), older and younger adults should exhibit positive anthropomorphic effects with automation that matches their age group. However, it may also be that an older-looking automated agent may prime negative stereotypes about age, particularly when the reliability of the automation is perceived to be low. This may explain why a previous study found that a young female agent enhanced younger adults' trust in automation but not older adults' when participants interacted with a health decision aid (Pak et al. 2012). The authors hypothesised that the dissimilarity between a younger female decision agent and an older participant may have muted any potential anthropomorphic effect on trust due to violation of the similarity-attraction. An alternative explanation is that older adults hold negative stereotypes of the capabilities of younger, female doctors but younger adults do not.

3. Age and gender stereotypes of physicians

People hold stereotypes that older workers have lower ability, are less motivated, and are less productive than younger workers (Posthuman and Campion 2009). Older workers are also seen as less adaptable to changing work situations and uncertainty than younger workers (DeArmond et al. 2006). Although aging studies show that these views may be exaggerated (e.g. see Czaja and Sharit 1998), they are widely held by people of all ages and affect workplace hiring decisions and evaluations (DeArmond et al. 2006; Posthuma and Campion 2009). Negative age stereotypes about older workers are even held by older adults themselves (Rosen and Jerdee 1976; Finkelstein and Burke 1998; Wrenn and Maurer 2004). Finally, these stereotypes may be activated without awareness (Devine 1989; Perdue and Gurtman 1990; Banaji and Hardin 1996).

Activation of age stereotypes may be moderated by individuating past behaviour or context (Kunda and Sherman-Williams 1993). Individuating information such as context (e.g. interacting with a doctor) may determine which aspect of a stereotype gets activated (Casper, Rothermund, and Wentura 2011). Knowing the occupation of an individual is a type of individuating information that seems to alter some negative age stereotypes. For example, although some occupations seem more negatively age stereotyped (e.g. Cleveland and Hollman 1990), the occupation of physician is moderately seen as a stereotypically older male occupation (Singer 1986) even though it is an occupation that may require adaptability and is faced with uncertainty. In contrast, when stereotypes of doctors were more recently assessed (Shah and Ogden, 2006), younger female doctors were perceived as having better personal manner and technical skill than older doctors of either gender. The scant literature on physician age stereotypes seems to suggest that the stereotype of older doctors is less negative than the stereotype for older adults in general, but still present (McKinstry and Yang 1994), demonstrating the power of individuating information on the otherwise powerful age stereotype.

In summary, person-judgment based on stereotypes can depend on individuating information, including profession, past performance (i.e. reliability), gender, and age. Similarly, assessment of computer-based automation with human-like characteristics may also be subject to pre-existing stereotypes consistent with the human-like qualities (e.g. age, gender). Anthropomorphic automation with ambiguous reliability may be more likely to activate pre-existing stereotypes. That is, when automation is unambiguously reliable or unreliable, stereotypes should not affect perceptions. But when automation is ambiguous, stereotypes will affect perceptions of the automation such as trust. The idea that imperfect automation may

engender the expression of implicit attitudes has been suggested by other automation researchers (Lee and See 2004; Merritt, Heimbaugh, and LaChapell 2012).

4. Anthropomorphism and automation characteristics

Studies of human–automation interaction have demonstrated that many factors related to the person, automated system, and task interact to determine trust in and performance with automation. For example, individual differences in attitudes towards automation (e.g. Mosier et al. 1998; Dzindolet et al. 2003; Merritt and Ilgen 2008) interacted with machine characteristics such as reliability and error types (e.g. Madhavan, Wiegmann, and Lacson 2006; Rovira, McGarry, and Parasuraman 2007) and task or situational factors such as workload (e.g. Röttger, Bali, and Manzey 2009) to affect behaviour with and perceptions of automation.

Research investigating the influence of anthropomorphic aspects specifically on human–automation interaction (Parasuraman and Miller 2004, Pak et al. 2012) found that various implementations of anthropomorphism such as etiquette (Bickmore 2011; Zhang, Zhu, and Kaber 2011) affected perceptions of trust and automation behaviour. For example, in aircraft engine diagnosis, the automation either presented advice in a rude or polite manner (Parasuraman and Miller 2004). As expected, perceived trust and performance in the diagnosis task was better when the automation was 80% reliable compared to 60% reliable. However, engine diagnosis performance and trust with polite but less reliable automation was the same as rude but highly reliable automation. It was not speculated why etiquette would interact with reliability but it may be that politeness affected an internal belief that artificially adjusted expectations of the automation that influenced attributions of responsibility (e.g. Marakas, Johnson, and Palmer 2000).

Thus, behaviour with anthropomorphic automation is affected by how it is perceived in addition to its reliability. The literature in computer-mediated communication has demonstrated the computers as social actors effect (e.g. stereotype elicitation, susceptibility to flattery) as well as the moderating influence of individual differences (e.g. cognitive style, ethnicity). Complementing these findings, the automation literature has shown that overt anthropomorphic elements (etiquette, human-like appearance) in automation can interact with machine-related factors such as automation reliability to influence trust and performance. The conceptual link between these two literatures is the finding that implicit attitudes about automation itself, or beliefs about the capabilities of automation held without conscious awareness, significantly affect trust in automation but only when reliability of the automation was uncertain (Lee and See 2004; Merritt, Heimbaugh, and LaChapell 2012).

Merritt, Heimbaugh, and LaChapell (2012) theorised that implicit general attitudes about automation affected the propensity to trust machines and an individual's trust in a specific automated system. Perceptions of the behaviour of any automation will be filtered through these explicit and implicit pre-existing beliefs about automation (Dzindolet et al. 2002). Merritt et al. found that when automation reliability was ambiguous, implicit beliefs about automation and stereotypes were more influential in determining trust than explicit beliefs. Presumably, in the face of ambiguity, individuals made attributions that were consistent with their implicit, schematic pre-existing beliefs about automation. This paralleled findings from the social cognition literature that stereotypic reasoning was common when an individual was faced with conflicting or ambiguous information (Kunda and Thaggard 1996).

Reframing the results of Parasuraman and Miller (2004) in light of the findings of Merritt, Heimbaugh, and LaChapell (2012), it may be that when automation performance was ambiguous/of low reliability participants fell back to their *newly* formed positive *implicit* beliefs about the automation (that the automation was polite), and the participants made more situational rather than dispositional attributions (i.e. attributed fault to the situation, not the automation). For the present study, Merritt et al's and Parasuraman and Miller's studies are crucial for several reasons. First, they showed that implicitly held beliefs influence explicit perceptions of trust in automation. Second, the implicit attitudes interacted with automation reliability to determine trust and behaviour. Factors at the person-level (stereotypes) and task-level (automation reliability) interacted to affect judgments and perceptions of technology. There is a wealth of research examining the role of etiquette on automation perceptions (Hayes and Miller 2011) but the current work extends the concept that another type of implicitly held perception (stereotypes) may affect how users perceive automation. The present study extended previous work on gender stereotypes on automation behaviour by examining another potential stereotype: age.

5. Overview of the study

Using participants in younger and older adult age groups, we collected judgments of trust of a simulated agent embedded within a decision aid that varied in gender, age, and reliability using a factorial survey with concrete health-related vignettes. Following the social cognition literature, we expected that age and gender stereotypes would most affect trust in the decision aid when system performance was ambiguous, but that there would be different effects for different age groups and genders of users. Specific research aims were as follows: (1) Determine the amount of variance in trust due to within-person variation

compared to between-person variation, (2) Determine how age of the agent, gender of the agent, and reliability of the decision aid agent affected judgments of trust in the aid, and (3) Determine how individual differences such as age and gender of the participant affected trust ratings of various decision aids. The results informed basic knowledge of how differing age and gender groups responded to stereotypes as well as informing the design of decision aids targeting particular groups of users.

We presented scenarios involving a decision aid (a smartphone ‘app’) for diabetes management via a factorial survey. The decision aid contained a simulated anthropomorphised agent. Factorial surveys have been widely used to examine how beliefs, judgments, and decision-making are influenced by situational factors (Rossi and Anderson 1982). Specific factors of the scenario were manipulated (in a factorial manner) and the participant rated all combinations of factors. The agent was a health-care provider offering advice on a specific diabetes-related dilemma. Because our dependent variable (trust) was a social judgment about a situation, a factorial survey was an ideal way to measure the influence of manipulated variables (age, gender, reliability of automation) as well as individual differences of the participants (Rossi and Anderson 1982; Hox, Kreft, and Hermkens 1991).

5.1 Method

5.1.1 Participants

Sixty younger adults and 47 older adults completed the study. The mean age of the younger group was 18.6 (SD = 0.9) while the older group was 72.7 (SD = 5.3). Younger adults were undergraduate college students whereas older participants were independently living, community-dwelling older adults. The younger participants chose to receive either course credit or \$7 per hour and the older participants received \$7 per hour. Descriptive statistics of participant characteristics are shown in Table 1.

5.1.2 Materials

Equipment. PC-compatible (Windows 7) computers running at 3.2 GHz with 4 GB of RAM were used with a 19-inch (48.3-cm) LCD monitor set at a resolution of 1024 × 1280 pixels. Participants were seated approximately 18 inches from the monitor and interacted primarily with a mouse (on the preferred side) and a keyboard.

Individual difference measures. In addition to participant age group and gender, we were interested in two individual difference measures: automation complacency and prior diabetes knowledge. The Complacency Potential Rating Scale (CPRS; Singh, Molloy, and Parasuraman 1993) is a 16-item scale designed to measure complacency towards common types of automation (e.g. automated teller machines). Participants responded to the extent they agreed with statements about automation on a scale of 1–5. The CPRS score was a sum of these responses and ranged from 16 (low complacency potential) to 80 (high complacency potential). We were primarily interested in CPRS to compare our sample to other studies that show higher complacency potential in older adults (e.g. Ho, Wheatley, and Scialfa 2005). Diabetes knowledge was assessed with the Diabetes Knowledge Test (DKT; Fitzgerald et al. 1998). The 23 questions of the DKT assessed basic knowledge about diabetes and diabetes management. Computerised versions of both the CPRS and DKT were used in this study.

Task. In a factorial survey, independent variables are called dimensions. The dimensions are orthogonal and can have multiple levels. Orthogonal dimensions allowed us to disentangle the unique effects of each dimension on judgments of trust. Our dimensions were agent gender (male, female), agent age (younger, older), and aid reliability (low, medium, high).

Table 1. Participant characteristics by age group and gender.

	Younger adults ($n = 60$)				Older adults ($n = 47$)			
	Female ($n = 37$)		Male ($n = 23$)		Female ($n = 25$)		Male ($n = 22$)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Age	18.49	0.72	18.74	1.15	72.00	5.29	73.45	5.27
CPRS ^{a*}	43.73	3.83	43.00	5.38	48.52	5.31	46.09	4.04
Diabetes knowledge ^{b*}	11.68	2.02	11.48	2.52	14.24	2.81	13.41	2.84

*Significant age group difference, $p < 0.05$ (no significant gender differences).

^a Scores could range from 16 indicating low complacency potential to 80 indicating high complacency potential (Singh, Molloy, and Parasuraman 1993).

^b The DKT scores could range from 0 indicating no knowledge to 23 indicating high knowledge (Fitzgerald et al. 1998).

Table 2. Dimensions (independent variables) of interest and resulting scenarios.

Scenario	Agent age (2)	Agent gender (2)	Stated reliability (3)
1	Young	Female	45%
2	Young	Female	70%
3	Young	Female	95%
4	Young	Male	45%
5	Young	Male	70%
6	Young	Male	95%
7	Older	Female	45%
8	Older	Female	70%
9	Older	Female	95%
10	Older	Male	45%
11	Older	Male	70%
12	Older	Male	95%

Note: Each scenario was presented twice resulting in 24 unique vignettes.

The dimensions of interest, their levels, and the factorial combinations resulting in 12 possible scenarios are shown in Table 2. Each scenario was replicated twice to create 24 unique vignettes. This resulted in 12 measurements of each dimension per participant. In their review of the literature, Wickens and Dixon (2007) proposed that an automation reliability of about 70% represented a critical inflection point; less than about 70% reliable was not relied upon while reliabilities higher than 70% led to complacency. For this reason, we chose high and low values that were well above and below 70% (45%, 70%, and 95%) to represent low, medium, and high reliabilities, respectively. Participants never actually experienced the levels of automation reliability; they were only told the past reliability of the particular app that was shown. No matter the stated past reliability of an app, the advice given by the app in every scenario was correct.

The possible combinations of agent age and gender are shown in Figure 1. An example vignette (containing older female, high reliability) is illustrated in Figure 2. The diabetes dilemma was presented in the upper left of screen. On the right, a diagnostic smartphone app gave a possible solution via an agent. The size of the smartphone was larger than actual size (approximately 30% larger) to be easily viewable from seated distance. Also, on the screen was a statement about the past reliability of the particular app (low, medium, or high). On the lower third, participants rated on a Likert scale their perception of trust and likelihood of following the advice of the aid.

The diabetes scenarios were used in a prior study (Pak et al. 2012) and were developed by adapting questions from a diabetes education workbook (Drucquer and McNally 1998), and reading diabetes support forums. They were designed to represent realistic scenarios that someone with Type II diabetes might experience. The presentation of the factorial survey was programmed in the Real Studio environment (Real Software 2013).

5.1.3. Design and procedure

The study was a 2 (age group of respondent: younger, older) \times 2 (gender of respondent: male, female) \times 2 (agent age: young, old) \times 2 (agent gender: male, female) \times 3 (aid reliability: low, medium, high) mixed-model design, with within-

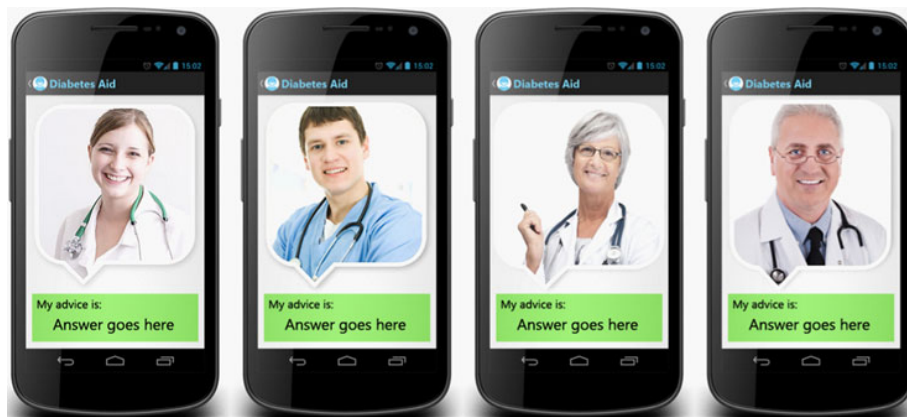


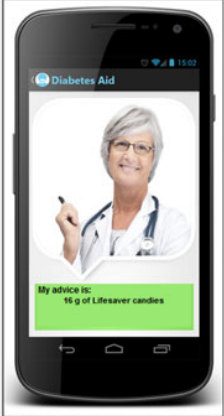
Figure 1. Illustration of the four possible smartphone agent conditions (young female, young male, older female, older male).

Remember: You are not trying to solve the problem below. You are giving your opinion on the smartphone app.

You were recently diagnosed with Type II diabetes and manage it with diet and medication. Your primary care doctor told you when your blood glucose gets low and you feel shaky, you should take a pinch of table sugar. However, you feel that taking a "pinch" of table sugar is not a precise enough measurement, so you want to eat something else. Your doctor approves this and reminds you that you need approximately 15 g of carbohydrates to substitute for one pinch of table sugar.

	2% Milk	Lifesaver Candy	Tropicana Orange Juice	Nature's Own White Bread
Serving Size	250 g	24 g	250 g	59 g
Calories	130	90	110	110
Total Fat	5 g	0 g	0 g	0.5 g
Saturated Fat	3 g	0 g	0 g	0 g
Cholesterol	20 mg	0 mg	0 mg	0 mg
Total Carbohydrate	13 g	24 g	26 g	25 g
Sugars	12 g	22 g	22 g	2 g

What should you eat instead of sugar to adjust your blood glucose levels?



Past reliability of THIS app's advice has been: **95%**

How much do you trust the smartphone helper?

1 Not at all 2 3 4 Neutral 5 6 7 Completely

Please BRIEFLY explain your rating in 100 characters or less.

How likely are you to follow the smartphone helper's recommendation?

1 Not at all 2 3 4 Neutral 5 6 7 Completely

Please BRIEFLY explain your rating in 100 characters or less.

Next

Figure 2. Image of the factorial survey response screen.

participant factors manipulated in the factorial survey. The first two variables (age group and gender of respondent) were quasi-independent grouping variables while the last three were within-groups manipulations of the decision aid and agent. The dependent variables were trust, likelihood of following advice, and diabetes knowledge.

Participants first completed a diabetes knowledge questionnaire administered on a computer. Next, participants started the factorial survey and were told:

You are playing the part of a newly diagnosed diabetic. Your doctor has given you a variety of different smartphone apps that may help you with your diabetes care. Your task involves giving us your opinion of the different smartphone apps. Just like many technological aids, the different apps will only sometimes seem reliable. Your performance is not being tested so you do not have to try to solve every problem. Instead, you are making judgments of the smartphone apps as quickly as possible.

After acknowledging the instructions and answering any remaining questions they began the survey.

In the survey, participants viewed a randomly presented vignette and were asked the following questions: (1) how much they trusted the smartphone app on a scale from 1 (not at all) to 7 (very much), and (2) whether they would follow or actually use the advice of the app (1–7). After the trust and decision aid usage questions, participants were also asked to briefly explain their ratings. To reinforce the notion that the smartphone app was a real decision aid and not just a pre-computed image, the smartphone app did not reveal its answer for 1.5 seconds (in the interim the message, 'Analysing the scenario. Just a moment . . .' appeared on the smartphone screen). After responding to 24 vignettes, participants completed the CPRS. Finally, participants answered the question, 'What do you think the study was about?' to assess whether they were aware of the purpose of the study. None of our participants were able to accurately state the purpose of the study other than what was told to them in the instructions (evaluating different apps). Because the trust and likelihood to follow ratings were highly correlated ($r = 0.83$, $p < 0.05$) only trust ratings were analysed.

5.2 Results

5.2.1 Hypothesised model

To answer our original research questions a two-level hierarchical model assessed the effects of agent gender and age, decision aid reliability, and diabetes knowledge on perceptions of trust in the decision aid. To review, our questions were (1) How is trust in an anthropomorphic decision aid affected by a user's age and gender?, (2) How is trust in the smartphone app affected by its appearance and reliability?, and (3) How is trust affected by domain knowledge?

Multiple responses were nested within the 107 participants: Each participant judged 24 vignettes resulting in a total of 2568 judgments for analysis. These judgments were nested within the manipulations performed on the survey (agent age, agent gender, reliability), which were in turn nested within the attributes of the participant (participant age, participant gender, diabetes knowledge score, CPRS score). Multi-level modelling was implemented through SAS, version 9.2.

Multi-level models are appropriate for data that exhibit hierarchical structure as they account for variability between and within participants and allow for examination of cross-level interactions (Raudenbush and Bryk 2002). Because respondents repeatedly made judgments on varying vignettes, those judgments of trust were not independent of each other; in fact, they were highly likely to be correlated which violates the independence of error variances assumption of analysis of variance (ANOVA) and regression (Hox and Bechger 1998; Tabachnick and Fidell 2007). There were also likely to be correlations between different levels (response level, group level). For example, trust responses on a vignette would likely be correlated to the responders group (gender, age group). That is, males may have a different stereotype than females (or older respondents versus younger ones) that they applied to the situation. Ignoring this hierarchical structure, or nesting, (i.e. by using ordinary least squares regression) can lead to an inflated Type I error rate, or detecting effects when there are none (Tabachnick and Fidell, 2007). Multi-level modelling solves this problem by allowing intercepts and slopes between levels to vary. Variability at one level is treated as a dependent variable at the next level. Hoffman and Rovine (2007) provided an accessible description of the usefulness of multi-level linear models in experimental psychology and human factors and Hox, Kreft, and Hermkens (1991) detailed why multi-level modelling is preferred for the analysis of factorial surveys.

A fully unconditional (non-multivariate) model (Model 1) was used to discover the amount of variance in trust found within participants at the survey level (Level 1; variance due to app appearance) and the amount of variance at the person level (Level 2; variance due to individual differences). This model represented a baseline to assess the fit of subsequent multivariate models (Models 2 and 3; equations in Appendix). Results (Table 3) revealed significant variance at both levels, with 94% of the variance at the survey level ($\sigma^2 = 3.04$, $z = 35.08$, $p < 0.0001$) and 6% of the variance at the person level ($\tau_{00} = 0.19$, $z = 4.39$, $p < 0.0001$).

Model 2 examined the effects of the survey manipulations on judgments of trust: agent gender, agent age, reliability, and all Level 1 interactions. Results revealed significant effects for all survey manipulations. Participants trusted male agents

Table 3. Unstandardised coefficients of multi-level models of the within and between-person effects of predictors on trust.

	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	4.75***	0.05	3.52***	0.11	3.43***	0.13
<i>Between-person</i>						
Age Group					0.35*	0.14
Gender					-0.13	0.12
Diabetes knowledge score					-0.06**	0.02
CPRS					-0.01	0.01
<i>Within-person</i>						
Agent gender			0.67***	0.14	0.67***	0.14
Agent age			0.38**	0.14	0.38**	0.14
Reliability of agent			1.09***	0.08	1.09***	0.08
Agent gender × agent age			-0.42*	0.20	-0.42*	0.20
Agent gender × reliability			-0.43***	0.11	-0.44***	0.12
Agent age × reliability			-0.36***	0.11	-0.19	0.12
Agent gender × agent age × reliability			0.47**	0.15	0.47**	0.15
<i>Cross-level</i>						
Age group × agent age group × reliability					-0.35***	0.09
Gender × agent gender × reliability					0.09	0.09
Age group × agent gender × reliability					-0.06	0.09
Gender × age group × reliability					-0.04	0.09
R^2 within-person			16.02		16.55	
R^2 between-person			<0.01		<0.01	
<i>Random effects</i>						
σ^2	3.04***	0.09	2.56***	0.07	2.54***	0.07
τ_{00}	0.19***	0.04	0.21***	0.04	0.20***	0.04

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. All between-person predictors were grand-mean centred.

more than female ones, older agents more than younger ones, and more reliable apps than less reliable ones. However, multiple significant interactions further refined this story. The three-way interaction of agent gender, agent age, and app reliability was significant – illustrated in Figure 3 – such that when the app was of low reliability, the younger female agent was trusted significantly less than the younger male aid, $F(1,1272) = 24.64, p < 0.05, \eta_p^2 = 0.2$, although there were no significant differences of agent gender for the younger agent at other reliability levels. For the older aid, the female agent was rated as less trusted, but this difference occurred only at the medium reliability level, $F(1,1272) = 13.91, p < 0.05, \eta_p^2 = 0.01$. These findings are consistent with our hypothesis that stereotypes would affect trust judgments when the reliability of a system was ambiguous (i.e. low or medium reliability).

A third model was conducted to include the individual difference predictors of participant age group, participant gender, CPRS, and diabetes knowledge and to examine hypothesised cross-level interactions. Our hypothesis was that participant

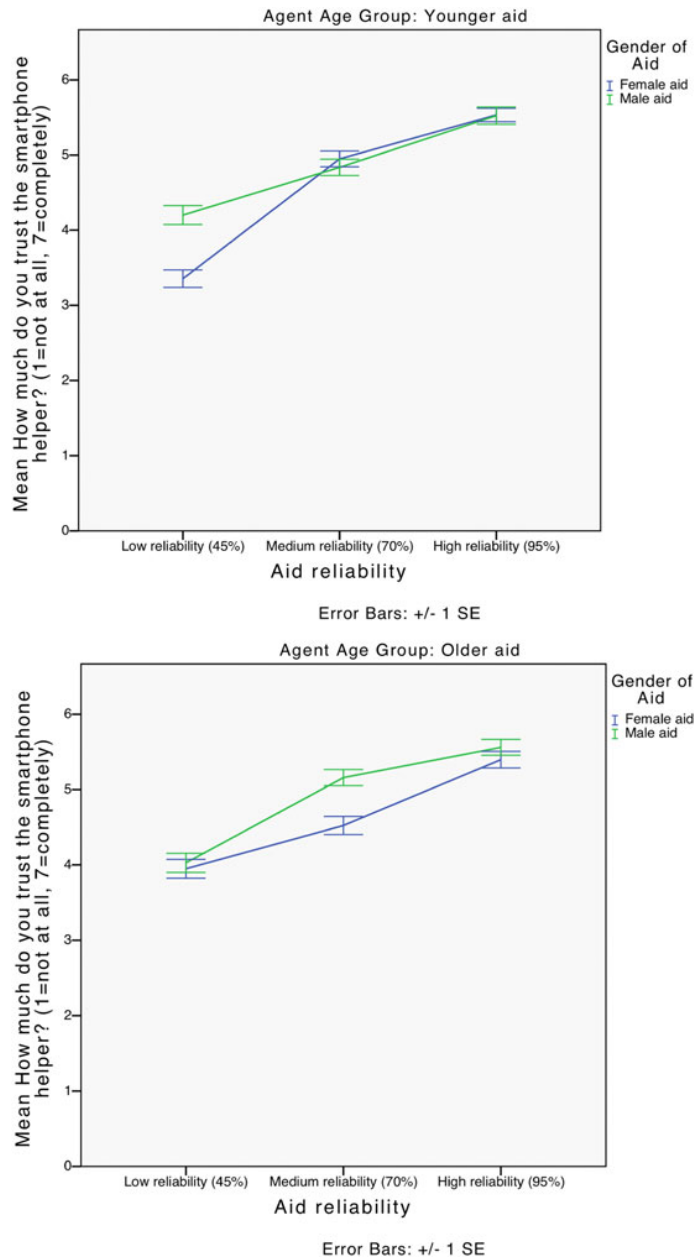


Figure 3. Three-way interaction of agent age group, agent gender, and reliability (from Model 2).

age group would interact with the age of the agent to differentially affect trust. The similarity-attraction hypothesis (Byrne 1971) would predict that the user's trust would be highest with agents that appear similar to them, particularly in age-appearance. We examined all cross-level interactions in Model 3.

In Model 3, those with higher diabetes knowledge rated the agents as less trusted overall. Older participants generally rated the agents as more trusted than did younger participants. This may be a manifestation of the generally higher complacency that older adults have with automation than younger adults (Ho, Wheatley, and Scialfa 2005). Gender of the participant and CPRS score had no effect on trust ratings. By entering these variables in the model they were controlled for when examining the cross-level interactions. Using the Akaike's information criterion, Model 3 was determined to better fit the data than Model 2 (it accounted for variance beyond Model 2). The three-way interaction among participant age group, agent age, and app reliability was significant (Figure 4). The source of the interaction was that younger adults in the low reliability condition tended to trust older agents significantly more than younger agents, $F(1,1434) = 16.88, p < 0.05$,

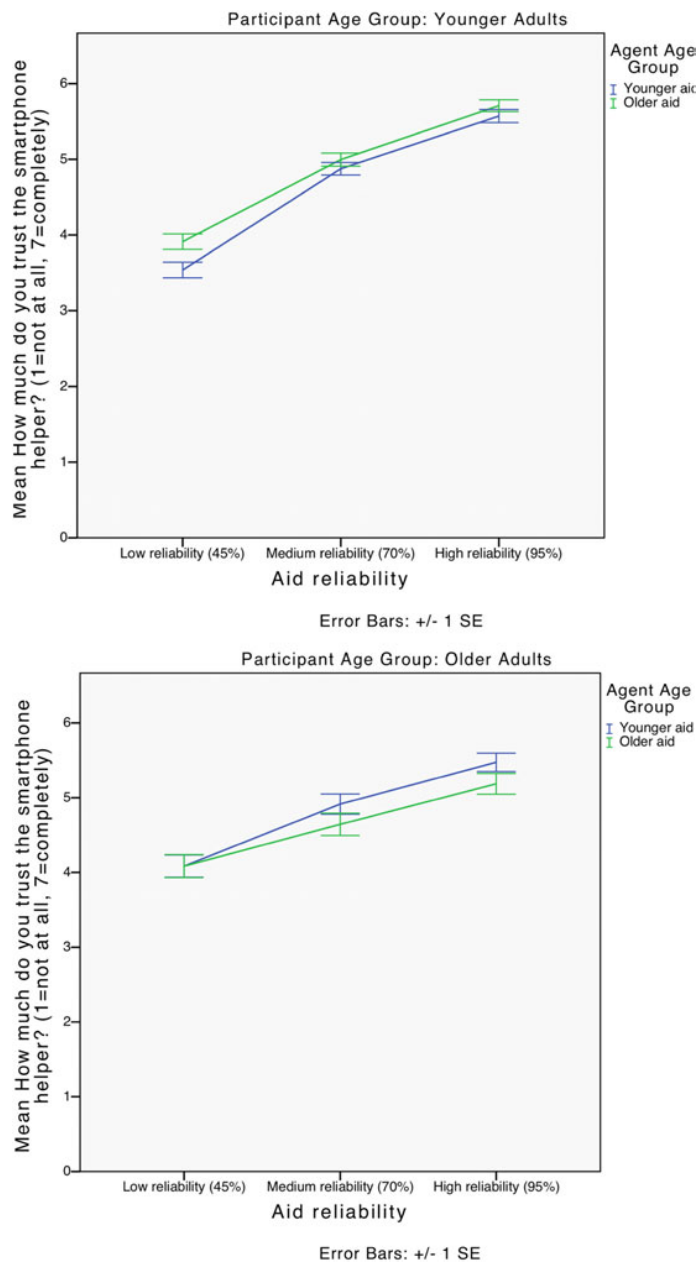


Figure 4. Cross-level interaction of participant age group, agent age group, and reliability (from Model 3).

$\eta_p^2 = 0.006$. There was no significant difference in trust by younger adults as a function of agent age in the medium or high reliability conditions. For older adults, there was no significant difference in trust as a function of agent age in any of the reliability conditions. Finally, to more directly test the possibility, presented in the introduction, that older adults may specifically hold negative stereotypes of young female agents, we examined the four-way interaction of agent age, agent gender, age group, and gender and found it to be not significant.

6. General discussion

As automation in consumer products and systems embodies human-like traits (e.g. anthropomorphic agents), stereotypes that users hold of age and gender may play an important role in trust and use of that automation. Prior research established that people apply gender stereotypes to computers but the purpose of this study was to examine if powerful and pervasive age stereotypes, as well as gender stereotypes, would be applied to anthropomorphic agents.

The finding that trust varies with reliability is not surprising; with higher levels of perceived reliability, users, particularly older adults, may become complacent (Mouloua et al. 2002; Ho, Wheatley, and Scialfa 2005). What is surprising is that this relationship between trust and complacency interacts with attributes of technology and individual differences in a way that is roughly consistent with the stereotype literature, specifically, age and gender stereotypes of doctors. However, perceived age group and gender of the agent and its reliability moderated the application of stereotypes (Model 2). When the agent appeared young, male agents were more trusted than female agents only when reliability was low. This gender difference disappeared at other levels of reliability. This pattern might suggest that unless the reliability of the system is catastrophically low (45%), most participants do not exhibit gender stereotypic thinking; perceptions of trust are primarily driven by reliability. However, when the reliability is very low, participants clearly shift to more stereotypic thinking and seem to attribute low performance to gender.

When the agent appeared older, male agents were more trusted than female agents only at medium levels of reliability. That is, stereotypic judgments appear at more moderate levels of reliability (70% versus 45%) if the aid is older rather than younger. The finding of gender stereotypic effects at 45% reliability when the agent is young, but at 70% when the agent is old seems to suggest that older female agents are judged more harshly than younger female agents. Given this finding one design recommendation is that when it is crucial for users to maintain high levels of trust in imperfect automation, a younger male agent is optimal because it seems less susceptible to large fluctuations in perceptions of trust as a function of gender (i.e. gender stereotypic thinking). More specifically, if it is undesirable to have users exhibit gender differences (or bias) in trust then using younger agents was preferable to older agents. A male agent was recommended over female because trust in female agents appeared more erratic as a function of reliability compared to male agents (e.g. the steep plunge in trust at 45% reliability for young females). However, this design recommendation does not take into account the gender or age group of the user. As the significant cross-level interaction of Model 3 shows, individual differences also seem to interact with the agent characteristics.

Model 3 showed that some anthropomorphic aspects of the aid did interact with participant individual differences to affect trust. Younger adults in low reliability conditions tended to trust older agents over younger agents while older adults did not show any significant differences in trust as a function of agent age. Based on Model 3, if the goal is to maintain high levels of trust in imperfect automation in young adult users, older agents (regardless of agent gender) are preferred. For older adult users, there was no significant difference in trust as a function of agent age group. However, there did appear to be a trend towards higher trust of younger agents with increasing reliability so for older users, a young agent may be optimal.

One caveat is that we did not assess *a priori* the pre-existing stereotypes held by our participants (as such an assessment might have influenced their behaviour in the experiment.) However, the stereotype literature is replete with research that shows the pervasiveness of the ‘warm but not competent’ stereotype of older adults not only in the USA but worldwide (Cuddy, Norton, and Fiske 2005). Another limitation is the use of a diabetes scenario. Although none of the participants in our study reported having diabetes, older adults may be more aware of diabetes simply because it is more common in their cohort than among younger adults (26.9% versus 11.3%, respectively; American Diabetes Association, 2011). Thus, simply being in a cohort that is more affected by diabetes may influence how one perceives diabetes advice. Another limitation was that because we assessed subjective perceptions of the automation (trust) because it is uncertain if trust translates to behaviour. However, past research has shown that perceptions of trust in automation are strongly correlated with behaviour (e.g. Lee and Moray 1994).

Acknowledgements

We thank Meghan Goodwin and Kayla Brennan for their help in collecting data and Peg Tyler for her help in manuscript preparation.

Funding

This research was supported by a grant from the Air Force Office of Scientific Research [award number FA9550-12-1-0385].

Notes on contributors

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

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

Brock Bass received his MS in Human Factors from Clemson University.

References

- American Diabetes Association. 2011. "Statistics About Diabetes." Retrieved from <http://www.diabetes.org/diabetes-basics/statistics/?loc=db-slabnav>
- Banaji, M., and C. D. Hardin. 1996. "Automatic Stereotyping." *Psychological Science* 7 (3): 136–141.
- Banaji, M. R., C. Hardin, and A. J. Rothman. 1993. "Implicit Stereotyping in Person Judgment." *Journal of Personality and Social Psychology* 65 (2): 272–281.
- Bickmore, T. 2011. "Etiquette in Motivational Agents." In *Human-Computer Etiquette: Cultural Expectations and the Design Implications They Place on Computers and Technology*, edited by C. C. Hayes and C. A. Miller, 206–226. Boca Raton: Auerbach.
- Byrne, D. E. 1971. *The Attraction Paradigm (Vol. 11)*. New York: Academic Press.
- Casper, C., K. Rothermund, and D. Wentura. 2011. "The Activation of Specific Facets of Age Stereotypes Depends on Individuating Information." *Social Cognition* 29 (4): 393–414.
- Cleveland, J. N., and G. Hollmann. 1990. "The Effects of the Age-Type of Tasks and Incumbent Age Compositions on Job Perceptions." *Journal of Vocational Behaviour* 36 (2): 181–194.
- Cuddy, A. J., M. I. Norton, and S. T. Fiske. 2005. "This Old Stereotype: The Pervasiveness and Persistence of the Elderly Stereotype." *Journal of Social Issues* 61 (2): 267–285.
- Czaja, S. J., and J. Sharit. 1998. "Age Differences in Attitudes toward Computers." *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 53 (5): 329–340.
- DeArmond, S., M. Tye, P. Y. Chen, A. Krauss, D. Apryl Rogers, and E. Sintek. 2006. "Age and Gender Stereotypes: New Challenges in a Changing Workplace and Workforce." *Journal of Applied Social Psychology* 36 (9): 2184–2214.
- Devine, P. G. 1989. "Stereotypes and Prejudice: Their Automatic and Controlled Components." *Journal of Personality and Social Psychology* 56 (1): 5–18.
- Drucquer, M. H., and P. G. McNally. 1998. *Diabetes Management: Step by Step*. Osney Mead, Oxford: Blackwell Science.
- Dzindolet, M. T., S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck. 2003. "The Role of Trust in Automation Reliance." *International Journal of Human Computer Studies* 58 (6): 697–718.
- Dzindolet, M. T., L. G. Pierce, H. P. Beck, and L. A. Dawe. 2002. "The Perceived Utility of Human and Automated Aids in a Visual Detection Task." *Human Factors* 44 (1): 79–94.
- Finkelstein, L. M., and M. J. Burke. 1998. "Age Stereotyping at Work: The Role of Rater and Contextual Factors on Evaluations of Job Applicants." *The Journal of General Psychology* 125 (4): 317–345.
- Fiske, S. T. 1998. "Stereotyping, Prejudice, and Discrimination." In *The Handbook of Social Psychology*, edited by D. T. Gilbert, S. T. Fiske, and G. Lindzey. 4th ed. New York: McGraw-Hill.
- Fitzgerald, J. T., R. M. Anderson, M. M. Funnell, R. G. Hiss, G. E. Hess, W. K. Davis, and P. A. Barr. 1998. "The Reliability and Validity of a Brief Diabetes Knowledge Test." *Diabetes Care* 21 (5): 706–710.
- Greenwald, A. G., and M. R. Banaji. 1995. "Implicit Social Cognition: Attitudes, Self-esteem, and Stereotypes." *Psychological Review* 102 (1): 4–27.
- Hayes, C. C., and C. A. Miller. 2011. *Human-Computer Etiquette: Cultural Expectations and the Design Implications They Place on Computers and Technology*. Boca Raton: Auerbach.
- Hilton, J. L., and W. von Hippel. 1996. "Stereotypes." *Annual Review of Psychology* 47 (1): 237–271.
- Ho, G., D. Wheatley, and C. T. Scialfa. 2005. "Age Differences in Trust and Reliance of a Medication Management System." *Interacting with Computers* 17 (6): 690–710.
- 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 Modeling." *Family Science Review* 11: 354–373.
- Hox, J. J., I. G. G. Kreft, and P. L. J. Hermkens. 1991. "The Analysis of Factorial Surveys." *Sociological Methods & Research* 19: 493–510.
- Johnson, D., J. Gardner, and J. Wiles. 2004. "Experience as a Moderator of the Media Equation: The Impact of Flattery and Praise." *International Journal of Human-Computer Studies* 61 (3): 237–258.
- Kite, M. E., K. Deaux, and M. Miele. 1991. "Stereotypes of Young and Old: Does Age Outweigh Gender?" *Psychology and Aging* 6 (1): 19–27.
- Kite, M. E., G. D. Stockdale, B. E. Whitley, and B. T. Johnson. 2005. "Attitudes toward younger and Older Adults: An Updated Meta-analytic Review." *Journal of Social Issues* 61 (2): 241–266.

- Kunda, Z., and B. Sherman-Williams. 1993. "Stereotypes and the Construal of Individuating Information." *Personality and Social Psychology Bulletin* 19 (1): 90–99.
- Kunda, Z., and P. Thagard. 1996. "Forming Impressions from Stereotypes, Traits, and Behaviours: A Parallel Constraint Satisfaction Theory." *Psychological Review* 103 (2): 284–308.
- Lee, E. J. 2003. "Effects of 'Gender' of the Computer on Informational Social Influence: The Moderating Role of Task Type." *International Journal of Human-Computer Studies* 58 (4): 347–362.
- Lee, E.-J. 2010. "The More Humanlike, the Better? How Speech Type and Users' Cognitive Style Affect Social Responses to Computers." *Computers in Human Behaviour* 26 (4): 665–672.
- Lee, J. D., and N. Moray. 1994. "Trust, Self-confidence, and Operators' Adaptation to Automation." *International Journal of Human Computer Studies* 40 (1): 153–184.
- Lee, J. D., and K. A. See. 2004. "Trust in Automation: Designing for Appropriate Reliance." *Human Factors* 46 (1): 50–80.
- Madhavan, P., D. A. Wiegmann, and F. C. Lacson. 2006. "Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids." *Human Factors* 48 (2): 241–256.
- Marakas, G. M., R. D. Johnson, and J. W. Palmer. 2000. "A Theoretical Model of Differential Social Attributions toward Computing Technology: When the Metaphor Becomes the Model." *International Journal of Human-Computer Studies* 52 (4): 719–750.
- McKinstry, B., and S. Y. Yang. 1994. "Do Patients Care about the Age of Their General Practitioner? A Questionnaire Survey in Five Practices." *The British Journal of General Practice* 44 (385): 349–351.
- Merritt, S. M., H. Heimbaugh, and J. LaChapell. 2012. "I Trust It, but I Don't Know Why: Effects of Implicit Attitudes toward Automation on Trust in an Automated System." *Human Factors* 55 (3): 520–534.
- Merritt, S. M., and D. R. Ilgen. 2008. "Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human Automation Interactions." *Human Factors* 50 (2): 194–210.
- Mosier, K. L., L. J. Skitka, M. D. Burdick, and S. T. Heers. 1996. "Automation Bias, Accountability, and Verification Behaviours." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 40 (4): 204–208.
- Mouloua, M., J. A. -A. Smither, D. A. Vincenzi, and L. Smith. 2002. "Automation and Aging: Issues and Considerations." *Advances in Human Performance and Cognitive Engineering Research* 2: 213–237.
- Nass, C., J. Steuer, and E. R. Tauber. 1994. "Computers Are Social Actors." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating Interdependence*, 72–78.
- Pak, R., N. Fink, M. Price, B. Bass, and L. Sturre. 2012. "Decision Support Aids with Anthropomorphic Characteristics Influence Trust and Performance in Younger and Older Adults." *Ergonomics* 55 (9): 1059–1072.
- Parasuraman, R., and C. A. Miller. 2004. "Trust and Etiquette in High-Criticality Automated Systems." *Communications of the ACM* 47 (4): 51–55.
- Perdue, C. W., and M. B. Gurtman. 1990. "Evidence for the Automaticity of Ageism." *Journal of Experimental Social Psychology* 26 (3): 199–216.
- Posthuma, R. A., and M. A. Campion. 2009. "Age Stereotypes in the Workplace: Common Stereotypes, Moderators, and Future Research Directions?" *Journal of Management* 35 (1): 158–188.
- Qiu, L., and I. Benbasat. 2009. "Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems." *Journal of Management Information Systems* 25 (4): 145–182.
- Raudenbush, S. W., and A. S. Bryk. 2002. *Hierarchical Linear Models*. 2nd ed. Thousand Oaks, CA: Sage.
- Real Software. 2013. [Computer software]. Austin, TX.
- Rosen, B., and T. H. Jerdee. 1976. "The Nature of Job-Related Age Stereotypes." *Journal of Applied Psychology* 61 (2): 180–183.
- Rossi, P. H., and A. B. Anderson. 1982. "The Factorial Survey Approach: An Introduction." In *Measuring Social Judgments*, edited by P. H. Rossi and S. L. Nock. Beverly: Hills Sage.
- 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 (1): 76–87.
- Röttger, S., K. Bali, and D. Manzey. 2009. "Impact of Automated Decision Aids on Performance, Operator Behaviour and Workload in a Simulated Supervisory Control Task." *Ergonomics* 52 (5): 512–523.
- Shah, R., and J. Ogden. 2006. "What's in a Face? The Role of Doctor Ethnicity, Age and Gender in the Formation of Patients' Judgements: An Experimental Study." *Patient Education and Counseling* 60 (2): 136–141.
- Singer, M. S. 1986. "Age Stereotypes as a Function of Profession." *The Journal of Social Psychology* 126 (5): 691–692.
- Singh, I. L., R. Molloy, and R. Parasuraman. 1993. "Individual Differences in Monitoring Failures of Automation." *The Journal of General Psychology* 120 (3): 357–373.
- Skitka, L., K. Mosier, and M. Burdick. 1999. "Does Automation Bias Decision-Making?" *International Journal of Human-Computer Studies* 51 (5): 991–1006.
- Tabachnick, B. G., and L. S. Fidell. 2007. "Multi-Level Linear Modeling." In *Using Multivariate Statistics*. 5th ed., 781–857. Boston, MA: Pearson.
- 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 (3): 201–212.
- Wrenn, K. A., and T. J. Maurer. 2004. "Beliefs about Older Workers' Learning and Development Behavior in Relation to Beliefs about Malleability of Skills, Age-Related Decline, and Control." *Journal of Applied Social Psychology* 34 (2): 223–242.
- Yee, N., J. N. Bailenson, and K. Rickertsen. 2007. "A Meta-Analysis of the Impact of the Inclusion and Realism of Human-Like Faces on User Experiences in Interfaces." In *Presented at the CHI 07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY: ACM.
- Zhang, T., B. Zhu, and D. B. Kaber. 2011. "Anthropomorphism and Social Robots." In *Human-Computer Etiquette: Cultural Expectations and the Design Implications They Place on Computers and Technology*, edited by C. C. Hayes and C. A. Miller, 231–255. Boca Raton: Auerbach.

Appendix. Multi-level model

Model 2:

Level 1: $TRUST_{it} = \beta_{0it} + \beta_{1it}(AgntGndr) + \beta_{2it}(AgntAge) + \beta_{3it}(Reliab) + \beta_{4it}(AgntGndr*AgntAge) + \beta_{5it}(AgntGndr*Reliab) + \beta_{6it}(AgntAge*Reliab) + \beta_{7it}(AgntGndr*AgntAge*Reliab) + r_{it}$

Level 2:

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

$$\beta_{1i} = \gamma_{10}$$

$$\beta_{2i} = \gamma_{20}$$

$$\beta_{3i} = \gamma_{30}$$

$$\beta_{4i} = \gamma_{40}$$

$$\beta_{5i} = \gamma_{50}$$

$$\beta_{6i} = \gamma_{60}$$

$$\beta_{7i} = \gamma_{70}$$

Model 3:

Level 1: $TRUST_{it} = \beta_{0it} + \beta_{1it}(AgntGndr) + \beta_{2it}(AgntAge) + \beta_{3it}(Reliab) + \beta_{4it}(AgntGndr*AgntAge) + \beta_{5it}(AgntGndr*Reliab) + \beta_{6it}(AgntAge*Reliab) + \beta_{7it}(AgntGndr*AgntAge*Reliab) + r_{it}$

Level 2:

$$\beta_{0i} = \gamma_{00} + \gamma_{01}(AGE) + \gamma_{02}(GENDER) + \gamma_{03}(DKS) + \gamma_{03}(CPRS) + u_{0i}$$

$$\beta_{1i} = \gamma_{10}$$

$$\beta_{2i} = \gamma_{20}$$

$$\beta_{3i} = \gamma_{30} + \gamma_{31}(AGE*GENDER)$$

$$\beta_{4i} = \gamma_{40}$$

$$\beta_{5i} = \gamma_{50} + \gamma_{51}(GENDER) + \gamma_{52}(AGE)$$

$$\beta_{6i} = \gamma_{60} + \gamma_{61}(AGE)$$

$$\beta_{7i} = \gamma_{70}$$

Low Memory, Mo' Problems: Effects of individual differences on types and levels of automation

Ericka Rovira¹, Richard Pak², Anne McLaughlin³

¹U.S. Military Academy Department of Behavioral Sciences & Leadership, West Point, NY 10996

²Clemson University Department of Psychology, Clemson, SC 29634

³North Carolina State University Department of Psychology, Raleigh, NC 27695

Running head: Individual Differences with Imperfect Automation

Word count: 5210

Corresponding author:

Ericka Rovira,
U.S. Military Academy
Department of Behavioral Sciences & Leadership
267 Thayer Hall/Engineering Psychology
MADN-BS&L/646 Swift Rd.
West Point, NY 10996, USA
e-mail: Ericka.Rovira@usma.edu

Acknowledgements

The second author was supported by a grant from the Air Force Office of Scientific Research (award number FA9550-12-1-0385).

APPENDIX 2: Rovira, E., Pak, R., & McLaughlin, A. C. (under review). Low Memory, Mo' Problems: Effects of individual differences on types and levels of automation. Human Factors.

Abstract:

Objective: We explored the extent to which individual differences in cognitive ability affected the use of types and levels of automation support in a complex decision-making task.

Background: Studies show performance benefits with reliable automation but performance costs with imperfect automation, particularly as automation support increases. Cognitive abilities are also critical to decision-making and correlate with automation reliance.

Method: We examined decision-making performance with varying types and levels of imperfect automation that supported 86 participants performing a simulated command and control task. Participants completed measures of attentional control and spatial working memory.

Results: Automation reliability, support, and task load interacted to affect accuracy. Additionally, working memory ability interacted with reliability and automation support. Reliable automation with increased automation support resulted in higher accuracies. When automation was imperfect, the reverse was true: increased automation support resulted in lower accuracy, especially for those with lower working memory ability. Those with higher working memory were less susceptible to the detrimental effects of increasingly supportive, but imperfect, automation. Further, lower working memory was associated with more trust in automation.

Conclusion: These results confirm the link between automation performance and individual differences, but also demonstrate the limits of the conventional wisdom that higher, reliable automation support unilaterally helps performance while higher, imperfect automation support harms performance.

Application: Optimizing human system performance requires understanding how individual variability contributes to performance with automation. These results may apply to the design of systems that accommodate individual differences in abilities through interface design and personnel selection.

Keywords: human automation interaction, types and levels of automation, individual differences, working memory, attention, trust, task load, and mental workload

Precis: The extent to which individual differences cognitive ability affected the use of imperfect types and levels of automation in complex decision-making was investigated. It was found that increased working memory capacity buffered the performance costs of imperfect decision automation and enhanced the benefit of automation support.

INTRODUCTION

Background

Commercial pilots are supported by sophisticated technology in the cockpit, soldiers are supported with automated targeting systems, drivers are supported by rear cameras cars, and many mobile phones use voice recognition software to assist users in searching for information. Each of these examples of automation could be characterized along two dimensions: what stage of information processing they support (type of automation: information acquisition, information analysis, decision-making, or action implementation) and how much they support the operator (level of automation: from a low level to a highly autonomous level; Sheridan & Verplank, 1978; Parasuraman, Sheridan, & Wickens, 2000).

A growing body of research has examined how human performance is differentially affected by various types and levels of highly reliable but imperfect automation (Crocoll & Coury, 1990; Endsley & Kaber, 1999; Galster, Bolia, & Parasuraman, 2002; Lorenz, Di Nocera, Röttger, & Parasuraman, 2002; Sarter & Schroeder, 2001; Wickens & Xu, 2002; Rovira, McGarry, & Parasuraman, 2007; Onnasch, Wickens, Li, & Manzey, 2014). The interest is motivated by the severe negative human performance consequences of *imperfect* automation such as: out of the loop unfamiliarity (Wickens, 1992), automation complacency (Parasuraman, Molloy, & Singh, 1993), loss of situation awareness (Endsley & Kiris, 1995), and skill degradation (Bainbridge, 1983).

In a meta analysis of 18 automation studies examining the differential effects of types and levels of automation, Onnasch et al. (2014) found performance benefits with reliable automation and performance decrements with higher types and levels of automation. Of most interest, were the decrements in performance found when automation support moved across the critical

boundary of information automation to decision automation. Thus, an important goal for designers is to mitigate performance costs associated with higher types and levels of automation by facilitating appropriate trust calibration (e.g., Rovira, Cross, Leitch, & Bonaceto, 2014). One approach is to better understand how individual differences in cognitive ability affect the appropriate use of imperfect types and levels of automation in complex decision-making tasks.

Individual Differences

Some early research has explored sources of individual differences and performance with automation (e.g., Singh, Molloy, & Parasuraman, 1993). However, these early investigations have focused on what could be considered personality characteristics (e.g., complacency potential; Singh et al., 1993). Another source of individual differences may be cognitive abilities specifically working memory capacity (Baddeley, 1986) and visuospatial attention (Gopher, 1982). It is also well established that working memory and attention are critical abilities that underlie effective decision-making (Lohse, 1997) and reliance on automation (Parasuraman & Manzey, 2010). Therefore, optimizing human system performance necessitates the assessment and understanding of how individual cognitive variability contributes to operational performance and automation usage.

In one of the earlier studies examining automation performance and individual differences in cognitive abilities, Chen and Terrence (2009) investigated the effects of imperfect automation and individual differences in a military multitask environment. Specifically they were interested if individual differences in perceived attentional control impacted how operators interacted with miss vs. false alarm prone automation. Attentional control was assessed using a survey that measured individuals' perceived attentional focus and shifting. They found that individuals with high perceived attentional control were more negatively affected by false

alarms, whereas for individuals with low perceived attentional control, miss prone automation was more harmful. In the context of their task (military gunner and robotics operator), perceived attentional control was clearly an important moderator of how operators reacted to automation false alarms and misses.

Individual differences in working memory also seem to play a role in mediating operator performance with automation. Parasuraman, de Visser, Lin, and Greenwood (2012) examined whether certain genotypes could predict an individual's susceptibility to automation bias; in other words operators adhering to imperfect automation. Researchers looked at two specific single nucleotide polymorphisms (SNPs) or variants of the DBH gene that regulate Dopamine (DA) and norepinephrine (NE). DA and NE levels are associated with DBH enzyme activity (low, high) that contributes to neural activity in the prefrontal cortex known to play a critical role in working memory ability. Using a command and control task (Rovira, et al., 2007), Parasuraman et al. (2012) varied the automation support (manual, reliable, and imperfect) that low and high DBH enzyme groups experienced. They found no difference between the low and high DBH enzyme groups with manual and reliable automation, but with imperfect automation individuals in the low DBH enzyme group performed better compared to individuals in the high DBH enzyme group. Parasuraman et al. (2012) attributed this effect to individual differences in working memory induced by enhanced DA availability in the low DBH enzyme group. However, because they did not measure working memory or other cognitive abilities, it is still unclear if individual differences in working memory interact with automation reliability to affect performance.

The importance of individual differences in working memory was examined in another study (de Visser, Shaw, Mohamed-Ameen, & Parasuraman, 2010). Researchers investigated the

role of working memory in an automated UAV task by varying task load (low, high) and automation reliability (manual, reliable, and imperfect). Participants completed both the Operation Span (OPSPAN) and Spatial Span (SSPAN) working memory tests (Engle, 2002). Researchers found a significant correlation with OSPAN scores and performance on the automated task. For each automation task performance measure, they found that linear models that included working memory accounted for more of the variance in performance as compared to the linear models without the individual differences OSPAN measure. Thus, when individual differences in working memory are accounted for, more variation in performance with automation can be explained. However, these researchers did not investigate types and levels of automation.

Research Hypotheses

This research was aimed at understanding the sources of performance differences underlying human-automation interaction with imperfect automation across different types and levels of automation. Many studies have investigated this same topic (for a review see Onnasch et al, 2014), however our work is distinct because it examines individual differences. In addition, the current research extended previous work in this area in two specific ways. First, we explicitly varied types and levels of imperfect automation and task load. Secondly, we more directly measured individual differences in cognitive abilities by using well-accepted working memory and visuospatial attention tasks compared to self-reported measures of abilities, complex proxy tasks (e.g., video game performance), or genetic predictors of cognitive performance. Finally, while evidence from a review of 20 automation reliability studies suggested that dependence on imperfect automation would be stronger with increased task demand (because the operator's limited resources are expended; Wickens & Dixon, 2005) this is

the first study that investigated the effects of individual differences in working memory and visuospatial attention on types and levels of imperfect automation and varying task demand.

We hypothesized that individual differences in working memory and visuospatial attention would differentially impact reliance on varying types and levels of automation.

Specifically:

1. First, consistent with previous literature, we hypothesized that:
 - a. operators would perform better with reliable automation compared to manual control.
 - b. there would be no difference between task load conditions when the automation was reliable.
 - c. the differential impact of information and decision automation would be evident with imperfect automation especially with high task load.
2. Second, as suggested by Parasuraman et al. (2012) we expected individuals with higher working memory capacity to show less of a decrement with higher forms of imperfect decision automation as compared to individuals with less working memory capacity. Specifically, with imperfect automation or high task load it was predicted that the benefits of better working memory capacity would be highlighted.
3. Third, we expected individuals with high visuospatial ability to perform better with high task demand and information automation as compared to individuals with lower levels of visuospatial ability. This would be interesting as researchers currently recommend lower types and levels of automation when 100% automation reliability cannot be guaranteed and return to manual control is of concern (Onnasch et al, 2014), however integrating large amounts of data may be difficult for some individuals.

4. Finally, we expected a relationship between variations in cognitive ability and self-report measures of trust. Specifically, individuals with low cognitive abilities would trust the automation more compared to individuals with high cognitive abilities.

METHODS

Participants

A total of 86 cadets (18 women) from the U.S. Military Academy volunteered and participated in this study for extra credit. Ages ranged from 18 to 24 ($M = 20.27$, $SD = 1.25$).

Stimuli and Task Procedures

Participants completed this study in two hours including training and breaks. Participants first completed two cognitive measures followed by a simulated artillery sensor-to-shooter targeting task. Response time and accuracy were collected for all measures. An anti-saccade task (Unsworth, Schrock, & Engle, 2004) was also administered to participants, but data loss prevented analysis and so it will not be discussed further.

Visuospatial Attention Task. A spatially cued letter discrimination task developed by Greenwood et al. (2000) was used to measure attentional control. First, a fixation point was displayed for 500 ms followed by a cue (an arrow pointing either left, right, or in both directions). The cue was either valid, predicting the subsequent target location on 61.5% of the trials, invalid on 15%, neutral on 15%, or no cue appeared on 8.5% of the trials. The location cue appeared for a cue–target SOA of 500 or 2,000 ms (Figure 1). Next, a letter target appeared to the right or left of the fixation point. Participants categorized the target letter as either a consonant or a vowel by using their index fingers to select one of two responses on a keyboard.

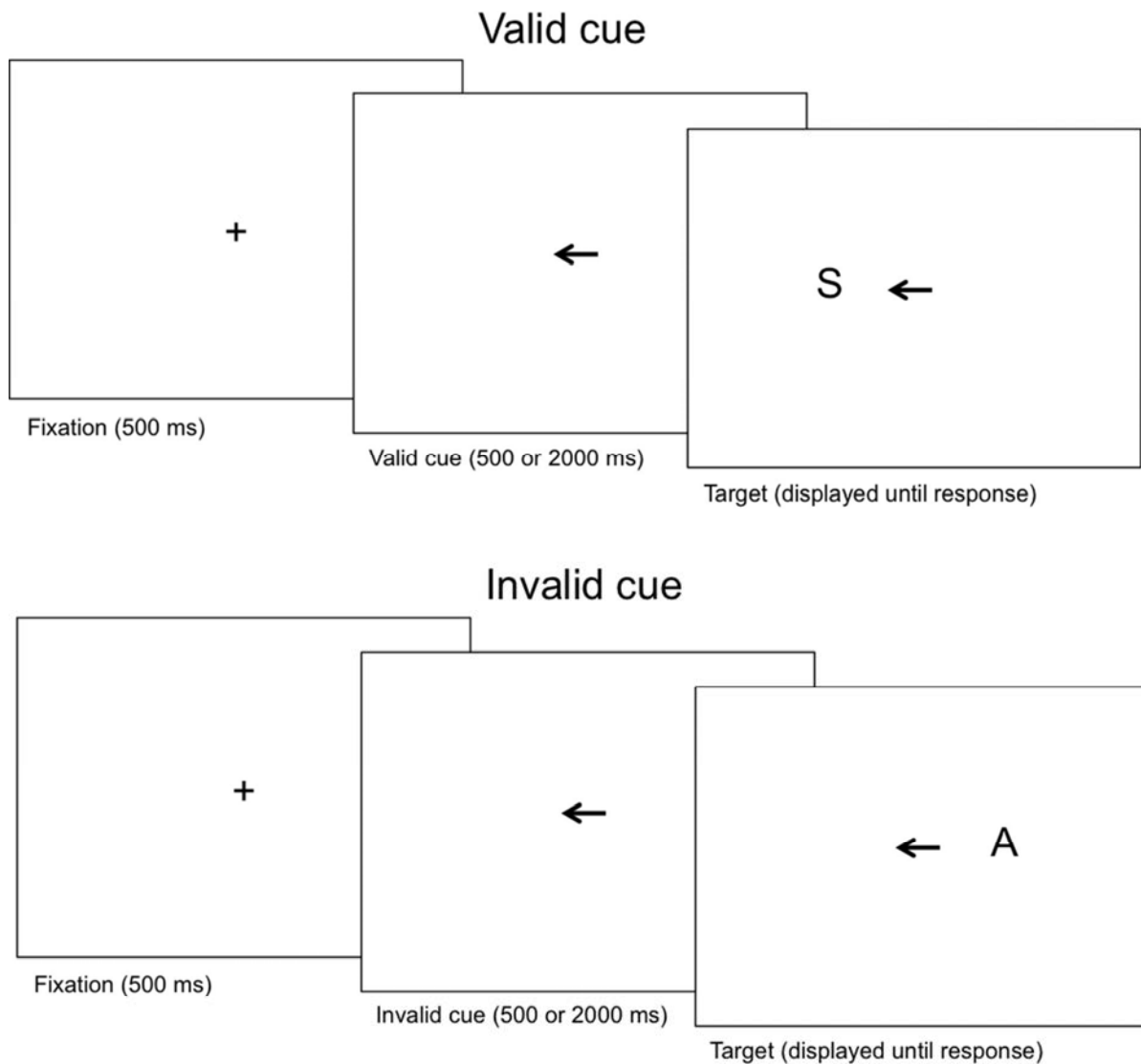


Figure 1. Visuospatial attention measure used to indicate individual differences in attentional control at various SOAs.

Working Memory Task. A spatial working memory task assessed working memory capacity (Figure 2; 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 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 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.

Because the working memory task generated several dependent variables (performance at different memory loads), a composite score was created consisting of accuracy on trials at three levels of memory load, in both match and non-match conditions. Z-scores were computed for each of the six conditions and a mean was taken to form a composite for each individual. Thus, this composite score was not standardized, but reflected the average of the standardized scores.

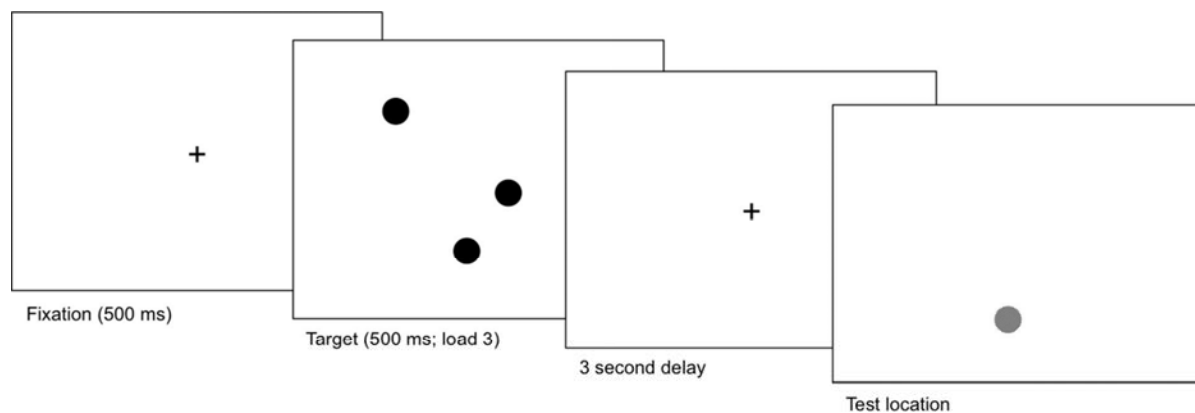


Figure 2. Working memory measure used to indicate spatial working memory capacity at 3 levels of load.

Artillery Sensor-to-Shooter Targeting Task

A low-fidelity software simulation of an artillery sensor-to-shooter targeting system was used with varying levels of automation support (Rovira et al., 2007). The artillery task consisted of three components in separate windows: a terrain view, a task window, and a communications module (Figure 3). A two-dimensional terrain view of a simulated battlefield displayed red enemy units (labeled 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). In the

task window, users made enemy-friendly engagement selections. The participants were required to identify the most dangerous enemy target and select a corresponding friendly unit to engage in combat with the target, known as an enemy-friendly engagement selection.

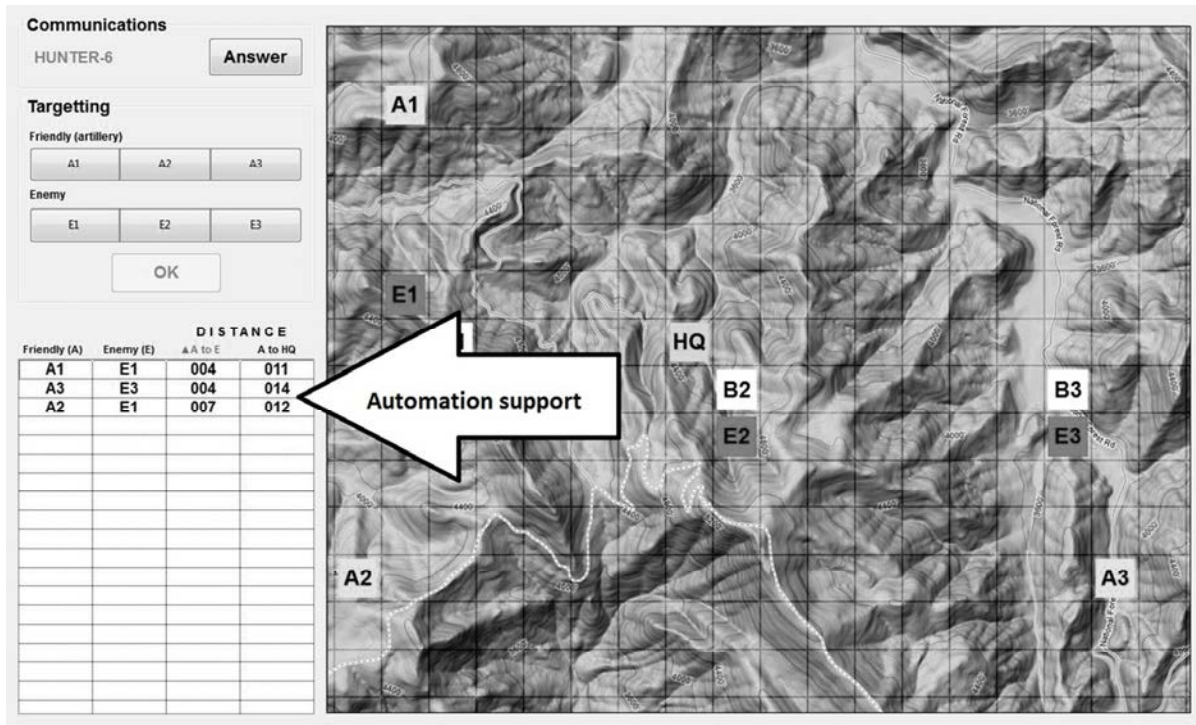


Figure 3. The sensor-to-shooter task interface, shown in the low task load, medium-decision automation condition.

The bottom left of the task window provided varying types and levels of automation support. The lowest support (above the fully manual condition) was *information automation*, which provided a list of all possible engagement combinations, including the distances between enemy targets, 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. The next level of automation, *low-decision automation*, gave a list of all possible engagement combinations, including the distances between enemy targets, friendly units, and headquarters; however, the listings were prioritized with the best selection first and the worst choice last, making this a form of decision automation. In the *medium-decision automation*

condition, the participant was provided the top three options for engagement, including the distances between enemy targets, friendly units, and headquarters (Figure 3). Unless the trial was imperfect, the first line was always the best enemy-friendly selection.

Participants could either follow the automation (select the top pairing in the ordered list) or make their own enemy-friendly unit engagement selection, but were required to make a decision within 10 s. Participants were able to cross-verify the automation by reviewing the terrain view. 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 overall difficulty of completing the sensor-to-shooter task, a random 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. Their call sign occurred randomly every 50 and 90 seconds.

Experimental Design

A 4 (Automation Support: manual, information automation, low-decision automation, medium-decision automation) x 2 (Task Load: low, high) x 2 (Trial Reliability: reliable, imperfect) within-subjects design was used. *Task load* was manipulated by increasing the number of friendly and enemy units from three to six each. *Trial reliability* was manipulated for each of the automation support conditions and referred to a correct automated assessment (reliable) versus an incorrect automated assessment (imperfect). Participants were informed that although the automation was highly reliable, it was not 100% reliable (actual overall reliability was 80%). However, no further information on reliability was given.

Each participant practiced with each of the eight conditions: manual at (1) low task load and (2) high task load; information automation at (3) low task load and (4) high task load; low-

decision automation at (5) low task load (6) high task load; and medium-decision automation at (7) low task load and (8) high task load. During practice, participants completed trials at both task load conditions for each level of automation support tool before a new level of automation support was introduced. After completing the practice trials, participants completed 8 blocks with each block representing a particular combination of task load (low, high; counterbalanced) or automation support (manual, information automation, low-decision, medium-decision; counterbalanced via partial Latin square). Each block consisted of 40 trials. In all, each participant completed 320 test trials.

Dependent variables included the accuracy and speed of enemy-friendly engagement selections. Accuracy was calculated by the percentage of trials in which the participant correctly selected the most dangerous enemy target and a corresponding friendly unit to engage in combat. Secondary task measures of performance included accuracy on the communications (call sign) task. To obtain subjective measures of mental workload, participants completed a computerized version of NASA-Task Load Index (TLX) after each block (Hart & Staveland, 1988). Participants also rated their trust in automation after each automation-present block (history-based trust) using an on-screen visual analog scale ranging from 0 to 100 (adapted from Lee & Moray, 1994) and at the completion of the study (dispositional trust; adapted from Jian, Bisantz, & Drury, 2000).

RESULTS

Repeated measures analyses of variance (ANOVAs) were conducted to evaluate effects of automation support, task load, and trial reliability on performance, subjective mental workload, and trust. Multilevel linear models were conducted to measure the role of individual differences in cognitive ability on task performance under the various manipulations.

Manual Control versus Automation Support

Decision-making accuracy was computed under manual control and automation support. For these analyses, we collapsed across the three forms of automation support (information, low-decision, and medium-decision) and then segregated by trial reliability (reliable, imperfect). Figure 4 shows that performance with reliable automation improved compared to manual and degraded with imperfect automation and high task load. A 3 (automation support: manual, reliable automation, imperfect automation) x 2(task load: low, high) repeated measures ANOVA revealed a main effect of automation type, $F(1,84)=272.7, p<.05, \eta_p^2 = .76$, task load $F(1,85)=82.0, p<.05, \eta_p^2 = .49$, and the interaction between the two, $F(1,84)=51.9, p<.05, \eta_p^2 = .38$.

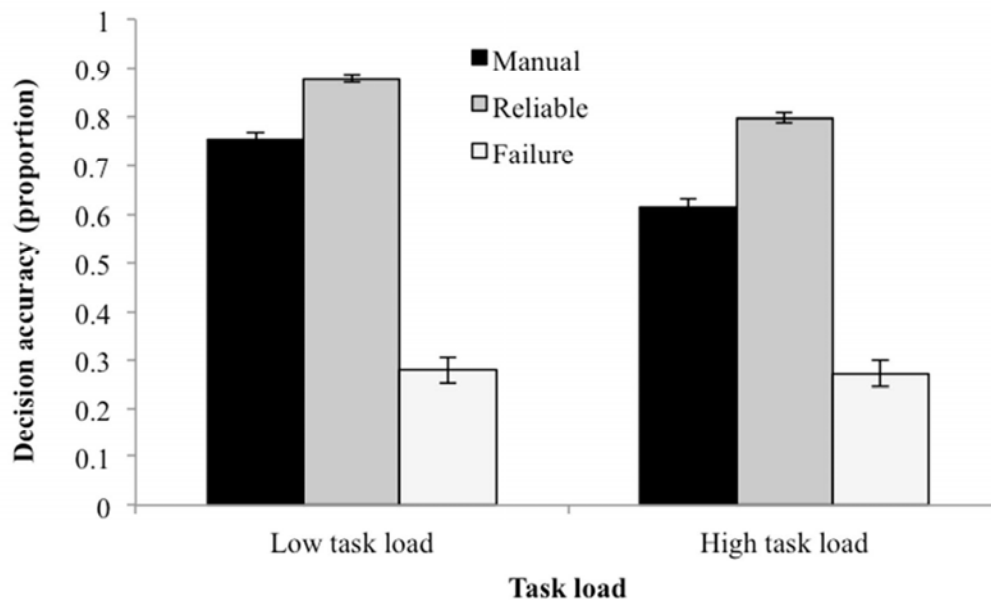


Figure 4. Decision making accuracy as a function of task load and automation support. Bars indicate standard error.

Pairwise comparisons showed the source of the interaction was due to performance decrements with increased task load with manual (low task load $M=.75, SD=.14$; high task load $M=.61, SD=.16$) and reliable automation (low task load $M=.88, SD=.07$; high task load $M=.79,$

$SD=.10$), but not with imperfect automation ($p<.05$). Due to space limits RT data was not included though it varied similarly to accuracy data and did not demonstrate a speed accuracy trade-off.

Multilevel Models

A two-level hierarchical model assessed the effects of the within-person variables of automation support, task load, automation reliability, the between-person predictor of working memory score, and their interactions on decision-making accuracy in the sensor-to-shooter task. It was expected that decision-making accuracy would be related to automation support, task load, reliability, and individual differences in working memory ability. Automation support was included as an interval-level variable in the model.

Multiple responses were nested within the 85 participants as each participant performed the sensor-to-shooter task under varying automation support, task load, and trial reliability. Accuracy represented the ratio of their correct to incorrect trials under each combination of those conditions. These scores were nested in the within-person manipulations (automation support, task load, reliability), which were in turn nested within the attributes of the participant (working memory ability). These nested observations were unlikely to be independent, violating the independence of error variances assumption of logistic regression (e.g., responses by a participant are likely to be correlated to that person's ability). MLMs are preferred over regression especially for within-subjects experimental designs that produce hierarchically structured data (Raudenbush & Bryk, 2002). Multivariate regression ignores this hierarchical structure, or nesting, which can lead to inflated Type I error rates (Hox & Bechger, 1998; Tabachnick & Fidell, 2007) while MLM allows each individual to act as his or her own control, accounts for variability between and within participants, and allow for examination of cross-level

Individual differences and automation use 16

interactions (Raudenbush & Bryk, 2002). Hoffman and Rovine (2007) provided an accessible discussion of the usefulness of multilevel linear models in human factors research. Multilevel modeling was implemented using PROC MIXED through SAS, version 9.4.

Table 1. Unstandardized Coefficients of Multilevel Models of the Within- and Between-person Effects of Predictors on Accuracy in a Sensor-to-Shooter task

Fixed effects	Model 1		Model 2		Model 3	
	Unconditional Model		Random Coefficients Regression		Slopes and Intercepts	
	Estimate	SE	Estimate	SE	Estimate	SE
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
<i>Cross-level</i>						
Task load x AutoSupp			-0.081 ***	0.024	-0.080 ***	0.024
Task load x Reliability			-0.501 ***	0.073	-0.502 ***	0.072
AutoSupp x Reliability			0.154 ***	0.024	0.159 ***	0.024
Task load x AutoSupp x Reliability			0.210 ***	0.034	0.209 ***	0.033
<i>Task load x WM</i>						
Task load x WM					-0.036	0.065
<i>AutoSupp x WM</i>						
AutoSupp x WM					0.011	0.026
<i>Reliability x WM</i>						
Reliability x WM					0.080	0.065
<i>Task load x AutoSupp x WM</i>						
Task load x AutoSupp x WM					0.013	0.030
<i>Reliability x AutoSupp x WM</i>						
Reliability x AutoSupp x WM					-0.089 **	0.030
Random Effects						
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
Model fit statistic						
A1C	972.2		-12.6		-28.6	

Note. * $p < .05$, ** $p < .01$, *** $p < .001$; Working memory composite score was grand-mean centered. SE indicates standard error.

Model 1: No predictors. A fully unconditional model (Table 1: Model 1) was first used to discover the amount of within and between –person variance in accuracy and provide a baseline to assess the fit of multivariate models (Models 2 and 3). The unconditional model revealed significant variance at both levels, with 97% of the variance at the within-person level ($\sigma^2 = 0.149$, $z = 21.48.08$, $p < .001$) and 3% of the variance at the person level ($\tau_{00} = 0.005$, $z = 1.81$, $p = .034$).

Model 2: Within-person variables. Model 2 examined the effects of the within-person manipulations on accuracy (Table 1). When using this model that included the within-subjects

manipulated variables, 67% of the 97% within-subject variance was accounted for. Model fit using the Akaike's information criterion (AIC) improved from 976.2 to -12.6 (lower values indicate better fit).

The model revealed a significant three-way interaction of automation support, task load, and trial reliability (Table 1). Data were divided into reliable and imperfect trials to examine the effects and interactions of automation support and task load and decompose the interaction. For *reliable* automation, pairwise comparisons showed that increased task load decreased accuracy only under information automation support (from $M=.70$, $SD=.19$ to $M=.44$, $SD=.25$; $p < .05$). This can be seen on the left panel of Figure 5 where accuracy in the information automation condition declined as task load increased while low and medium automation accuracy were unaffected. For trials with *imperfect* automation, pairwise comparisons showed that increasing task load significantly decreased accuracy only with medium-decision automation (from $M=.29$, $SD=.31$ to $M=.16$, $SD=.26$; $p < .05$). This decline in the medium-decision condition with increased task load can be seen on the right panel of Figure 5.

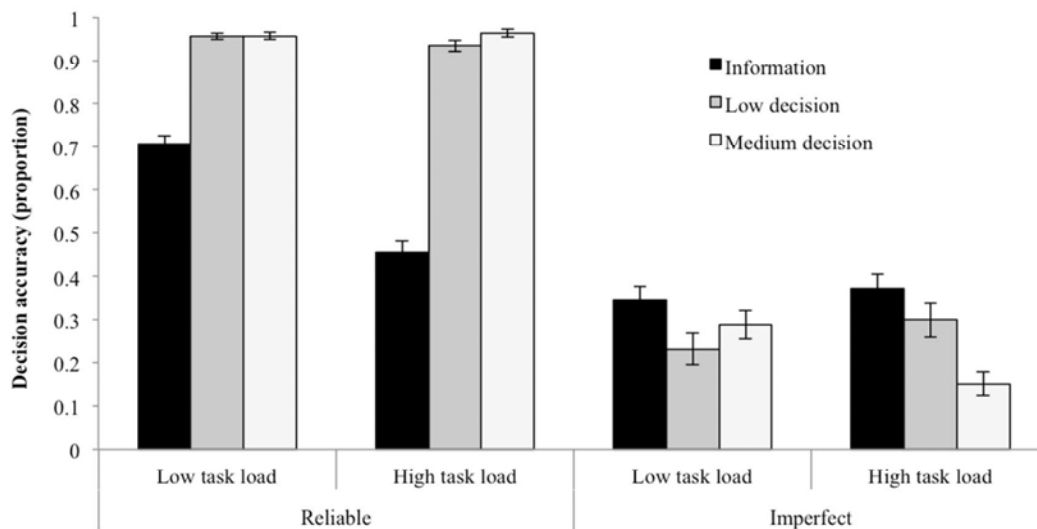


Figure 5. Decision accuracy as a function of trial reliability, task load, and automation support. Bars indicate standard error

In sum, the source of the 3-way interaction of reliability, automation condition, and task load appears to be: First, there was a large effect of reliability, where imperfect automation harmed performance. Further, imperfect automation harmed performance more with higher automation support while reliable automation improved performance with increasing automation support. Last, these effects were exacerbated by task load, which had less of an effect on performance when automation was reliable and the most effect on performance when automation was unreliable and at increased automation support.

Model 3: Cross-level interactions. We expected individuals with higher working memory capacity to show less of a decrement with higher forms of imperfect decision automation as compared to individuals with less working memory capacity. Specifically, with imperfect automation or high task load it was predicted that the benefits of better working memory capacity would be highlighted (equation available in appendix). A third model was conducted to include working memory ability to examine these hypothesized cross-level interactions. Attentional ability was not included in the model as it showed no correlation with accuracy ($r = -.071, p > .05$).

The model revealed a 3-way cross-level interaction of reliability, automation support, and working memory ability (Table 1). 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. To decompose the interaction, data were divided into reliable and imperfect trials to examine the effects and interactions of automation support and working memory (Figure 6). Task load was controlled for in these models.

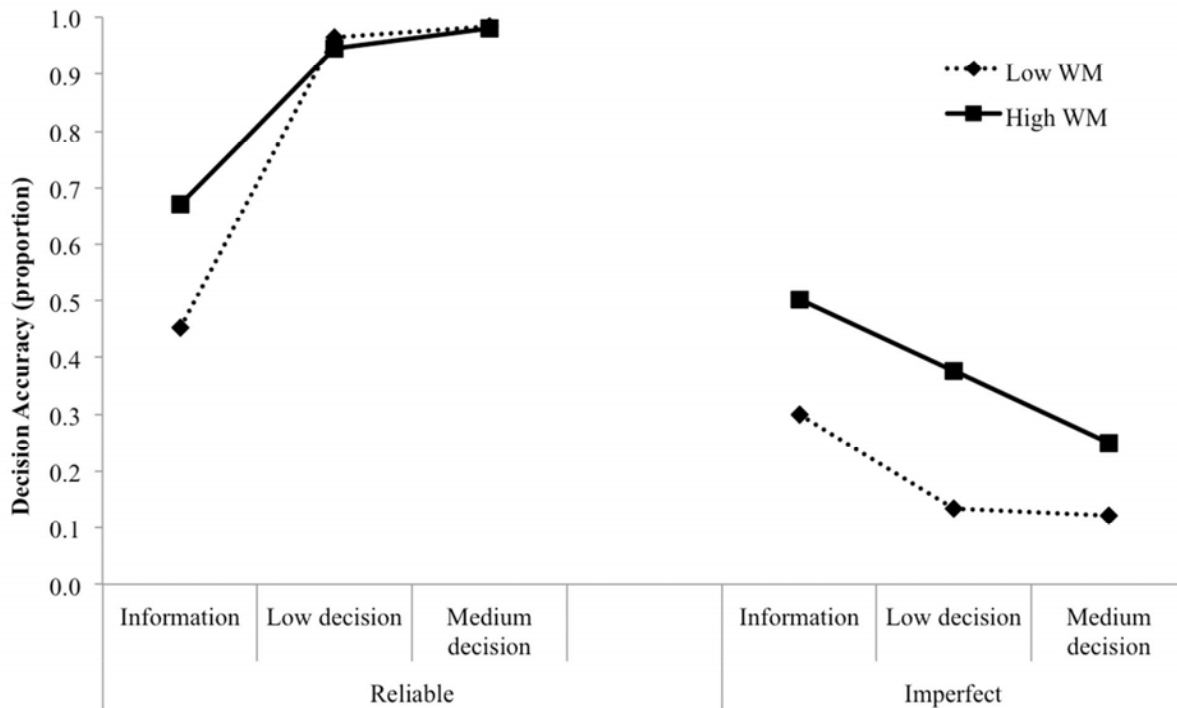


Figure 6. Proportion of correct responses across automation support. Points were plotted by calculating estimates based on low (1 SD below the mean) and high (1 SD above the mean) values for the composite working memory measure. Note that both groups had high accuracy overall with more reliable automation.

When automation was **reliable**, simple slopes analyses revealed that low and high working memory participants differed significantly from each other with varying automation support ($t(1,410)=-8.01, p<.001$ and $t(1,410)=-12.87, p<.001$). Significance contrasts also revealed that accuracy differed between low and high working memory participants for information automation support ($t(1,410)=-11.6, p<.001$) and the predicted data points for high decision support (e.g., beyond medium-decision support) ($t(1,410)=-17.57, p<.001$).

When automation was **imperfect**, those with high working memory were able to maintain some level of performance (~52% higher than low working memory participants), although just as those with lower working memory, their accuracy declined as automation support increased. A simple slopes analysis revealed that low and high working memory

participants differed significantly from each other at all types and levels of automation support ($ps < .001$) and accuracy declined as automation support increased ($ps < .01$).

In sum, working memory interacted with reliability and automation support to affect accuracy. When automation was reliable, increasing automation support resulted in higher accuracy for all participants. When automation was imperfect, the reverse was true: increasing automation support resulted in worse accuracy. This was especially true for those with lower working memory ability. These results show that those with higher working memory were less susceptible to the detrimental effects of increasingly supportive but imperfect automation.

Trust

History-based trust (after every block). The multivariate main effects of automation support and task load were significant, Wilks' lambda = .37, $F(8,66)=13.9$, $p < .05$, $\eta_p^2 = .63$, Wilks' lambda = .82, $F(4,70)=3.81$, $p < .05$, $\eta_p^2 = .18$. The interaction of task load and automation support 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.

Correlations between trust measures and abilities were computed to examine the effects of individual differences in working memory and attention on trust. In the information automation condition, having better attentional control (lower attention costs) was associated with more positive beliefs about automation (trust, reliance; $r = -.16$, $r = -.13$, respectively, all $ps < .05$). However, in the low decision automation, attention was no longer correlated but working memory was negatively correlated to trust such that working memory (higher WM scores) was associated with lower trust, reliance, and beliefs that automation improved

performance ($r = -.14$, $r = -.23$, and $r = -.13$, respectively, all $ps < .05$). Finally, in the medium-decision condition, working memory negatively correlated to reliance showing that higher working memory was associated with less self-reported reliance on automation ($r = -.14$).

Dispositional trust. Trust negatively correlated with working memory (lower working memory scores was associated with more agreement with positive statements about automation; $r = -.22$, $p < .05$) while distrust positively correlated with working memory (higher working memory was associated with more agreement with negative statements about automation; $r = .24$). Attention was not correlated with positive or negative statements about automation.

Subjective Ratings of Mental Workload

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

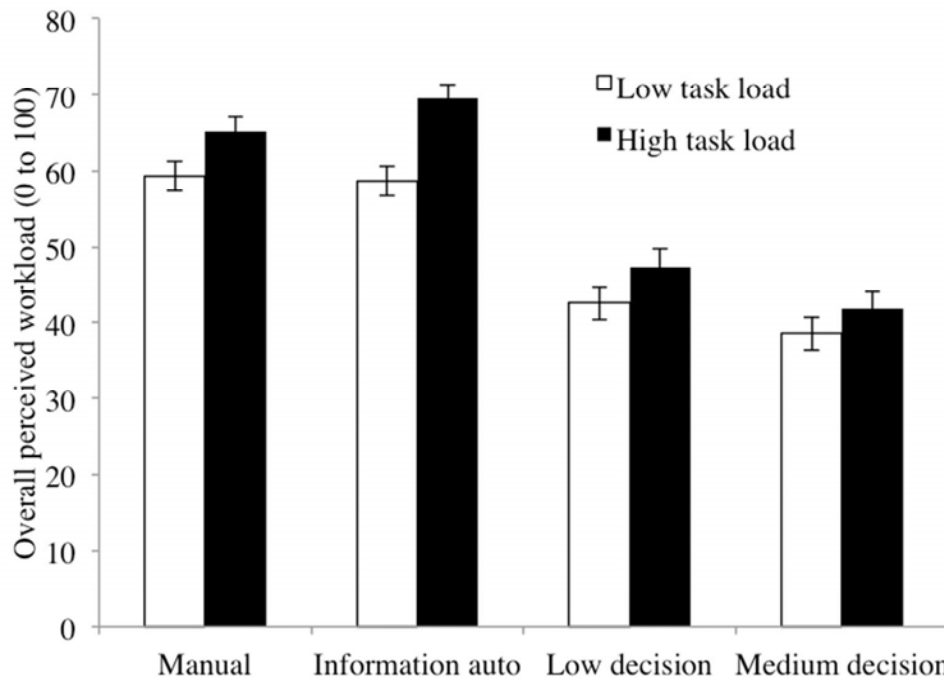


Figure 7. Mental workload as a function of automation support and task load.

DISCUSSION

The extent to which automation enhances decision-making depends on individual differences in cognitive ability. Using a simulated automated targeting task, we showed that the extent to which an operator experienced both the costs of imperfect automation and the benefits of reliable automation depended on individual differences in working memory. This finding may help optimize human-automation interaction.

Our study replicated prior research that operators would perform better with reliable automation compared to manual control (hypothesis 1a). In addition, task load did not differentiate performance when the automation was reliable (hypothesis 1b). Finally, our study showed that with imperfect automation, 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 automation (information and low-decision) and higher automation (medium-decision). It appears that lower automation can mitigate some of the performance penalty of increased task load when automation is imperfect while performance significantly declines with imperfect and higher automation.

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

When automation was imperfect, those with low and high working memory showed declines in accuracy as the type and level of automation increased. However, those with lower working memory were more severely impacted by the unreliability than those with higher working memory. Taken together, these results confirmed hypothesis 2 regarding the effects of type and level of automation and working memory. These results also added detail to the conventional wisdom that increasing automation benefits performance but can lead to catastrophic performance when automation is imperfect (i.e., the lumberjack effect; Onnasch et al., 2014). When automation was reliable, those with higher working memory benefitted more than those with lower working memory, and when automation was imperfect, those with lower working memory suffered more than those with higher working memory. These results

confirmed the link between automation performance and individual differences in working memory as suggested by previous researchers (Parasuraman et al., 2010, Parasuraman, 2012), but also extend the literature by further specifying the automation conditions (type and level of automation support and reliability) under which working memory affects performance.

The lack of any effect of attention on performance was puzzling (hypothesis 4). There may be several differences that explain our disparate results. First, Chen and Terrence (2009) used a *subjective* “perceived attentional control” measure to assess attentional ability whereas we used a spatial cueing task. Second, Chen and Terrence manipulated reliability by adjusting false alarms and miss rates while our task paradigm did not allow for false alarms or miss rates (the automation was always on in the automation-present conditions). Third, the choice of attention measure (a spatial cueing task) and resultant dependent variable (attentional cost from median reaction time) may have not been a sensitive indicator of individual differences in attention in our sample of college students as it was in middle-aged adults (Greenwood et al., 2005). Further, although multiple targets needed to be kept in memory during the sensor-to-shooter task, there were no distractor targets on the screen, meaning the task did not require high levels of attentional control.

More broadly, these results should be put into context with some other possible limitations that may affect the generalizability of the results. First, though college students are typical participants, our sample was from a military academy that possibly made them less representative. Although, the automated task was a simulated command and control task and the participants had completed Army basic training.

Practical Implications

Knowing how operators will perform with reliable, but imperfect types and levels of automation at different task loads is enhanced if we understand the impacts of individual differences in working memory and attention on human automation interaction. One way this knowledge may be useful is in automated systems that alter the types and levels of automation support based on the operators working memory ability, so called adaptive automation. Both working memory and attentional capacity may change as a function of the current task load. Our results provide some information that suggests how different levels of working memory and attention affect performance. These results also provide some guidance in the design of new automated systems. These results showed that the level of working memory demand varies as a function of the type and level of automation and automation reliability. Finally, our results suggest that designers should design interfaces that support individuals by matching their working memory abilities.

Key Points

- It was hypothesized that individual differences in working memory and attention would affect human automation interaction with varying types and levels of imperfect automation or high task load in a simulated command and control task.
- Participants performed a simulated command and control task with manual, information automation, low-decision automation, or high decision automation differing in two levels of task load: low or high. Participants also completed a spatial working memory task and a visuospatial attention task.
- Increased, reliable automation support reduced the differences between those with low and high working memory abilities. Higher working memory ability buffered the costs of imperfect decision automation. Lower working memory was associated with more trust in automation.
- Designers may mitigate some of the performance decrements experienced with imperfect automation by designing interfaces that support individual differences in working memory and attention.

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. & Terrence, P. I. (2008). Effects of tactile cueing on concurrent performance of military and robotics tasks in a simulated multitasking environment. *Ergonomics*, *51*, 1137-1152.
- Crocoll, W. M., & Coury, B. G. (1990). Status or recommendation: Selecting the type of information for decision aiding. *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 1524–1528). Santa Monica, CA: Human Factors and Ergonomics Society.
- de Visser, E., Shaw, T., Mohamed-Ameen, A., & Parasuraman, R. (2010). Modeling human-automation team performance in networked systems: Individual differences in working memory count. *Proceedings of the Human Factors and Ergonomics Society 54th Annual Meeting*, 1087-1091.
- Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, *42*, 462–492.
- Endsley, M. R., & Kiris, E. O. (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., Bolia, R. S., & Parasuraman, R. (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* (pp. 438–442). Santa Monica, CA: Human Factors and Ergonomics Society.
- Gopher, D. (1982) A selective attention test as a predictor of success in flight training. *Human Factors*, *24*, 173-183.
- Greenwood, P. M., Lambert, C., Sunderland, T., & Parasuraman, R. (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
- Greenwood, P. M., Sunderland, T., Friz, J. L., & Parasuraman, R. (2000). Genetics and visual attention: Selective deficits in healthy adult carriers of the E4 allele of the apolipoprotein E gene. *Proceedings of the National Academy of Sciences, USA*, *97*, 11661–11666.
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam: Elsevier Science.
- Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behaviour Research Methods*, *39*(1), 101-117.
- Jian, J., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53-71.
- Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human- Computer Studies*, *40*, 153–184.
- Lohse, G. (1997). The role of working memory on graphical information processing. *Behaviour & Information Technology*, *16*(6), 297-308.

- Lorenz, B., Di Nocera, F., Röttger, S., & Parasuraman, R. (2002). Automated fault management in a simulated space flight micro- world. *Aviation, Space, and Environmental Medicine*, *73*, 886–897.
- Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. *Human Factors*, *56*(3), 476–488.
- Parasuraman, R., & Manzey, D. (2010). Complacency and bias in human use of automation: A review and attentional synthesis. *Human Factors*, *52*, 381–410.
- Parasuraman, R., de Visser, E., Lin, M.-K., & Greenwood, P. M. (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., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced “complacency.” *International Journal of Aviation Psychology*, *3*, 1–23.
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (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 Bryk, A.S. (2002). *Hierarchical linear models (2nd Ed.)*. Thousand Oaks, CA: Sage Publications.
- Rovira, E., Cross, A., Leitch, E., & Bonaceto, C. (2014). Displaying Contextual Information Reduces the Costs of Imperfect Decision Automation in Rapid Retasking of ISR Assets. *Human Factors*.
- Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors*, *49*, 76–87.
- Sarter, N., & Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human Factors*, *43*, 573–583.
- Singh, I. L., Molloy, R., & Parasuraman, R. (1993). Automation- induced “complacency”: Development of the complacency potential rating scale. *International Journal of Aviation Psychology*, *3*, 111–121.
- Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators* (Technical report). Cambridge, MA: MIT, Man Machine Systems Laboratory.
- Tabachnick, B. G., & Fidell, L. S. (2007). Multi-level linear modeling. In *Using Multivariate Statistics* (5 ed., pp. 781–857).
- Unsworth, N., Schrock, J.C., & Engle, R.W. (2004) Working memory capacity and the antisaccade task: Individual differences in voluntary saccade control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1302-1321.
- Wickens, C. D., & Xu, X. (2002). *Automation trust, reliability and attention* (Tech. Rep. AHFD-02-14/MAAD-02-2). Savoy: University of Illinois, Aviation Research Lab.

BIOGRAPHIES

Ericka Rovira is an Associate Professor in the Department of Behavioral Sciences & Leadership at the U.S. Military Academy, West Point. She received her Ph.D. 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 Ph.D. in psychology in 2005 from the Georgia Institute of Technology.

Anne McLaughlin is an Associate Professor in the Department of Psychology at North Carolina State University. She received her Ph.D. in psychology in 2007 from the Georgia Institute of Technology.

APPENDIX

Equation for Model 1

$$\text{Level 1: Accuracy}_{it} = \beta_{0it} + r_{it}$$

$$\text{Level 2: } \beta_{0i} = \gamma_{00} + u_{0i}$$

Equation for Model 2

$$\text{Level 1: Accuracy}_{it} = \beta_{0it} + \beta_{1it}(\text{Task load}) + \beta_{2it}(\text{AutoSupport}) + \beta_{3it}(\text{Reliab}) + \beta_{4it}(\text{Task load*AutoSupport}) + \beta_{5it}(\text{AutoSupport*Reliab}) + \beta_{6it}(\text{Reliab*Task load}) + \beta_{7it}(\text{AutoSupport*Reliab*Task load}) + r_{it}$$

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

Equation for Model 3

$$\text{Level 1: Accuracy}_{it} = \beta_{0it} + \beta_{1it}(\text{Task load}) + \beta_{2it}(\text{AutoSupport}) + \beta_{3it}(\text{Reliab}) + \beta_{4it}(\text{Task load*AutoSupport}) + \beta_{5it}(\text{AutoSupport*Reliab}) + \beta_{6it}(\text{Reliab*Task load}) + \beta_{7it}(\text{AutoSupport*Reliab*Task load}) + r_{it}$$

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

Investigating Older Adults' Trust, Attributions, and, Capability Perceptions of Robots

Jessica Branyon & Richard Pak
Clemson University, Clemson, SC



Introduction

- Social anthropomorphize technology, applying human-like characteristics such as personality and social characteristics (Nass & Moon, 2000).
- Applying social attributes to technology makes it susceptible to stereotyping.
- Stereotypes are pervasive beliefs about characteristics and behaviors of a particular "group". Stereotypes are more likely to be activated in domains that are inconsistent with prescriptive social gender or age roles (e.g., Lee, 2008).
- Although gender stereotypes have been well established in the human-computer interaction literature, evidence suggests that technology can also be vulnerable to aging stereotypes (Pak, McLaughlin, & Bass, 2014).
- Physical appearance, facial features, and perceived age are known to play a role in the activation of age stereotypes (Hummert, Gansler, & Shiner, 1997). Age is one of the first and most salient attributes we notice about other people (Fiske, Kitayama, Markus, & Nisbett, 1998), which may also be true of other anthropomorphic technology such as robots.
- Stereotypes influence the way individuals make social judgments about others, including the types of causal attributions people make about the performance of others. When trying to determine the causality of an event, individuals rely on dispositional qualities of the actor or the external influences of the situation.
- Older adults tend to make dispositional attributions when the outcome of a situation is negative or when their personal beliefs are violated (Blanchard-Fiddis, Herzog, & Hohoda, 2012).
- Stereotypes influence the perceived capabilities of others as well as trust in those capabilities (Mair, 1987).
- Trust in robots is influenced by performance based factors like reliability (Hancock et al., 2011) as well as appearance (Goetz, Kessler, & Powers, 2002).

Current Study

- The purpose of the current study was to investigate whether a robots' appearance, reliability, and task type would influence trust in the robot, the perceived capabilities of the robot, and the causal attributions of the robots' performance.
- We employed a factorial survey methodology where a series of slideshow vignettes were presented to the participants. Participants attitudes toward the robots' behavior and appearance were rated on Likert scales.



Figure 1. Younger adult facial stimuli



Figure 2. Older adult facial stimuli

APPENDIX 3: Branyon, J., & Pak, R. (2015). Investigating older adults' trust, attributions, and capability perceptions of robots. Presented at the American Psychological Association 123rd Annual Meeting, Toronto, ON, American Psychological Association

Methods & Procedure

Participants
• 27 older adults aged 65 to 79 ($M = 70.89$, $SD = 3.86$) participated in the current study

Design

• The study was a 2 (age of robot: young, older; Figure 1 & 2) X 2 (robot reliability: high, low) X 4 (task: changing light bulb, sorting laundry, moving boxes, sorting recycling; Figures 3-6) within-subjects design.
• The dependent variables were trust, perceived capabilities of the robot, and causal attribution ratings

Procedure

• The survey was programmed using Qualtrics and was completed remotely.
• During a 60-minute session, participants viewed randomly presented slideshow vignettes of one of the four tasks and two robot types (reliability, gender). Participants made trust ratings, perceived capability ratings, and causal attribution ratings.
• After responding to 16 unique scenarios, participants completed the CPRS (Singh, Malloy, & Parasuraman, 1993).

Variations of Robot Appearance and Task



Figure 3. Older adult robot sorting laundry



Figure 4. Younger adult robot sorting recycling



Figure 5. Older adult robot stacking boxes



Figure 6. Younger adult robot changing a light bulb

Results

• A 2 (age of robot) X 2 (reliability) X 4 (task) repeated measures ANOVA was conducted.
• There was a significant 2-way interaction of reliability X task on trust, $F(3,78)=5.10$, $p=.03$. When reliability was low, sorting laundry yielded significantly higher trust ratings more compared to the other tasks. When reliability was high, moving boxes produced significantly lower trust ratings than the other tasks (Figure 7).
• There was a significant main effect of reliability on capability ratings such that robots that perform tasks reliably ($M = 3.47$, $SD = 0.71$) versus ($M = 3.47$, $SD = 1.04$) than those that perform tasks unreliably ($M = 2.72$, $SD = 1.06$), $F(1,27)=12.68$, $p=.001$.
• There was a significant 2-way interaction between robot age and task on capabilities, $F(3,78)=51.06$, $p<.001$. When the robot appeared young, sorting laundry yielded significantly higher capability scores than stacking boxes. When the robot appeared old, there were no differences in capability scores by task (Figure 8).
• There was a significant 2-way interaction between robot age and task on situational attributions. When the robot appeared younger, participants attributed performance on the laundry sorting task to situational factors significantly more than performance on sorting recycling and moving boxes ($F(3,78)=2.81$, $p=.045$ (Figure 9).

Results Cont.

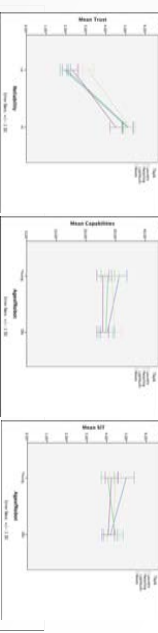


Figure 7. Trust ratings by task and reliability. Figure 8. Capabilities ratings by robot age and task.

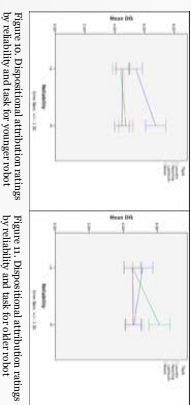


Figure 9. Situational attribution ratings by reliability and task for older robot

• There was a significant 3-way interaction of age of robot X reliability X task on dispositional attributions, $F(3,78)=12.406$, $p<.001$.

• When the robot appeared young, reliable laundry sorting yielded the highest dispositional ratings (Figure 10).
• When the robot appeared older, reliable recycling sorting yielded the highest dispositional ratings (Figure 11).

Discussion & Conclusions

• Although there were no main effects of robot age on the dependent variables, age moderated the effect of task on the robot's perceived capabilities as well as the types of causal attributions individuals made about the robot's performance.
• Individuals tend to have higher trust ratings when completing a fine motor task or light cognitive tasks than when performing a gross motor task (i.e. moving boxes).
• Reliable cognitive task performance yielded the highest dispositional attribution ratings regardless of robot appearance. This finding suggests that people might attribute outcomes differently in the context of human-robot interaction than in human-human interaction.
• These findings emphasize the importance of task type on older adults' perceptions of robots. In this context, users trust robots that perform cognitive and light motor tasks more than ones that perform gross motor tasks. It is also important to select the appropriate age appearance for robots based on the tasks they are to perform. Our results recommend selecting a younger appearance for a robot that will perform cognitive tasks.

References

Blanchard-Fiddis, S., Herzog, R., & Hohoda, M. (2012). The influence of age on trust in robots. *International Journal of Human-Computer Studies*, 68(1), 1-14.
Branyon, J., & Pak, R. (2015). Investigating older adults' trust, attributions, and capability perceptions of robots. Presented at the American Psychological Association 123rd Annual Meeting, Toronto, ON, American Psychological Association.
Clemson University. (2015). *Investigating older adults' trust, attributions, and capability perceptions of robots*. Retrieved from <http://www.clemson.edu/~branyon/>
Fiske, S. T., & Taylor, M. A. (1991). *Social cognition: From perception to action*. New York, NY: Oxford University Press.
Hummert, M. A., Gansler, K., & Shiner, R. L. (1997). Age and gender differences in the perception of facial expressions. *Journal of Nonverbal Behavior*, 21, 287-309.
Kitayama, M., Markus, H. R., & Liberman, N. P. (2004). Cultural differences in the perception of facial expressions. *Journal of Nonverbal Behavior*, 28, 287-309.
Lee, K. S. (2008). *Age and gender differences in the perception of facial expressions*. *Journal of Nonverbal Behavior*, 32, 287-309.
Nass, J. B., & Moon, J. (2000). *Machine mind: The psychology of computer-mediated communication*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
Pak, R., & Branyon, J. (2015). *Investigating older adults' trust, attributions, and capability perceptions of robots*. Presented at the American Psychological Association 123rd Annual Meeting, Toronto, ON, American Psychological Association.
Singh, H., Malloy, J., & Parasuraman, R. (1993). *The computerized personality research scale (CPRS)*. In *Personality and individual differences* (Ed. by S. Ozeri, pp. 1-14). New York, NY: Praeger.

Acknowledgements

This research was supported by a grant from the Air Force Office of Scientific Research (award number A9550-1-13-1-0083) and the Clemson Human Factors Research Institute.



The Effects of Age and Working Memory Demands on Automation-Induced Complacency

William Leidheiser & Richard Pak
Clemson University
Department of Psychology

Complacency refers to a type of automation use expressed as insufficient monitoring and verification of automated functions. Previous studies have attempted to identify the age-related factors that influence complacency during interaction with automation. However, little is known about the role of age-related differences in working memory capacity and its connection to complacent behaviors. The current study aims to examine whether working memory demand of an automated task and age-related differences in cognitive ability influence complacency. Higher degrees of automation (DOA) have been shown to reduce cognitive workload and may be used to manipulate working memory demand of a task. Thus, we hypothesize that a lower DOA (i.e. information acquisition stage with lower level) will demand more working memory than a higher DOA (i.e. decision selection stage with higher level) and that a lower DOA will result in a greater difference in complacency between age groups than a higher DOA.

INTRODUCTION

The World Health Organization (WHO, 2011) estimates that by 2050, there will be approximately 1.5 billion elderly (age 65 and over) in the world. A host of automated services and devices are or will be designed to help older adults maintain independence (e.g., medication reminder apps). Despite this availability of automation and its seemingly utility to maintain independent living (Haigh & Yanco, 2002), research has shown that older adults may be more complacent with automated systems compared to younger age groups (so called automation-induced complacency).

Automation-induced complacency is the “self-satisfaction that may result in non-vigilance based on an unjustified assumption of satisfactory system state” (Billings, Lauber, Funkhouser, Lyman, & Huff, 1976). It is the state in which a user fails to notice imperfect automation. The fault is not detected because the user is poorly monitoring the system, which can result in acceptable performance with reliable automation or diminished performance with unreliable automation (Parasuraman & Manzey, 2010). For instance, an older adult with diabetes may monitor their blood glucose levels with an automated tool. If the older adult perceives the device as reliable and trusts that the blood glucose readings are accurate, they may rely on the reading even when starts to falter. As older adults begin to adopt automated technologies, it is important to understand the age-related factors that contribute to increased complacency and the performance costs associated with those behaviors.

Older Adults, Working Memory, and Complacency

Older adults have been found to be more complacent with automation relative to younger adults (Ho, Wheatley, & Scialfa, 2005b). Various studies have suggested several possible explanations for older adults increased complacency. Some person-related variables range from issues such as higher levels of trust (Johnson, Sanchez, Fisk, & Rogers, 2004; Pak, Fink, Price, Bass, & Sturre, 2012), or age-related differences in abilities (e.g., working memory; Ho et al.,

2005b) while some system-related variables are reliability of the automation (Sanchez et al., 2004) and workload (McBride, Rogers, & Fisk, 2011).

Research investigating age differences in cognitive ability as a possible explanation for changes complacency has found that in a high working memory demanding automated task, older adults relied more on the automation, committed more errors, had greater trust in the system, and were less confident in their own abilities compared to younger adults (Ho et al., 2005b). Based on their findings, they concluded that age-related differences in working memory might be a potential reason for age differences in complacency due to the memory dependent automated task. For instance, the younger adults were able to hold more information about the task in their working memory (Ho et al., 2005b). Since they could actively store and recall this information when needed, younger adults could more easily identify an automation failure compared to older adults.

Researchers theorized there are two main factors that contribute to older adults’ complacent behavior with automated technologies (Ho, Kiff, Plocher, & Haigh, 2005a). The first is that while using automation, older adults form an inaccurate mental representation of the correct values used in the decision making process due to reduced working memory capacity. The second is that due to their reduced working memory capacity, older adults are unable to judge the accuracy of automation. In both cases, it is assumed older adults’ relative complacency with automation is due to a mismatch between the working memory demands of the task and working memory capacity of the person (Ho et al., 2005a). If working memory capacity plays such a central role in automation complacency, we should observe the opposite relationship as well: reduced complacency in older adults when the automation has been designed to demand relatively less working memory resources (or working memory resources are less constrained). The design of Ho et al.’s (2005b) study precludes this determination because it is unclear whether the high working memory demands of the task or the degree of automation (DOA) contributed to the difference in complacency.

How Complacency is Influenced by Automation-Related Factors

Reliability. Automation reliability is the overall accuracy of the system and is an important factor of automation-induced complacency because the number of errors it produces can impact dependence on automation.

Across different levels of reliability, age is known to produce increased effects on trust in automation. For instance, several studies found that higher reliability led to higher subjective trust in the system for both age groups, but older adults had significantly higher trust than younger adults (Sanchez et al., 2004; Ho et al., 2005b). Highly reliable automation is problematic because users can become accustomed to its high level of performance and may not expect it to fail.

Research on age differences in automation use has found that older adults tend to overestimate the actual automation reliability (Olson et al., 2009). With known differences in working memory, older adults have difficulty detecting errors and perceiving overall automation performance. A combination of unnecessarily high trust in the system and a lack of working memory may produce a lack of error prone awareness consistent with complacent behavior.

Workload. The workload or demand of a task can be taxing on an individual's cognitive resources, especially when a task is performed over a long period of time. Greater complacency has been shown in a multitask environment instead of a single task or monitoring role for younger adults (Parasuraman, Molloy, & Singh, 1993).

Older adults have a greater tendency to monitor automation and verify the accuracy of the information, even under taxing conditions (Ho et al., 2005b). Exerting more cognitive resources to complete a task may lead the user to rely on automation after task demands become too difficult to manage. There are also age differences in complacency that have occurred under equivalent high workload conditions, where older adults display greater complacency than younger adults (Hardy, Mouloua, Dwivedi, & Parasuraman, 1995; Vincenzi, Muldoon, Mouloua, Parasuraman, & Molloy, 1996; Ho et al., 2005b). If workload only partially contributes to increases in complacency, other age-related factors must be involved as well.

Working memory capacity has been found to significantly predict younger adult performance in an automated task with varying workload (de Visser, Shaw, Mohamed-Ameen, & Parasuraman, 2010). Since working memory plays a role in predicting performance, this cognitive ability may explain some age-related differences in complacent behaviors.

Degree of Automation. Automation comes in a variety of forms, which can execute different functions for the user based on their capabilities and limitations. However, automation is not simply an all or none concept because any individual task can feature varying degrees of automation that take into account the use of various stages and levels (Wickens, Li, Santamaria, Sebok, & Sarter, 2010).

Parasuraman, Sheridan, and Wickens (2000) identified several stages of automation that are based on an existing

model of human information processing: information acquisition (stage 1), information analysis (stage 2), decision and action selection (stage 3), and action implementation (stage 4). Each stage is designed to support a different aspect of the cognitive process.

Levels of automation differ from stages because they affect the role of humans and automated systems in a given task. These levels exist on a spectrum of automation, where each level between manual and fully automated changes the designation of authority for decision-making tasks. A low level of automation grants authority to the human, making the individual an active participant in the task and giving the system a secondary role of the passive monitor. These roles are reversed under a high level of automation.

Along each stage of automation, varying levels can be applied to achieve a lower or higher DOA. More automation or a greater DOA can be achieved with both higher levels within a stage and later stages (Manzey, Reichenbach, & Onnasch, 2012). Also, higher DOAs are associated with greater performance in addition to diminished workload (Wickens et al., 2010). Since workload is reduced under a higher DOA, the automation is taking on more of the cognitive demand for those tasks than the operator. This leaves the operator with more cognitive resources at higher DOAs. Thus, working memory demands should lessen as the user moves from a lower DOA towards a higher DOA.

Higher complacency can take the form of performance detriments under unreliable systems and performance gains for increasingly reliable automation. For instance, a meta-analysis found that higher DOAs lead to greater accuracy for younger adults, but only when the automation performed optimally (Onnasch, Wickens, & Manzey, 2013). However, there was a greater performance cost for imperfect automation as DOA increased. For younger adults, these findings reveal differences in performance across DOAs, which seem to indicate changes in complacent behavior. In this context of comparing performance across lower and higher DOAs, research on the older adult population has not been performed. In terms of research by Ho et al. (2005b), it is still unclear whether the high working memory demands of the task or the high DOA contributed to age-related differences in complacency.

Current Study

The aim of this study is to examine the relationship between automation-induced complacency and working memory. Age-related differences in working memory have been implicated as a possible cause of age-related differences in automation-induced complacency. However, prior automation studies (e.g., Ho et al., 2005b) have not manipulated working memory demands of the task to observe how complacency is affected. Therefore, we will use two DOAs that vary in working memory demand. This study will analyze speed and accuracy of user selections at each DOA. Performance under reliable and unreliable trials can provide information to infer the degree to which users are complacent with automation.

METHOD

Participants

Thirty-six undergraduate students will be recruited for this research and given course credit for participation. Thirty-six older adults from the local area will be recruited and will be compensated for their time.

Task

The tasks for this study will be adapted from prior research that uses an automated system in the context of a low-fidelity UAV simulation (Rovira, McGarry, & Parasuraman, 2007). The primary task for this study will be to quickly and accurately find the closest combination of friendly (green units) and enemy units (red units) in terms of distance apart on the grid (Figure 1). Automation will be presented as a table in the bottom left-hand corner of the screen, which will display the distances and unit combinations needed by participants to complete the primary task. The secondary task will consist of checking for a specific call sign and clicking a corresponding button when it appears on screen. The call sign is comprised of a single word and number combination (e.g. Hunter-6). The program will randomly alternate between 14 different call signs every 5 seconds as the participant completes the primary task.

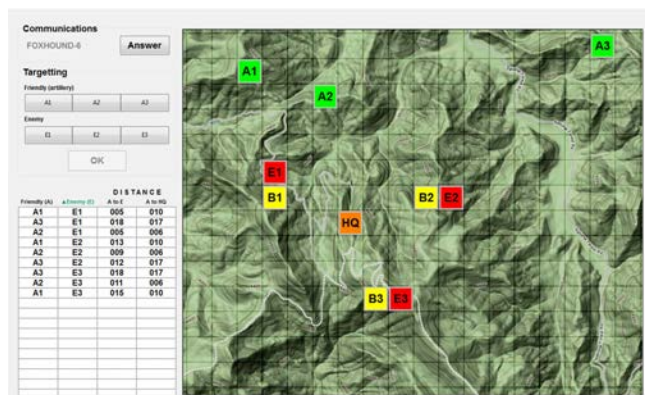


Figure 1. Screenshot of a low degree of automation (DOA) and low workload trial within the targeting system that features the communications panel (top-left), targeting input panel (top-left), automation table (bottom-left), and grid (right).

Participants will complete blocks of trials in a random counterbalanced order, where each block will consist of a different DOA and workload level. The DOA manipulation will change the stage and level of the automation table used in the task. The lower DOA will use the information acquisition stage, which presents all possible friendly and enemy unit combinations from the grid, with a low level of automation that does not sort the information in any meaningful way. The higher DOA will use the decision and action selection stage, which will present the top 3 friendly and enemy unit combinations. In addition, the high level of automation will sort the information based on importance, so that the shortest distance combination is presented at the top. The workload

manipulation will change the number of units presented in the grid. Low workload will present 3 friendly and 3 enemy units, while high workload will show 6 friendly and 6 enemy units. Each combination of DOA and workload will be presented twice for a total of 8 blocks and 240 trials.

The overall automation reliability will be set at 80%, which is above the threshold for imperfect reliability acceptance (Wickens & Dixon, 2007). In each block of 30 trials, 24 trials will be reliable and the remaining 6 trials will be unreliable. An unreliable trial will contain inflated distance values between unit combinations or incorrect optimal suggestions in the automation support table. The first aid failure will not occur until the 10th trial, so that users can build rebuild trust after each block. Also, the automation failures will be distributed randomly throughout the remaining trials.

Measures

Ability measures. The following abilities will be assessed: perceptual speed (digit-symbol substitution; Wechsler, 1997), working memory (automated operation span (Aospan); Unsworth, Heitz, Schrock, & Engle, 2005), and vocabulary (Shipley vocabulary; Shipley, 1986). These measures were chosen because they are reliable indicators of their respective abilities (e.g., Czaja et al., 2006; Unsworth et al., 2005). The cognitive ability measures were selected to confirm age differences in fluid and crystallized intelligence. Specifically, the working memory ability measure serves to control for differences in targeting task performance between age groups.

NASA-Task Load Index (NASA-TLX). Subjective workload will be measured with the NASA-TLX (Prichard, Bizo, & Stratford, 2011). A computer version of the task will present 6 items that constitute overall workload: mental demand, physical demand, temporal demand, performance, effort and frustration. Each item is rated on a Likert scale of 0 to 20, where higher values indicate increased workload. Subjective workload will be calculated as the average of the 6 combined items. The NASA-TLX will be used as a manipulation check for DOAs and age differences in perceived workload.

Trust Questionnaires. Subjective trust will be measured with a general rating of trust in automation (Jian, Bisantz, & Drury, 2000). This measure is a 12-item survey that is rated on a Likert scale of 1 (not at all) to 7 (extremely). The first 5 questions are negatively framed and the last 7 are positively framed. Trust is the sum of normal and reverse coded responses. Higher scores on this measure indicate greater trust in the automated system. The measure will be analyzed for age-related differences in trust towards automation.

In addition, we will use a survey adapted from Lee and Moray (1992) to measure subjective trust specifically towards each DOA and working memory manipulation. This trust measure will pose 3 questions, rated from 0 (not at all) to 100 (extremely), about the automated aid used in each set of trials. For example, the questions will ask participants to answer how much they trusted, relied upon, or benefited from using the automated aid. The overall score will consist of an average of those questions and higher scores will indicate higher trust.

Additionally, this questionnaire will be used to examine trust differences between age groups, level of workload, and DOA.

Complacency Potential Rating Scale (CPRS). The CPRS measures individual potential complacency behavior (Singh, Molloy, & Parasuraman, 1993). This 20-item scale contains 4 filler items and is rated on a Likert scale of 1 (strongly disagree) to 5 (strongly agree). The CPRS scores is a sum of these responses except for the filler responses, where higher values on this measure indicate an increased complacency potential. The CPRS was selected in order to examine age differences in complacency potential.

Design

The current study is a 2 (age group: young or old) x 2 (DOA: low or high) x 2 (automation reliability: unreliable or reliable) x 2 (workload: low or high) mixed-subjects design. Age group will be a between-subjects independent variable. These groups will differ in working memory capacity because older adults have been shown to have less of this ability than younger adults. DOA, automation reliability, and workload will be within-subjects independent variables. The DOAs serve as our working memory demand manipulation.

The dependent variables will be targeting task accuracy, targeting task completion time, complacency potential, subjective trust, subjective workload, and working memory capacity. *Targeting task accuracy* will be measured by the mean rate of optimal responses for each automation block. An optimal response is the identification of the closest pair of friendly and enemy units on the targeting task grid. *Targeting task time* will be measured by the average duration (in milliseconds) it takes participants to complete each trial. *Complacency potential* will be comprised of scores on the CPRS. *Subjective trust* will be measured by the sum of subjective ratings on the trust questionnaire for each combination of DOA and workload level. *Subjective workload* will consist of an average of the 6 items on the NASA-TLX and will be measured for each combination of DOA and workload level. *Working memory capacity* will be measured as the sum of perfectly recalled sets of letters on the Aospa task.

Procedure

Participants will be seated at individual PC-computers and provided with informed consent. They will be instructed to complete the demographics form and the cognitive ability measures. The experimenter will then tell participants to open and observe the targeting task instructions screen. Participants will be told the following: "In this experiment, you will have two tasks. The first task will be to monitor the communications panel for the call sign Hunter-6. When you see Hunter-6, you should click the answer button. The second task will be to target enemy units with the closest artillery unit as quickly as you can. You will do this by first selecting an artillery unit and then select an enemy target from the list of buttons. The computer will sometimes help you with this task by showing you the distances between friendly and enemy units. Sometimes, two sets of targets will have the same distance. In this case, you will pick the one with the shortest

distance to the headquarters. Sometimes the computer aid will give you lots of information, other times it will give you much less information. The computer can be very reliable but it is not perfect all the time." After these instructions, the experimenter will answer questions before continuing.

As the participants complete the tasks, the units in the grid and the values within automation table will change for each subsequent trial. Between each block of trials, participants will fill out the NASA-TLX and a brief subjective trust measure. During the experiment, a screen will appear to indicate when participants linger too long on a particular trial. If participants do not input friendly and enemy unit combinations within the set time limit, the program will automatically continue to the next trial. Younger adults will have 10 seconds to complete each trial, while older adults will have 15 seconds. Older adults will have more time for the task because of normative age-related differences in psychomotor speed (Salthouse, 1985).

Participants will proceed through each block of trials and the computer will notify them when they are finished. When they complete the automation program, participants will be presented with a general subjective measure of trust in automation and the CPRS. At the conclusion of the experiment, participants will be debriefed and provided compensation for their time.

EXPECTED RESULTS

Repeated measures ANOVAs will be performed to test these expected results. We anticipate main effects of DOA as well as age group on targeting task accuracy and task time, where younger adults should outperform older adults. Overall, we expect participants to perform better under a higher DOA (i.e. decision selection stage with higher level) than a lower DOA (i.e. information acquisition stage with lower level). Also, we will measure differences in subjective workload and trust towards specific DOAs and levels of workload. For those variables, we expect to find main effects of workload and DOA.

Since we expect an inverse relationship between DOA and cognitive demand, we hypothesize that older adults will have a greater tendency to become complacent under a lower DOA. We can infer the extent to which participants are complacent by analyzing their pattern of performance at different reliability levels. A greater difference between performance with unreliable and reliable automation indicates higher complacency because the user is relying heavily on the system without monitoring for failures. Therefore, we will perform a repeated measures ANOVA to examine targeting task accuracy for unreliable and reliable trials across DOAs and age groups. We hypothesize a lower DOA will result in a greater difference in complacency between age groups than a higher DOA. We anticipate this result because a higher DOA should support working memory ability by taking on more cognitive demanding tasks that would otherwise burden the user. Consistent with previous findings, younger adults should be more inclined to become complacent with a higher DOA. When taking into account age group differences in working

memory ability, we expect that age-related performance effects will not be present.

Finally, we anticipate that older adults will have higher general trust and complacency potential than younger adults. We will conduct two independent samples t-tests to compare differences in complacency potential and general trust in automation between age groups.

DISCUSSION

It is important to understand the factors that contribute to complacent behaviors within the human-automation interaction. For the design of automated systems, it is necessary to consider factors such as reliability and workload. Since high system reliability is common in most automated technologies today and thus makes users more susceptible to complacent behaviors, it is essential to alert the user to potential automation-related failures that can occur. In terms of task demands, keeping the task manageable for the user is critical for detecting and correcting inaccuracies.

Designers should select the appropriate DOA for the known population of users. Specifically, the design of automated tasks should consider the age of the user. Automation can be presented in many different ways and can perform a wide range of tasks for the user. Depending on the type of task, some forms may demand more working memory than others. Limiting working memory demand through automation can be beneficial to both younger and older adults. This may help to reduce the occurrence of complacent behaviors during interaction with automation.

REFERENCES

- Billings, C. E., Lauber, J. K., Funkhouser, H., Lyman, G., & Huff, E. M. (1976). *Aviation Safety Reporting System* (Technical Report TM-X-3445). Moffett Field, CA: National Aeronautics and Space Administration Ames Research Center.
- Czaja, S. J., Charness, N., Fisk, A. D., Hertzog, C., Nair, S. N., Rogers, W. A., & Sharit, J. (2006). Factors predicting the use of technology: Findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). *Psychology and Aging, 21*(2), 333-352.
- de Visser, E., Shaw, T., Mohamed-Ameen, A., & Parasuraman, R. (2010). Modeling human-automation team performance in networked systems: Individual differences in working memory count. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 54*(14), 1087-1091.
- Haigh, K. Z., & Yanco, H. (2002). Automation as caregiver: A survey of issues and technologies. In *AAAI-02 Workshop on Automation as Caregiver: The Role of Intelligent Technology in Elder Care*, 39-53.
- Hardy, D. J., Mouloua, M., Molloy, R., Dwivedi, C. B., & Parasuraman, R. (1995). Monitoring of automation failures by young and old adults. In *International Symposium on Aviation Psychology* (pp. 1382-1386). Columbus, OH: Association of Aviation Psychology.
- Ho, G., Kiff, L. M., Plocher, T., & Haigh, K. Z. (2005a). A model of trust & reliance of automation technology for older users. In *AAAI-2005 Fall Symposium: "Caring Machines: AI in Eldercare"*, 45-50.
- Ho, G., Wheatley, D., & Scialfa, C. T. (2005b). Age differences in trust and reliance of a medication management system. *Interacting with Computers, 17*(6), 690-710.
- Jian, J., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*(1), 53-71.
- Johnson, J. D., Sanchez, J., Fisk, A. D., & Rogers, W. A. (2004). Type of automation failure: The effects on trust and reliance in automation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 48*(18), 2163-2167.
- Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*(10), 1243-1270.
- Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making, 6*, 57-87.
- McBride, S. E., Rogers, W. A., & Fisk, A. D. (2011). Understanding the effect of workload on automation use for younger and older adults. *Human Factors, 53*(6), 672-686.
- Olson, K. E., Fisk, A. D., & Rogers, W. A. (2009). Collaborative automated systems: Older adults' mental model acquisition and trust in automation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 53*(22), 1704-1708.
- Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2013). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human Factors, 56*(3), 476-488.
- Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics, 55*(9), 1059-1072.
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381-410.
- Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced 'complacency'. *The International Journal of Aviation Psychology, 3*(1), 1-23.
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 30*(3), 286-297.
- Prichard, J. S., Bizo, L. A., & Stratford, R. J. (2011). Evaluating the effects of team-skills training on subjective workload. *Learning and Instruction, 21*(3), 429-440.
- Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors, 49*(1), 76-87.
- Salthouse, T. (1985). Speed of behavior and its implications for cognition. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 400-426). New York: Van Nostrand Reinhold.
- Sanchez, J., Fisk, A. D., & Rogers, W. A. (2004). Reliability and age-related effects on trust and reliance of a decision support aid. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 48*(3), 586-589.
- Shipley, W. C. (1986). *Shipley Institute of Living Scale*. Los Angeles: Western Psychological Services.
- Singh, I. L., Molloy, R., & Parasuraman, R. (1993). Automation-induced "complacency": Development of a complacency-potential scale. *International Journal of Aviation Psychology, 3*(2), 111-122.
- Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior research methods, 37*(3), 498-505.
- Vincenzi, D. A., Muldoon, R., Mouloua, M., Parasuraman, R., & Molloy, R. (1996). Effects of aging and workload on monitoring of automated failures. In *Proceedings of the 40th Human Factors and Ergonomics Society Annual Meeting* (pp. 1556-1561). Santa Monica, CA: Human Factors and Ergonomics Society.
- Wechsler, D. (1997). *Wechsler Memory Scale III*. (3rd Ed.). San Antonio, TX: The Psychological Corporation.
- Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science, 8*(3), 201-212.
- Wickens, C. D., Li, H., Santamaria, A., Sebok, A., & Sarter, N. B. (2010). Stages and levels of automation: An integrated meta-analysis. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 54*(4), 389-393).
- World Health Organization, National Institute on Aging, National Institutes of Health, and U.S. Department of Health and Human Services. (2011). *Global Health and Aging* (NIH Publication No. 11-7737). Retrieved from http://www.who.int/ageing/publications/global_health.pdf

Effects of Age and Gender Stereotypes on Trust in an Anthropomorphic Decision Aid

Brock Bass¹, Meghan Goodwin¹, Kayla Brennan¹, Richard Pak¹, & Anne McLaughlin²
¹Clemson University, ²North Carolina State University

Stereotypes are beliefs about the capabilities of another group. Previous research indicates stereotypes can affect how users interact with anthropomorphic computer aids. User perception can be affected by gender and age stereotypes elicited by the appearance of the computer system. Other research has shown that perceptions of automation (e.g., implicit ones such as propensity to trust automation, or perceptions of etiquette) interact with reliability to influence automation trust behavior. The current study built upon these ideas to examine whether implicit beliefs (i.e., stereotypes) about the perceived age and gender of automation interacted with reliability to affect perceptions of trust in automation. We employed a factorial survey where we presented scenarios of automation to younger adults. The anthropomorphized automation had a perceived age and gender, and was stated to be variably reliable.

INTRODUCTION

Stereotypes in Human-Computer Interaction (HCI)

Stereotypes are preconceptions about the traits, behavior, or abilities of another group. They help set our expectations of individuals that we meet. For example, a commonly held stereotype of athletes is that they are unintelligent, but have social prowess. As the example shows, stereotypes can have both negative and positive connotations that may be inconsistent with real group attributes (i.e., not all athletes may be unintelligent or have social prowess). Stereotypes have adaptive value because they function as schemas by filtering and organizing incoming information thereby easing processing and interpretation (Hilton & Von Hippel, 1996). However, when the stereotype is highly simplified or inaccurate, it can lead to errors in perceptions and behavior.

Stereotypes do not just affect person-perception, but also computer-perception. Computers, intentionally or not, can exhibit anthropomorphic characteristics. Anthropomorphism can be defined as the attribution of human characteristics (e.g., mental states, motives, and emotions) to non-human agents, such as computers (Epley, Waytz, & Cacioppo, 2007). Previous research has investigated the phenomenon of human users imputing human social characteristics (e.g., stereotypes) to computer systems (Nass, Steurer, & Tauber, 1994). This phenomenon is addressed by the Computers are Social Actors experimental paradigm (Nass et al., 1994). The CASA experimental paradigm described by Nass et al. is as follows: pick a social science finding, replace the human with a computer, design the computer with characteristics associated with humans, and determine if the rule still applies. A wide range of studies using the CASA paradigm have shown that users tend to treat

computers with the same social rules and heuristics as they would other people (Fogg & Nass, 1997; Katagiri, Nass, & Takeuchi, 2001; Zambaka, Goolkasian, & Hodges, 2006).

Gender and Age Stereotypes in HCI

Previous research has shown that gender stereotypes are present in HCI. Nass et al. (1994) confirmed that humans apply gender stereotypes relating to “knowledgeability” in HCI situations. That is, when the perceived gender of the computer voice matched the stereotypical topic (love and relationships for the female voice; computers and technology for the male voice), subjects rated the computer as a better teacher. This finding confirms the pre-existing gender stereotype between gender and appropriateness of topic. An implication from this finding is that computers can be perceived as “gendered” just as we assign gender stereotypes to humans. At a broader level, the study supports the influential power that gender stereotypes carry (Nass et al., 1994). Further supporting users’ tendency to gender-type computers, Lee (2003) showed that people conform to the advice of an aid that is sex-stereotypically matched (feminine aid for fashion, masculine aid for sports). These gender stereotype studies demonstrate that even when users may be unaware of applying stereotypes; these stereotypes nonetheless become activated and affect perceptions in a task with a gendered computer.

Interestingly, perceived ethnicity has been shown to contribute to user perceptions of benefits from anthropomorphic agents more than perceived gender (Benbasat, Dimoka, Pavlou, & Qui, 2010). Users perceived agents as more enjoyable and useful when there was a perceived ethnicity match (Asian users with Asian aids, Caucasian users with Caucasian aids), but there was no

effect for gender matching. Benbasat et al. concluded that the significant perceived ethnicity effect was due to the similarity-attraction hypothesis (Byrne, 1971), which states that people are more attracted to those who are similar to themselves.

The vast majority of CASA studies examined a single age group (younger adults) and thus have not examined or manipulated perceived age of the computer agent. Age is one of the first and most salient attributes we notice of other people (Fiske, Kitayama, Markus, & Nisbett, 1998), which may also be true of anthropomorphic agents in HCI. Therefore, examining age stereotypes among younger and older adults is relevant in HCI. There is evidence that age stereotypes (i.e., stereotypes about older adults) are much stronger (Kite, Deaux, & Miele, 1991) and more complex than gender stereotypes (Kite, Stockdale, Whitley, & Johnson, 2005). Kite et al. (1991) assessed age and gender stereotypes and showed that when negative stereotypes were generated, they were more likely due to the age of the target person than the gender (approximately 3 times greater in magnitude). This suggests that, according to the similarity-attraction hypothesis (Byrne, 1971), older adults should exhibit positive anthropomorphic effects with automation that matches their age group. A previous study (Pak, Fink, Price, Bass, & Sturre, 2012) found that a young female agent affected trust in automation in younger adults, but not in older adults. One explanation for the age difference was that the dissimilarity between a younger female decision aid and an older participant may have muted any potential anthropomorphic effect on trust due to the similarity-attraction hypothesis. An alternative explanation is that older adults hold negative stereotypes of the capabilities of younger, female doctors.

Stereotype Activation Depends on Individuating Information

Individuating information such as context (e.g., interacting with a doctor) is also known to determine which aspect of a stereotype gets activated (Casper, Rothermund, Wentura, 2011). Kunda and Sherman-Williams (1993) found whatever the stereotype; its ultimate construal and effect on judgment will depend on the individuating information. Knowing the occupation of an individual is a type of individuating information that seems to alter some negative age stereotypes. For instance, people hold stereotypes that in general older workers have lower ability, are less motivated, and are less productive than younger workers (Posthuma & Campion, 2009). Older adult workers are seen as less adaptable to changing work situations and uncertainty than younger workers (DeArmond, Tye, Chen, & Krauss, 2006). However, the occupation of physician is

moderately seen as a stereotypically older male occupation (Singer, 1986), even though it is an occupation that may require adaptability and facing uncertainty. The individuating information (i.e., occupation of a doctor) allows certain aspects of the stereotype to be activated but not others.

In addition to aspects like occupation, another type of individuating information is past behavior, or more relevant to the current study, a past history of ambiguous system performance (i.e., history of moderate reliability). The assumption is ambiguous system performance will lead to stereotype activation, while unambiguous system performance (i.e., history of unambiguously low or high reliability) will not. Merritt and Ilgen (2008) theorized that implicit attitudes about automation affect whether an individual trusts automation. The user's explicit (e.g., reliability) and implicit (e.g., stereotypes) beliefs (schemas) about automation will shape their perceptions of automation behavior (Dzindolet, Pierce, Beck, & Dawe, 2002). Merritt and Ilgen found that when automation reliability was ambiguous, implicit, pre-existing beliefs about automation were more influential in determining trust than explicit beliefs. Presumably, in the face of automation ambiguity, individuals made attributions that were consistent with their implicit, schematic pre-existing beliefs about automation. This parallels findings from the social cognition literature which shows causal reasoning is common when an individual is faced with conflicting or ambiguous information (Kunda & Thagard, 1996). That is, when automation is unambiguously good or bad, stereotypes should not affect perceptions. But when automation is ambiguous, stereotypes will exert an effect on perceptions of the automation (i.e., trust). Previous human factors research has shown that automation reliability has a "crossover point" or threshold that affects human operator performance. This threshold occurs when automation is below 70% reliable, and results in operator performance similar to situations with no automation. That is, when automation is much less than 70% reliable, the operator begins to behave as if there was no automation present (Wickens & Dixon, 2007).

Current Study

The purpose of the current study was to investigate how trust in automation is affected by stereotypes (age and gender) and how these stereotypes interact with machine factors (reliability) to affect user trust. We manipulated anthropomorphic aids on the following variables: perceived gender (male, female), perceived age (young, old), and automation reliability (45%, 70%, 95%) to investigate their effect on user trust in HCI. The 70% reliability level reflects the "crossover point" described

in previous literature, and the 45 % and 95 % reliability levels were chosen as substantial deviations from 70 % reliability (i.e., a 25 % increase or decrease in reliability). In the current study, we used a factorial survey methodology, which is a type of survey that contains elements of a factorial experiment. In a factorial survey, a respondent evaluates a scenario and then is asked to make a judgment of interest. The scenario can be a short story or a snapshot of a situation, which in the current study is in the context of diabetes management. The important aspect is that specific factors of the scenario are being manipulated (in a factorial manner). The respondent is repeatedly exposed to all combinations of factors in a series of scenarios. Because our dependent variable (trust) is a social judgment about a situation, a factorial survey is an ideal way to measure how judgment is influenced by perceptions of the automation (i.e., age, gender, reliability) as well as individual differences. Factorial surveys have been widely used in various domains to examine how beliefs, judgments, and decision-making are influenced by situational factors.

METHOD

Participants

The participants for the study were Clemson University undergraduate students ($N = 50$). The age range for these participants was 17 to 23 ($M = 18.58$, $SD = .93$). These introductory psychology students received class credit for participating in the study.

Apparatus

Participants viewed the experiment on desktop computers situated in cubicles. The computers presented stimuli on 19-inch LCD monitors and participants made all responses using the keyboard and mouse. They were seated in office chairs about 18-24 inches from the screen in an office environment.

Design

The study was a 2 (gender of respondent: male, female) \times 2 (perceived aid age: young, old) \times 2 (perceived aid gender: male, female) \times 3 (automation reliability: low, medium, high) mixed factorial survey. The first variable (gender of respondent) was a quasi-independent grouping variable, while the last 3 were within-groups manipulations of the automation. The dependent variables were trust, likelihood of following advice, complacency potential rating scale (CPRS; Singh, Molloy, & Parasuraman, 1993), and the participant's diabetes knowledge.

Procedure

Participants first completed a diabetes knowledge questionnaire administered on a computer. The 23 questions assessed basic knowledge about diabetes and diabetes management. Next, participants started the factorial survey portion. Participants were told the following: "You are playing the part of a newly diagnosed diabetic. Your doctor has given you a variety of different smartphone apps that may help you with your diabetes care. Your task involves giving us your opinion of the different smartphone apps. Just like many technological aids, the different apps will only sometimes seem reliable. Your performance is not being tested so you do not have to try to solve every problem. Instead, you are making judgments of the smartphone apps as quickly as possible." After acknowledging the instructions and asking any remaining questions they began the survey. In the survey, participants viewed each vignette and were asked the following questions: 1) how much they trusted the smartphone app on a Likert scale from 1 (not at all) to 7 (very much), and 2) whether they would follow the advice of the app (Likert scale, 1-7). After each question, participants were also asked to briefly explain their ratings by typing a brief explanation in a subsequent field. To reinforce the notion that the smartphone app was real automation (and not just a pre-computed image), the smartphone app did not reveal its answer for 1.5 seconds (in the interim the message "Analyzing the scenario. Just a moment..." appeared on the smartphone screen). After responding to 24 vignettes, participants completed the complacency potential rating scale. Finally, after completing the CPRS, participants answered the question, "What do you think the study was about?" in order to assess whether participants realized the purpose of the study. This question was to determine if participants were aware of our experimental manipulation and thus prone to demand characteristics.

RESULTS

The results from this study are presented in two sections aligned with data type: quantitative (Likert ratings) and qualitative (explanations for Likert ratings).

Quantitative Data

To examine differences in trust as a function of aid characteristics, a 2 (age group of aid) \times 2 (gender of aid) \times 3 (reliability of aid) ANOVA was conducted. There was a significant 3-way interaction of age group of aid \times gender of aid \times reliability of aid, $F(1, 1440) = 3.84$, $p < .05$. This finding came from analyzing the trust ratings given by the participants concerning each aid via Likert

scales. The trust ratings were further analyzed as a function of the reliability of the aid (see Figure 1). In the low reliability (45 %) condition it was found that the older female aid was the most trusted. The younger male aid was the second most trusted in the low reliability condition. In the moderate reliability (70%) condition it was found that the older male aid was the most trusted, while the younger female aid was the second most trusted. This replicated previous research findings (Pak et al., 2012) that showed younger adults trusted a younger female aid significantly more than a non-anthropomorphic aid. In the high reliability (95 %) condition there was no significant difference found in trust ratings. Regardless of the aid characteristics (i.e., gender, age) the participants indicated similar trust ratings for each aid in the high reliability condition.

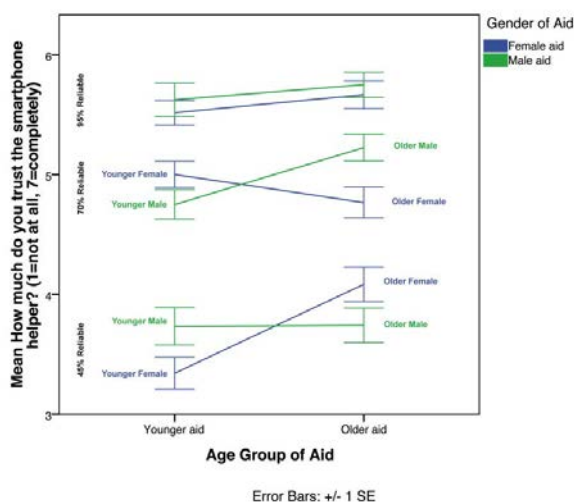


Figure 1. Trust ratings for anthropomorphic aids.

Qualitative Data

The participants’ explanations for their trust ratings were used to help interpret the numerical trust ratings presented above. Two coders achieved reliability above 70% on the simple coding scheme. Trust explanations were coded by their dominant theme using a coding scheme generated from a subset of a random number of statements. We have currently only examined 1/4 of the qualitative data (approximately 340 of the 1200 statements), but the trends (Figure 2) seem to show that subjects did not overtly attribute trust ratings to perceived age or gender (categories A and B). However, there did seem to be a trend to attribute trust ratings more to a general tendency to trust/mistrust machines when the aid was younger (category C) compared to older. In addition, subjects stated that they trusted/mistrusted the aid because of their double-checking efforts most for the older male aid (category E). This may be reflective of

the stereotype of older male doctors (i.e., older men are trustworthy doctors).

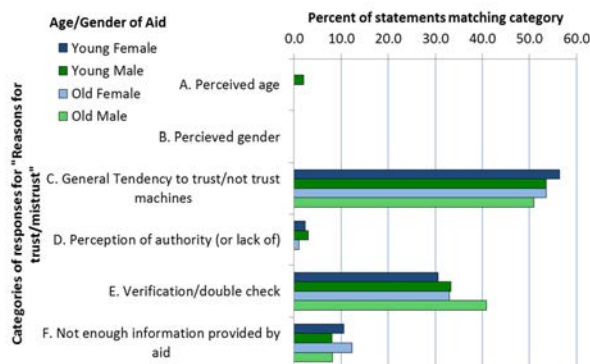


Figure 2. Categorization of trust explanations.

DISCUSSION

The findings of the current study extend the literature about how people treat and behave with anthropomorphic computer aids. We found that age stereotypes can be elicited and can affect trust, but in a complex way. The significant 3-way interaction (age group of aid × gender of aid × reliability of aid) can be thought of as a 2-way interaction (age group of aid × gender of aid) that varied across the independent variable of the reliability of aid.

Although highly reliable automation is desired, many automated systems would be classified as having moderate reliability. In light of our findings, moderately reliable automation would activate user stereotypes and subsequently affect the user’s trust ratings. This finding shows the necessity of proper use of stereotypes in automation, specifically when it is ambiguous in reliability. Designing automation to contain anthropomorphic aids that activate users’ stereotypes could be a future area of dispute and present difficult questions. For example, if automation is only moderately reliable should there be an anthropomorphic aid that may cause users’ trust to increase for this automation? This question may have to be answered according to the context in which the automation is aiding the user, and the consequences associated with following the aid. A future study could examine the current study’s finding of a higher level of trust in the older female aid in low reliability conditions. Potentially, there is a “motherly” aspect to some aids that cause trust when the conditions warrant this behavior. The current study has provided more evidence that HCI is similar to human-human interaction, and that the interaction between stereotypes and individuating information (e.g., automation reliability) is an area rich for exploration.

REFERENCES

- Benbasat, I., Dimoka, A., Pavlou, P. A., & Qiu, L. (2010). Incorporating social presence in the design of the anthropomorphic interface of recommendation agents: Insights from an fMRI study. *ICIS 2010 Proceedings Paper 228*.
http://aisel.aisnet.org/ficis2010_submissions/228.
- Byrne, D. E. (1971). *The attraction paradigm* (Vol. 11). Academic Pr.
- Casper, C., Rothermund, K., & Wentura, D. (2011). The activation of specific facets of age stereotypes depends on individuating information. *Social Cognition, 29*(4), 393-414.
- DeArmond, S., Tye, M., Chen, P. Y., Krauss, A., Apryl Rogers, D., & Sintek, E. (2006). Age and gender stereotypes: New challenges in a changing workplace and workforce. *Journal of Applied Social Psychology, 36*(9), 2184-2214.
- Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L.A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 44*(1), 79-94.
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864-886.
- Fiske, A. P., Kitayama, S., Markus, H. R., & Nisbett, R.E. (1998). The cultural matrix of social psychology. *The handbook of social psychology, 1 & 2*(4), 915-981. New York, NY, US: McGraw-Hill, x, 1085 pp.
- Fogg, B. J., & Nass, C. (1997). Silicon sycophants: The effects of computers that flatter. *International Journal of Human Computer Studies, 46*(5), 551-561.
- Hilton, J. L., & Von Hippel, W. (1996). Stereotypes. *Annual review of psychology, 47*(1), 237-271.
- Katagiri, Y., Nass, C., & Takeuchi, Y. (2001). Cross cultural studies of the computers are social actors paradigm: The case of reciprocity. *Usability evaluation and interface design: Cognitive engineering, intelligent agents, and virtual reality, 1558-1562*.
- Kite, M. E., Deaux, K., & Miele, M. (1991). Stereotypes of young and old: Does age outweigh gender?. *Psychology and Aging, 6*(1), 19-27.
- Kite, M. E., Stockdale, G. D., Whitley, B. E., & Johnson, B. T. (2005). Attitudes toward younger and older adults: An updated meta-analytic review. *Journal of Social Issues, 61*(2), 241-266.
- Kunda, Z., & Sherman-Williams, B. (1993). Stereotypes and the construal of individuating information. *Personality and Social Psychology Bulletin, 19*(1), 90-99.
- Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel constraint satisfaction theory. *Psychological Review, 103*(2), 284-308.
- Lee, E. J. (2003). Effects of "gender" of the computer on informational social influence: The moderating role of task type. *International Journal of Human-Computer Studies, 58*(4), 347-362.
- Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human automation interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*(2), 194-210.
- Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence, 72-78*.
- Posthuma, R. A., & Campion, M. A. (2009). Age stereotypes in the workplace: Common stereotypes, moderators, and future research directions?. *Journal of Management, 35*(1), 158-188.
- Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics, 55*(9), 1059-1072.
- Singer, M. S. (1986). Age stereotypes as a function of profession. *The Journal of social psychology, 126*(5), 691-692.
- Singh, I. L., Molloy, R., & Parasuraman, R. (1993). Automation induced "complacency": Development of the complacency-potential rating scale. *The International Journal of Aviation Psychology, 3*(2), 111-122.
- Wickens, C.D., & Dixon, S.R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science, 8*(3), 201-212.
- Zanbaka, C., Goolkasian, P., & Hodges, L. (2006). Can a virtual cat persuade you?: The role of gender and realism in speaker persuasiveness. In *Proceedings of the SIGCHI conference on Human Factors in computing systems, 1153-1162*.

Faces as Ambient Displays: Assessing the attention-demanding characteristics of facial expressions

Brock M. Bass & Richard Pak
Clemson University
Department of Psychology

Ambient displays are used to provide information to a user in a non-distracting manner. The purpose of this research is to examine the efficacy of facial expressions as ambient displays. Facial emotion recognition requires very little if any conscious attention, which makes it an excellent candidate for the ambient presentation of information. This study will investigate whether using facial expressions as an ambient display permits humans to gain information with ease. This study will assess the attention-demanding characteristics of Chernoff faces in a dual-task experiment. The data from this study could be helpful in understanding whether humans are able to use facial expressions for gaining quick and concise information about a particular system or device.

INTRODUCTION

Ambient displays convey information to the user without being very cognitively demanding—they are in the background. For example, the battery meter icon of a computer interface, or a dangling string from the ceiling to represent network traffic on a computer network (Weiser & Brown, 1995). Some important characteristics of ambient displays are: useful and relevant information, sufficient information design, consistent and intuitive mapping, and the match between the system and the real world (Mankoff, Dey, Hsieh, Kientz, Lederer, & Ames, 2003). Using these heuristics as a benchmark, facial expressions could be considered a type of ambient display. The purpose of this study is to examine the ambient quality of facial expressions; that is to measure their attention-demanding qualities when conveying simple numerical information. We will study this in the context of user-system automation calibration.

When users are interacting with computerized decision support systems or automated aids, the user must, over time, determine how much they should trust the system. Optimally, the user would calibrate their trust to match the level of actual system reliability. That is, to be highly trusting of a highly reliable automated system, or distrusting of a very unreliable system (Parasuraman, 1997). However, this scenario of human-automation interaction (HAI) can be problematic in some cases. For example, an operator may place too much trust in unreliable automation, also known as misuse of automation. Conversely, an operator may not place enough trust in reliable automation, which can lead to disuse of automation. An operator's misuse or disuse of automation is a function of their level of trust, which is a byproduct of their perceptions about the reliability of the automation (Parasuraman, 1997). The goal of this study is to determine if increasing the deducibility and transparency of trial-level automation reliability can enhance users ability to judge overall system reliability, and thus calibrate trust. This transparency of automation reliability may allow operators to interact with ambient displays more appropriately.

One plausible way an automated system can present more transparent information about its own reliability is if the

system presented its own confidence in its recommendation. This concept can be categorized in the ambient display heuristic of useful and relevant information. Many automated systems, particularly of the decision support type, are able to present to the user their level of confidence in the automated advice. For example, if the system is working from faulty data, it will weigh its advice as potentially unreliable. The exchange of information, in this case system confidence, is a way of diminishing the uncertainty that can exist in HAI (Bubb-Lewis & Scerbo, 1997). Trust is a malleable variable that can be shaped through interactions with a system (Antifakos, Kern, Schiele, & Schwaninger, 2005). If a system is presenting the operator with its system confidence level, then the operator should be able to build a more appropriate relationship with the automation. Some previous research has indicated that methods such as tactile output or auditory output may be helpful in conveying system confidence (Wisneski, 1999; Poupyrev, Maruyama, & Rekimoto, 2002; Sawhney & Schmandt, 2000). While these modalities are novel in certain capacities, a less intrusive and less attention demanding modality would be more beneficial to users (Antifakos, Kern, Schiele, and Schwaninger, 2005).

One novel information presentation format is the use of facial expressions. An interesting area of facial expression research involves Chernoff faces (Chernoff, 1973). These faces were created to represent multivariate data in a way that would allow the viewer to gain information in a quick, yet complete manner. Chernoff (1973) makes a point that humans are accustomed to viewing and interpreting faces. Differences in the configuration of a face, even small ones, can be noticed by humans (Chernoff, 1973). If this statement is in fact true, facial expression may act as a superb source of information output. Previous studies have investigated the effectiveness of Chernoff faces with mixed results. A previous study concluded that Chernoff faces are not processed pre-attentively, and do not benefit users more than other modes of visual information display (Morris, Ebert, & Rheingans, 2000). The process of identifying the characteristics (i.e., eye brow slant, eye size, nose length) of the Chernoff face was said to be a serial process (Morris, Ebert, & Rheingans, 2000). A similar study investigating perceptual sensitivities for

Chernoff faces found that children process Chernoff faces differently from adults (Tsurusawa, Goto, Mitsudome, Nakashima, & Tobimatsu, 2008). Children focus more on individual features, while adults process a face holistically (Tsurusawa, Goto, Mitsudome, Nakashima, & Tobimatsu, 2008). It was found that people encode the meaning of a face through the perceptual features of the face. Specifically, the eye brows and mouth are important for this encoding (Morris, Ebert, & Rheingans, 2000; Tsurusawa, Goto, Mitsudome, Nakashima, & Tobimatsu, 2008). If Chernoff faces are manipulated properly, giving the right amount of useful information, they will fulfill the heuristic of sufficient information design as an ambient display.

Human Emotion Decoding

Research has shown that humans are able to automatically recognize emotion through facial expressions. Given this information, using facial expressions as ambient displays would not add cognitive load and would enforce the heuristic of consistent and intuitive mapping. Studies have shown that tasks involving affective (emotional) stimuli may be responded to without awareness (Whalen, 1998; Morris, 1998). For example, it was found that the amygdala seems to have an automatic response to facial expressions. Data from the fMRI confirmed that participants experienced an increase in amygdala activation during the experiment (Whalen, 1998). This indicates that even though participants were unaware of the presentation of emotional facial expressions, they still processed this information. The conclusions of this study make a case that explicit knowledge is not necessary for a person to process emotional facial expressions. This process is done below the level of conscious awareness, or in other terms, automatically (Whalen, 1998).

Neuroimaging studies have supported the notion that emotional processing of faces is a more effective pathway than the processing of other stimuli. A previous study compared the automatic processing of emotional facial expressions versus emotional words. Rellecke (2011) hypothesized that facial expressions would be decoded more automatically than words, due to their perceptual features and human's natural ability to decode them. Based on the results of emotion-related brain potentials (ERPs), facial expressions were found to have a prolonged effect on the brain. This finding alludes to emotional facial expression processing as being automated to a higher extent versus emotional word processing (Rellecke, 2011). One point that this study also discussed is how there may be preconditions that are necessary for advanced automatic processing of emotional words. The two stimuli were tested in the same superficial stimulus analysis task, but only one (facial expression) led to advanced automatic processing. Facial expression seems to be a stimulus that needs no prompting or preconditions to allow fast, but also meaningful processing (Rellecke, 2011). With indications that facial expressions are a more effective pathway for the decoding of emotional data, we want to investigate the limits and capabilities of this potentially new modality for information transport.

In order for facial expression to be used as a means of relaying quantitative system/automation information, we must know if users are able to properly and consistently decode facial expression intensity into a consistent quantitative value (e.g., an intense smiling face represents 90%). Hess (1997) did a research study which investigated the issue of facial expression decoding with varying degrees of intensity for different emotional categories. It was determined that when participants were given an emotional facial expression stimulus, they were accurate at perceiving the stimulus' physical intensity. Graphically, this means that there is a positive linear trend for the perceived intensity of the expression by the human versus the actual physical intensity of the emotional facial expression (Hess, 1997). Understanding the effects that different emotional facial expressions and their intensities have on human's ability to decode is critical in determining the most effective stimuli to use as ambient displays.

Age-Related and Cultural Effects on Decoding

Despite the ease with which humans are able to decode emotional facial expressions, it is still moderated by age and cultural aspects. Age can alter a person's ability to correctly perceive and understand the facial expression that is presented before them. Neuropsychological research has shown that age-related issues in facial expression decoding may be a result of problems with the medial temporal lobe (Orgeta, 2007). The amygdala is housed here, which corroborates previous research that suggests the amygdala is necessary for facial expression decoding (Whalen, 1998; Morris, 1998). There is an interesting paradox that has been asserted for older adults involving their ability to decode emotional facial expressions. According to the socioemotional selectivity theory, older adults are actually more aware of certain emotional situations and images than non-emotional (Orgeta, 2007).

Some studies yielded results that showed older adults as being more aware of positive facial expressions, but not negative facial expressions (Orgeta, 2007). The results of this study indicated that there is an age-related difference when decoding positive versus negative facial expressions. Orgeta (2007) found that for the facial expressions of sadness and fear, there was not a larger age-effect based on the expression being higher in perceptual cost. An image that was only 50 % expressive did not show larger age-related effects than a 100 % expressive image. This compliments previous research because it indicates that the major issue in age-related decline with facial expression decoding comes from cognitive decline and not perceptual decline (Orgeta, 2007). Another issue that affects the decoding of facial expression is culture. There are six basic emotions that transcend culture. They are: anger, happiness, fear, surprise, disgust, and sadness (Ekman & Friesen, 1975). These emotions can be represented with facial expressions and are readily recognizable (Lee, 2006; Batty, 2003). Because these facial expressions are not confined to specific cultures, it puts no restraints on the ability of different people groups to successfully decode these facial expressions (Ekman & Friesen, 1971). It appears that increasing age is a

factor that may cause difficulty in facial expression decoding, while culture seems to be of no hindrance. Due to facial expressions prevalence and familiarity in human culture, making them an ambient display allows the heuristic of matching the system to the real world to be met.

LIMITATIONS OF PREVIOUS LITERATURE

The previous literature has provided a foundation for knowledge about facial expressions, but there are limitations to these studies. The Hess (1997) study presented emotional facial expressions in a single-task format. The participants viewed the image and rated it on the emotionality and intensity that they perceived. This methodology does not clarify whether facial emotion decoding is truly resource/attention-free as neuropsychological studies suggest. A dual-task design should be implemented to properly measure attention usage. In order to gain this data, measures of response time, accuracy, and subjective workload should be used. The Hess (1997) study also measured decoding accuracy for each facial expression image through the presentation of several emotion scales. The participant was presented with seven emotional labels, which they manipulated to show the intensity of emotion for the previous picture. Instead of presenting seven individual scales, it seems to be less complicated to present one scale or to have a quick input device (keyboard number keys) after the image is viewed. The Hess (1997) study presented facial expression intensity in increments of 20 % intensity. This intensity scale may not provide a complete spectrum of facial expression decoding data. The Orgeta (2007) study also presented only four intensity levels. The number of intensity levels may need to be increased to capture a more accurate representation of people's ability to decode facial expression. Another limitation in the Orgeta (2007) study was the facial images were presented in increasing order as the participant advanced through the experiment. This method may have led to participants forming an anticipation bias that the next facial image was going to be more expressive. The purpose of the current study is to examine the user's ability to accurately decode quantitative value from a facial expression. Previous studies have looked at human's ability to properly decode facial expression type (Ekman & Friesen, 1975; Ekman & Friesen, 1971), intensity (Tsurusawa, Goto, Mitsudome, Nakashima, & Tobimatsu, 2008; Hess 1997), and the effectiveness of Chernoff faces (Chernoff 1973; Tsurusawa, Goto, Mitsudome, Nakashima, & Tobimatsu, 2008; Morris, Ebert, & Rheingans, 2000). However, no study to date has fused these previously listed concepts into one holistic study; this is the intent of the current study.

OVERVIEW OF THE STUDIES

The current study will model itself in some areas after Hess (1997). However, our study will use the dual-task paradigm to precisely measure the attention-demanding characteristics of facial displays. The current study will use only one measurement scale (direct key entry) after each trial to eliminate any confusion for the participants about what the

scales are measuring. This will also allow for more precise response time data. In the Orgeta (2007) study the facial expressions were shown in increasing order. Chernoff facial expression stimuli will be shown in randomized intensity order in an effort to avoid any biases being formed by the participants. The Chernoff faces will be manipulated differently compared to previous research (Chernoff, 1973; Tsurusawa, Goto, Mitsudome, Nakashima, & Tobimatsu, 2008; Morris, Ebert, & Rheingans, 2000). Only the mouth will be manipulated in order to gain understanding about the affect of this one variable on decoding. Finally, the current study will use a facial expression intensity scale more precise than previous research (Hess, 1997; Orgeta, 2007). A facial expression scale presenting emotions in increments of 10 % will be used. Our hypothesis is that by making these modifications the current study will be able to address the research question with more accuracy.

METHOD

Participants

There will be 80 participants (40 younger adults, 40 older adults) tested for the current study. The age range for younger adults will be 18-24 years old, while the age range for older adults will be from 65-85 years old.

Design

This study will be a between-subjects, 2 (age group) x 2 (facial expression type) x 10 (facial expression intensity) factorial design. The dependent variables being measured are: the speed (ms) for the block task, the proficiency on the block task (amount of blocks cleared), the speed (ms) of response on the facial expression task, the amount of "misses" on the facial expression task, and the accuracy of response for the facial expression rating. Measures of subjective workload will be collected with the NASA-TLX and individual cognitive ability data will be collected with a battery of cognitive abilities tests.

Task and Materials

Participants will view the program on 19-inch LCD monitors and make all responses using the keyboard. They will be seated in office chairs about 18-24 inches from the screen in an office environment. Participants will initially take a computerized cognitive abilities test. These tests include the digit symbol test, reverse digit span test, and the Shipley vocabulary test. These tests will gather information on individual abilities such as working memory, perceptual speed, and vocabulary.

The primary task will be to play a block game similar to the game Tetris. The block game consists of moving multi-colored blocks. The main objective of the block task is to manipulate the blocks, and successfully clear them using the arrow keys and space bar. The blocks appear on the screen (moving from bottom to top) as the participant interacts with the program.

The secondary task will be to identify the level of emotion presented on a computer-generated character. The facial expression stimuli were rendered using the statistical program R. This allowed the experimenter to have systematic control over the faces and increase their facial expression intensity as desired. The facial expression stimuli are basic line drawings composed of black lines on a white background. This was done to eliminate any confounding variables that could appear due to gender, ethnicity, or age. There are 19 stimuli total: 9 happy stimuli (ranging from 10% expressive – 90% expressive), 9 sad stimuli (ranging from 10% expressive – 90% expressive), and one neutral stimulus (0% expressive). The dimensions of the stimuli are 170 pixels by 250 pixels.



Figure 1. 50 % happy Chernoff face.

Procedure

Participants will be randomly assigned to experimental conditions prior to the experiment. The participants will be given an informed consent document before any testing is conducted. The participant will then take a battery of cognitive abilities tests. Next, the experiment will be presented in three phases. The participants will complete two separate single-tasks (block task and facial expression task) to record baseline data on their abilities in each task. To examine the attentional demands of decoding Chernoff faces, participants will then engage in a dual-task. The primary task will be the block task. This spatial-manipulation task will be relatively cognitively taxing on the participants. The secondary task will be the facial expression task. This task will presumably be fairly automatic for the participants and will require little to no cognitive resources.

In phase 1, the participant will perform the block task in a single-task environment. The participant will have to reach a pre-set score (based on number of successful manipulations) to complete the task. In phase 2 of the experiment, the participants will be asked to respond to facial expressions that are flashed on the computer screen. The participant will be in one of two facial expression conditions (happy or sad). Once phase 2 begins, the facial expression will appear in the window for three to five seconds and then disappear. The facial expressions will be shown in a randomized order in regard to their intensity level. During this time interval the participant will try to respond to the facial expression using the number keys. If the participant does not hit a number key before time has elapsed then a “miss” will be recorded. Regardless of whether the participant has responded or missed making a response, after three to five seconds (randomized appearance time) the screen will go back to being blank until the next trial. There will be 60 trials in each condition. In phase 3, the participant will be exposed to both

phases 1 and 2 simultaneously. This will create a dual-task situation. After the participant has completed the experiment the computer will display the NASA-TLX questionnaire for completion.

PREDICTED RESULTS

The first hypothesis (H_{1a}) is that participant's performance (i.e., accuracy and speed) will be above chance levels for facial emotion decoding in the single-task phase. We are assuming that the younger and older participants will be able to rely on previously acquired innate facial expression knowledge to achieve high accuracy decoding. We expect to see a linear trend between the actual intensity of the emotional facial expression presented and the perceived intensity of the emotional facial expression. This hypothesis is based on the results of Hess (1997) and Orgeta (2007). The second part of our hypothesis (H_{1b}) is that all participant's performance on the facial decoding task will be affected due to the dual-task environment. As a consequence of divided attention, we expect facial expression response time and misses to increase in the dual-task environment. However, the current study hypothesizes that one condition will present itself as more decodable to the participant in the dual-task. This is based on the supposed automatic nature of the facial expression task and also the effects of facial expression type on decoding. If this modality is actually resource-free and allows for ease of decoding as some studies indicate (Whalen, 1998; Lee, 2006; Morris, 1998), then dual-task performance should not significantly deviate from single-task performance for the participant when the most effective facial expression type is presented, which we hypothesize to be the happy condition.

The second hypothesis (H_2) is that all participants will show a difference in facial recognition accuracy scores between the two conditions (i.e., happy and sad). We are expecting a main effect for condition. It is hypothesized that participants will have larger actual versus perceived differences (i.e., worse facial expression recognition) for the sad condition. This hypothesis is drawn from the socioemotional selectivity theory and research which supports positive expressions as more identifiable, referred to as the “happy face advantage”. (Orgeta, 2007; Ekman & Friesen, 1971).

The third hypothesis (H_3) is that the variables of age and condition will have a significant effect on the accuracy of the participants. We are expecting a two-way interaction between age and emotion of presented face on accuracy. Thus, when older adults are in the happy condition their accuracy scores will not be significantly different (i.e., differences in actual and perceived facial expression intensities) than younger adults in the happy condition. However, we expect to see younger adults produce significantly better performance scores in the sad condition versus the older adults in the sad condition. This hypothesis is driven by the socioemotional selectivity theory. It is expected that older adults will have significantly better performance in the happy condition than in the sad condition, while younger adults will yield less significant performance differences between the conditions.

DISCUSSION

The current research study is attempting to clarify the issue of whether emotion can be used as a reliable and resource-free modality. This research is relevant not just for applied psychology, but also for the pool of knowledge in psychology used for future studies. The potential importance of this research will continue to increase as the use of ambient displays, automation, and system confidence increases in our society. The ability for facial expression to be used in automation as a viable communication tool may be similar to that of the visual and spatial modalities presented in multiple resource theory (Lee, 2006). If the current study discovers that the emotional modality is capable of reliable and efficient information processing during a dual-task situation, then the concept of emotionally transparent automation may be implemented in future automation. One of the main goals of HAI is easily understood output from the automation for the human to interpret. This allows the human to understand automation behavior and predict future behavior. It has been noted that this process can be hindered by advanced automation. To help alleviate this disconnect between the automation and human, the use of facial expressions could bring interpretation clarity for the human (Lee, 2006). This clarity would allow for properly calibrated trust to be formed between the human and automation, and ultimately allow the human to have a realistic idea of the system's confidence and interact with it accordingly. This is important because automation that is assisting in critical situations (health management, aviation, nuclear power plants, etc.) needs to be trusted and used properly by the user. Widening the research question back out to ambient displays, it is evident that many domains could benefit from research explaining how facial expressions aid in information display. One interesting point proposed by Lee (2006) is the lack of attention that is required for emotional stimuli processing. The current study is trying to build on this idea and show that the use of facial expressions to deliver information requires almost no attention from the human. This finding would give evidence that human's are already equipped with a resource-free modality that can be used to gain information. One potential benefit of an innate modality for information processing would be the little to no training required for people to properly access this tool. Due to the large variety of users for most systems, implementing effective ambient displays can be difficult. However, if facial expression decoding proves to be an effective information processing method, then it could be critical to making ambient displays successful across demographic categories. The current study could be used to show a unifying aspect of human information processing that could be applied to research in multiple disciplines. In sum, the current study may find that the key to creating a viable ambient display is found within the human brain. To capitalize on this fact, ambient displays should be designed to display emotional facial expressions to take advantage of this untapped modality.

REFERENCES

- Antifakos, S., Kern, N., Schiele, B., & Schwaninger, A. (2005). Towards improving trust in context-aware systems by displaying system confidence. Proceedings from MobileHCI '05: *The 7th International Conference on Human Computer Interaction with Mobile Devices & Services*. Salzburg, Austria.
- Batty, M., & Taylor, M. J. (2003). Early processing of the six basic facial emotional expressions. *Cognitive Brain Research*, 17, 613-620.
- Bubb-Lewis, C., & Scerbo, M. (1997). Getting to know you: Human computer communication in adaptive automation. In M. Mouloua & J. M. Koonce (Eds.), *Human-automation interaction: Research and practice* (pp. 92-99). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Chernoff, H. (1973). The use of faces to represent points in k dimensional space graphically. *Journal of the American Statistical Association*, 68(342), 361-368.
- Ekman, P. & Friesen, W.V. (1971). Constraints across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2), 124-129.
- Ekman, P. & Friesen, W.V. (1975). *Unmasking the face: A guide to recognizing emotions from facial cues*. Oxford, England: Prentice-Hall.
- Hess, U., Blairy, S., & Kleck, R. E. (1997). The intensity of emotional facial expressions and decoding accuracy. *Journal of Nonverbal Behavior*, 21(4), 241-257.
- Lee, J. D. (2006). Affect, attention, and automation. In A. Kramer, D. Wiegmann & A. Kirlik (Eds.), *Attention: From theory to practice*. New York: Oxford University Press.
- Mankoff, J., Dey, A. K., Hsieh, G., Kientz, J., Lederer, S., Ames, M. (2003). Heuristic evaluation of ambient displays. Proceedings from SIGCHI '03: *Conference on Human Factors in Computing Systems*. Ft. Lauderdale, FL, USA.
- Morris, C. J., Ebert, D. S. & Rheingans, P. (2000). An experimental analysis of the effectiveness of features in chernoff faces. Proceedings from SPIE '00: *The 28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making*. Washington, D. C., USA.
- Morris, J. S., Friston, K. J., Buchel, C., Firth, C. D., Young, A. W., Calder, A. J., & Dolan, R. J. (1998). A neuromodulatory role for the human amygdala in processing emotional facial expressions. *Brain*, 121, 47-57.
- Orgeta, V., & Phillips, L. H. (2007). Effects of age and emotional intensity on the recognition of facial emotion. *Experimental Aging Research*, 34(1), 63-79.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230-253.
- Poupyrev, I., Maruyama, S., & Rekimoto, J. (2002). Ambient touch: Designing tactile interfaces for handheld devices. Proceedings from UIST '02: *The 15th Annual ACM Symposium on User Interface Software and Technology*. Paris, France.
- Rellecke, J., Palazova, M., Sommer, W., & Schacht, A. (2011). On the automaticity of emotion processing in words and faces: Event-related brain potentials evidence from a superficial task. *Brain and Cognition*, 77, 23-32.
- Sawhney, N., & Schmandt, C. (2000). Nomadic radio: Speech and audio interaction for contextual messaging in nomadic environments. *ACM Transactions on Computer-Human Interaction*, 7(3), 353-383.
- Tsurusawa, R., Goto, Y., Mitsudome, A., Nakashima, T., Tobimatsu, S. (2008). Different perceptual sensitivities for Chernoff's face between children and adults. *Neuroscience Research*, 60(2), 176-183.
- Weiser, M., & Brown, J. S. (1995). Designing calm technology. Retrieved from <http://www.ubiq.com/weiser/calmtech/calmtech.htm>.
- Whalen, P. J., Rauch, S. L., Etkoff, N. L., McInerney, S. C., Lee, M. B., & Jenike, M. A. (1998). Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *The Journal of Neuroscience*, 18(1), 411-418.
- Wisneski, C. (1999). The design of personal ambient displays. (Unpublished master's thesis). MIT, Boston, Ma.

APPENDIX 7: Bass, B. (2014). Faces as Ambient Displays: Assessing the Attention-Demanding Characteristics of Facial Expressions. Unpublished master's thesis. Available at: http://tigerprints.clemson.edu/all_theses/1941/

Clemson University
TigerPrints

All Theses

Theses

5-2014

Faces as Ambient Displays: Assessing the Attention-Demanding Characteristics of Facial Expressions

Brock Bass

Clemson University, brockb@g.clemson.edu

Follow this and additional works at: http://tigerprints.clemson.edu/all_theses

 Part of the [Psychology Commons](#)

[Please take our one minute survey!](#)

Recommended Citation

Bass, Brock, "Faces as Ambient Displays: Assessing the Attention-Demanding Characteristics of Facial Expressions" (2014). *All Theses*. Paper 1941.

This Thesis is brought to you for free and open access by the Theses at TigerPrints. It has been accepted for inclusion in All Theses by an authorized administrator of TigerPrints. For more information, please contact awesole@clemson.edu.

FACES AS AMBIENT DISPLAYS: ASSESSING THE ATTENTION-DEMANDING
CHARACTERISTICS OF FACIAL EXPRESSIONS

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Applied Psychology

by
Brock M. Bass
May 2014

Accepted by:
Dr. Richard Pak, Committee Chair
Dr. Leo Gugerty
Dr. Christopher Pagano

ABSTRACT

Ambient displays are used to provide information to users in a non-distracting manner. The purpose of this research was to examine the efficacy of facial expressions as a method of conveying information to users in an unobtrusive way. Facial expression recognition requires very little if any conscious attention from the user, which makes it an excellent candidate for the ambient presentation of information. Specifically, the current study quantified the amount of attention required to decode and recognize various facial expressions. The current study assessed the attention-demanding characteristics of facial expressions using the dual-task experiment paradigm. Results from the experiment suggest that Chernoff facial expressions are decoded with the most accuracy when happy facial expressions are used. There was also an age-effect on decoding accuracy; indicating younger adults had higher facial expression decoding performance compared to older adults. The observed decoding advantages for happy facial expressions and younger adults in the single-task were maintained in the dual-task. The dual-task paradigm revealed that the decoding of Chernoff facial expressions required more attention (i.e., longer response times and more face misses) than hypothesized, and did not evoke attention-free decoding. Chernoff facial expressions do not appear to be good ambient displays due to their attention-demanding nature.

ACKNOWLEDGEMENTS

Without my relationship with my personal Lord and Savior Jesus Christ, I truly would be devoid of my life's true meaning and joy. Therefore, I acknowledge Him in all I ever attempt to accomplish. During my time at Clemson University, I have been fortunate to encounter professors who have inspired and encouraged me. Consequently, I am deeply thankful to have as my mentor and advisor, Dr. Richard Pak, who not only served as my thesis committee chairman, but also my colleague and friend for these past three years. My committee members, Dr. Leo Gugerty and Dr. Christopher Pagano, made my committee complete and I am grateful for their willingness to serve and assist. I'm thankful for all of my fellow graduate students (past and present) who have truly become my friends and family. Finally, I want to thank my parents, Drs. John and Deborah Bass, as well as my sister, Alexis, for all their love and support throughout my life and especially while I've studied at Clemson.

TABLE OF CONTENTS

	Page
TITLE PAGE	i
ABSTRACT	ii
ACKNOWLEDGEMENTS	iii
LIST OF FIGURES	vi
INTRODUCTION	1
Human Emotion Decoding	3
Chernoff Faces	7
Age Related and Cultural Effects on Decoding	11
Limitations of Previous Literature	14
Overview of the Current Study	16
Hypothesis of the Current Study	17
METHODS	18
Participants	18
Design	19
Materials	19
Surveys & Abilities	19
Tasks	20
Procedure	20
RESULTS	22
Phase 2 (Single-task, Facial Expression Decoding Only)	24
Phase 3 (Dual-task, Block Task and Facial Expression Decoding)	33
DISCUSSION	48
CONCLUSION	58
APPENDICIES	60
A: Screenshot of Block Game Task (Phase 1)	61

Table of Contents (Continued)

	Page
B: Screenshot of Facial Expression Decoding Task (Phase 2)	62
C: Chernoff Facial Expression Stimuli Organized by Expression and Intensity	63
D: Screenshot of Block Task and Facial Expression Decoding Task (Phase 3)	64
REFERENCES	65

LIST OF FIGURES

Figure		Page
1	Mean intensity key pressed by facial expression condition for younger and older adults	27
2	Mean intensity key pressed by face presented for younger and older adults	28
3	Mean intensity key pressed by face presented for sad and happy facial expression conditions	29
4	Mean response time (ms) by age group for sad and happy facial expression conditions	31
5	Mean number of face misses by age group for sad and happy facial expression conditions	32
6	Mean intensity key pressed by task phase for younger and older adults	37
7	Mean intensity key pressed by face presented for younger and older adults (dual-task)	39
8	Mean intensity key pressed by face presented for sad and happy facial expression condition (dual-task).....	39
9	Mean intensity key pressed by face presented for younger and older adults (single-task, happy facial expression condition).....	40
10	Mean intensity key pressed by face presented for younger and older adults (dual-task, happy facial expression condition).....	41
11	Mean intensity key pressed by face presented for younger and older adults (single-task, sad facial expression condition)	42
12	Mean intensity key pressed by face presented for younger and older adults (dual-task, sad facial expression condition).....	42
13	Mean response time (ms) by task phase	43

List of Figures (Continued)

Figure		Page
14	Mean response time (ms) by age group	44
15	Mean number of face misses by task phase	45
16	Mean blocks cleared by age group.....	46
17	Mean computed workload by age group for sad and happy facial expression conditions	47

INTRODUCTION

Ambient displays can take many forms. For example, the battery meter icon of a computer interface, or a dangling string from the ceiling to represent network traffic on a computer network (Weiser & Brown, 1995). These examples are considered “ambient” because they convey information to the user without being substantially taxing on cognitive faculties (i.e., they are in the background and do not require the user to change focus or switch attention). Several important characteristics have been identified for the design of a good ambient display. Examples of these characteristics include: providing useful and relevant information, having a sufficient information design, using consistent and intuitive mapping, and appropriate matching between the system and the real world (Mankoff, Dey, Hsieh, Kientz, Lederer, & Ames, 2003). If these characteristics are adequately fulfilled by facial expressions, then facial expressions could be considered a good form of ambient display. The purpose of this study is to determine if face stimuli can serve as ambient indicators of quantitative information.

One situation where ambient displays may be helpful is in human-automation interaction (HAI). In some HAIs, users may become unaware of the hidden decision making processes or outcomes of automation. They may also lose track of the automation’s reliability over time (i.e., forget how reliable or unreliable it has been in the past). Such information (uncertainty of current processes, past reliability) can lead to fluctuations in trust that may not be justified (un-calibrated trust); that is trust that may be unwarranted. Un-calibrated trust can manifest itself as continued use of unreliable

automation (misuse) or unwarranted discontinued use of reliable automation (disuse) both of which cause non-optimal HAIs (Parasuraman, 1997).

One way in which an automated system can encourage proper calibration is by presenting as much information about its operation as possible. For example, it could present its own confidence in its recommendation, so called “system confidence”, or it could present a historical picture of its own reliability (both are information that are easily accessible by a system). This concept can be categorized in the ambient display heuristic of useful and relevant information. For example, if the system is working from faulty data, it will weight its advice as potentially unreliable. Presenting critical information, such as system confidence, is a way of diminishing the uncertainty that can exist in HAIs (Bubb-Lewis & Scerbo, 1997). Trust is a malleable variable that can be shaped through interactions with a system (Antifakos, Kern, Schiele, & Schwaninger, 2005). If a system is presenting the operator with its system confidence level, then the operator will be able to build a more appropriate trust relationship with the automation. However, this presentation needs to be salient and the automation state indicator should not add attentional demands to the user (Parasuraman, 1997). Some previous research has indicated that methods such as tactile output and auditory output may be helpful in conveying system confidence (Wisneski, 1999; Poupyrev, Maruyama, & Rekimoto, 2002; Sawhney & Schmandt, 2000). While these modalities are novel in certain capacities, a less intrusive and less attention demanding modality would be more beneficial to users. Thus, the ideal stimulus display type would be one that provides the user with meaningful information, while not becoming a distraction or a drain on the

user's attention (Antifakos, Kern, Schiele, and Schwaninger, 2005). Coding information as emotional expression in human-like faces may fulfill this role.

Human Emotion Decoding

Research has shown that humans have an ability to recognize emotional facial expressions with little attention allocation. Batty and Taylor (2003) had participants complete an implicit emotional task, which involved the presentation of target stimuli (non-faces) in a sequence with emotional faces. This experimental design allowed the researchers to test the participants' event-related potentials (ERPs) while viewing emotional faces, but without explicitly instructing the participant to look at the emotional faces. Through analysis of the ERPs, it was found that participants were processing the emotional face stimuli quickly (i.e., $M = 94$ ms for P1 component; $M = 140$ ms for N170 component). The results of this analysis of the P1 and N170 components suggest that participants were processing the emotional face stimuli pre-attentively (Batty & Taylor, 2003). Other studies have supported that tasks involving affective (emotional) stimuli may be responded to without awareness (Whalen, 1998). An fMRI study showed that participants experienced increased amygdala activation even when they were unaware of the presentation of emotional facial expressions (Whalen, 1998). The amygdala is a key area of the brain for the emotional facial recognition process. Previous research on animals has provided evidence that the amygdala is the brain area where facial and emotional processing occurs. A subsequent study built off of these findings and found the amygdala was crucial for humans' decoding of facial affect, especially the emotion of fear (Adolphs, Tranel, Damasio, & Damasio, 1994). The conclusions of Whalen (1998)

make a case that explicit knowledge is unnecessary for a person to process emotional facial expressions. This process occurs below the level of conscious awareness, or in other terms, automatically (Morris, 1998; Whalen, 1998). It can be inferred from these studies, that the use of facial expressions as ambient displays should not add cognitive load and would enforce the heuristic of consistent and intuitive mapping.

Neuroimaging studies have supported the notion that the emotional processing of faces is a more effective pathway than the processing of other stimuli. A previous study compared the automatic processing of emotional facial expressions versus emotional words. Rellecke (2011) hypothesized that facial expressions would be encoded more automatically than words, due to their perceptual features and humans' natural ability to encode them. This study was novel because it took two theoretically attention-free emotional processing stimuli (i.e., faces and words), and compared their efficiency and effect. The degree of encoding automaticity was being tested for each of these stimuli. Based on the results of the electroencephalogram (EEG), the event-related brain potentials (ERPs) recorded for the facial expression conditions were found to have a prolonged effect on the brain. This finding alludes to emotional facial expression processing as being automated to a higher extent than emotional word processing. Rellecke (2011) discusses the potential necessity for preconditions for the high automatic processing of emotional words. This was apparent because the two stimuli were tested in the same superficial stimulus analysis task, but only one (i.e., facial expression) led to advanced pre-attentive processing. Facial expression seems to be a stimulus that needs no prompting or preconditions to allow fast, but also meaningful processing (Rellecke,

2011). Data analysis found that happy faces were decoded earlier than other faces (i.e., 50-100 ms). This supports the theory that happy faces are advantageous in the early stages of emotional processing and may be instrumental in attention-free encoding. Also, data showed that angry faces were advantageous for later decoding (i.e., 150-450 ms). This coincides with previous research that states angry expressions, or threat-related expressions, have prolonged effects on the brain (Rellecke, 2011). These differences in emotion type on ERPs show that there may be a specific type of emotion that elicits faster decoding for humans.

Calvo and Lundqvist (2008) found the facial expression of happiness to be the stimuli best decoded by participants. Participants were presented with a happy facial expression and responded more accurately in its identification, and rarely mis-identified the expression as another emotion (i.e., neutral, angry, sad, disgusted, surprised, fearful). Response times for neutral and happy facial expressions were the fastest among all expressions. This indicates a fast, automatic form of facial expression decoding. Calvo and Lundqvist (2008) conducted a second experiment where the participants were exposed to the stimuli in a “fixed-pace mode”. Participants viewed the stimuli at fixed exposures of 25, 50, 100, 250, and 500 milliseconds. The results of this experiment paralleled the original findings, showing that the expression of happiness was consistently identified at a high accuracy level ($M = 98.4\%$) regardless of the exposure time. Having additional time to decode the happy expression did not result in accuracy gains. Thus, it can be inferred that humans are very quick and accurate at decoding happy facial expressions. With indications that facial expressions are an effective pathway for

the decoding of emotional data, we want to investigate the limits and capabilities of this potentially new modality for communication of quantitative information.

In order for facial expression to be used as a means of relaying quantitative system/automation information, we must know if users are able to properly and consistently decode facial expression intensity into a consistent quantitative value (e.g., a specific smiling face represents 90%). Hess (1997) investigated the issue of facial expression decoding with varying degrees of intensity for different emotional categories. When participants were given an emotional facial expression stimulus, they were accurate at perceiving its physical intensity; there was a linear trend for the perceived intensity of the expression by the human versus the actual physical intensity of the emotional facial expression (Hess, 1997). Analysis showed that when a facial expression was more intense (e.g., 80% and 100% expressive) the participant had a more accurate perception of the emotional stimulus. Happy expressions were the most recognizable across all intensity levels (Hess, 1997). This finding supports happy facial expressions as one of the most familiar and perhaps easiest of facial expressions to decode for humans. Bartneck and Reichenbach (2005) performed a similar study that sought to determine how the actual intensity of facial stimuli affected perceived intensity and accuracy. It was found that participants displayed high accuracy in perceiving happy face intensity, high recognition accuracy for happy faces, and gave low task difficult ratings for happy faces. It was also found that the happy facial expressions led to the fastest ceiling effect for recognition accuracy. Participants were able to recognize the happy facial expression starting at just 10% intensity. This reiterates quick decoding for happy facial expressions.

Understanding the effects that different emotional facial expressions and their intensities have on humans' ability to decode is critical in determining the most effective stimuli to use as ambient displays.

Chernoff Faces

Chernoff faces were created to represent multivariate data in a way that would allow the viewer to gain information in a quick, yet complete manner. For example, some of the original Chernoff faces were used to represent fossil data. The Chernoff faces displayed information pertinent to the fossils (i.e., inner diameter of embryonic chamber, total number of whorls, maximum height of chambers in last whorl, etc.) through variations including, but not limited to the faces: head shape, eye size, mouth size/shape, and eyebrow size/slant. Chernoff's rationale was that due to the extreme familiarity of faces, people would easily detect differences in the configuration of a face, even if the differences were small ones (Chernoff, 1973). It was expected that people would at least be able to examine faces more quickly than examining a row of numbers. Assuming that this is true, a schematic facial expression should act as a superb source of information output.

Chernoff faces have up to 18 characteristics that can be manipulated (Nelson, 2007). When representing multivariate data (e.g., the fossil data) it is beneficial to have multiple facial elements that can be manipulated and used for representing various data. However, when representing univariate data (i.e., a single percentage score) it seems that having a lower number of manipulated facial features is more beneficial. Therefore, it could be problematic to have several individual facial elements for the human to properly

decode. If a human naturally decodes a face as a whole rather than in parts; it may be counter-intuitive to present them with a face that requires the decoding of several features (parts) of the face. As Montello and Gray (2005) state, it is more beneficial to have a stimulus that communicates information univariately rather than multivariately when the goal is to give the user a single quantity. A pseudo-Chernoff face may be a remedy for this dilemma (Montello & Gray, 2005). This “pseudo-Chernoff” face could be created by systematically manipulating one facial characteristic, while holding all others constant. To properly convey a simple quantitative score the Chernoff face may only need to have one facial characteristic manipulated. Through this manipulation, the human may be more apt to decode the Chernoff face accurately and quickly, while noticing subtle changes (Kabulov, 1992).

The issue of whether interpreting Chernoff faces is a relatively less attention-demanding task is of primary importance to the current study. Previous studies have investigated the effectiveness of Chernoff faces as a pre-attentive stimulus with mixed results. A study concluded that Chernoff faces are not processed pre-attentively, and do not benefit users more than other modes of visual information display (Morris, Ebert, & Rheingans, 2000). The process of identifying the characteristics (eyebrow slant, eye size, nose length) of the Chernoff face was said to be a serial process. Participants’ accuracy of target stimuli identification improved when they were given more time and less distracters, indicating that the task was not pre-attentive (Morris, Ebert, & Rheingans, 2000). A similar study investigated data visualization and used Chernoff faces as one of the “glyph stimuli” to discover which data visualizations were the most effective (Lee,

Reilly, & Butavicius, 2003). Glyphs are data visualizations that are characterized by their attempt to display multivariate data through the manipulation of features on the glyph that correspond to raw data. It was found that participants had lower accuracy scores and took longer to answer questions when exposed to the glyph stimuli (Lee, Reilly, & Butavicius, 2003). This indicates a serial processing of information from the Chernoff faces, which is in agreement with the findings of Morris, Ebert, & Rheingans (2000).

A study investigating perceptual sensitivities found that children process Chernoff faces differently than adults (Tsurusawa, Goto, Mitsudome, Nakashima, & Tobimatsu, 2007). Children focus more on individual features, while adults process a face in a more holistic pattern. These findings seem to be discrepant with the previously mentioned studies. Perhaps adults do not decode Chernoff faces to the degree of serial processing as suggested by other studies. If adults decode in a faster more parallel manner, then Chernoff faces may allow for pre-attentive processing. Of particular interest is how the participants differed on their interpretation of the mouth angle presented. Children significantly differed from adults in their evaluation of the Chernoff face as a function of the angle of the stimuli's mouth. Children evaluated the faces as more emotional as the curvature of the mouth changed, while the adults were significantly below the children's evaluation score. Supposedly, this is a consequence of children's lack of holistic face processing ability (Tsurusawa, Goto, Mitsudome, Nakashima, & Tobimatsu, 2007). An additional finding bolstered Chernoff faces' potential value as a quantitative display. This was the participants' ability to evaluate the stimuli in discrete steps (Tsurusawa, Goto, Mitsudome, Nakashima, & Tobimatsu, 2007). Basically, participants could follow the

incremental facial feature changes in the Chernoff faces; similar to the hypothesis by Chernoff (1973). Although children and adults may process Chernoff faces differently, it can be inferred that Chernoff faces can demonstrate human facial expressions effectively.

A previous study used schematic faces (line faces similar to Chernoff faces) as stimuli to determine whether the “anger superiority effect” was apparent while using a visual search paradigm (Ohman, Lundqvist, & Esteves, 2001). The study found schematic faces to be identified quickly and accurately, with schematic faces representing anger/threatening emotion leading to the most pre-attentive reaction times. The visual search paradigm was reconfigured throughout the experiment by adding more distractor stimuli. This was done in an effort to make a more difficult visual search task, which would test for serial versus parallel search. Following each of these iterations, the threatening facial expression was shown to be the most decodable (faster and more accurate) stimuli (Ohman, Lundqvist, & Esteves, 2001). This is important because it indicates that the threatening schematic face is processed in parallel, or without using much attention. The results of this study show that schematic faces can be processed in parallel and that there is potentially an “anger superiority effect” for these types of stimuli (Ohman, Lundqvist, & Esteves, 2001).

If Chernoff faces are manipulated properly, giving the right amount of useful information, they will fulfill the heuristic of sufficient information design as an ambient display. To reiterate, the main issue concerning Chernoff faces is whether they can be interpreted pre-attentively, with minimal attentional resources. Once this issue is

understood with more clarity, the efficacy of facial expressions in the form of Chernoff faces to be ambient displays will be evident.

Age-Related and Cultural Effects on Decoding

Despite the ease with which humans are able to decode emotional facial expressions, it is still moderated by age. Age can alter a person's ability to correctly perceive and understand the facial expression that is presented to them. Neuropsychological research has shown that age-related issues in facial expression decoding may be a result of problems with the medial temporal lobe (Orgeta & Phillips, 2007). The amygdala is housed here, which corroborates with previous research that suggests the amygdala is necessary for facial expression decoding (Whalen, 1998; Morris, 1998). Despite these age-related issues; a competing theory has been asserted regarding older adult's ability to decode emotional facial expressions. The socioemotional selectivity theory asserts that social behavior is essentially a byproduct of time (Carstensen, Issacowitz, & Charles, 1999). In a sense, time can be thought of as the chronological age of a human. As the human ages, they essentially have less time to live and fulfill goals. This affects the way they view their decisions and weight their goals. The two types of goals that make up the socioemotional selectivity theory are knowledge-based and emotion-based goals (Carstensen, Issacowitz, & Charles, 1999). Younger adults are more likely to pursue knowledge-based goals because they have more time potential. The trade off for knowledge in lieu of emotional goals appears to be a worthy endeavor. Older adults supposedly take the opposite approach and view emotional-based goals as top priority. Older adults' view time as a non-renewable resource, and seek to

spend anytime they have left enjoying positive emotional experiences (Carstensen, Issacowitz, & Charles, 1999).

According to the socioemotional selectivity theory, older adults may actually be more aware of certain emotional situations and images than non-emotional (Orgeta & Phillips, 2007). Orgeta and Phillips (2007) showed older adults as being more accurate at identifying positive facial expressions, opposed to negative facial expressions. Older adults were found to identify positive emotions as accurately as younger adults. There was no significant difference between the older adults and younger adults in terms of identifying positive facial emotions (i.e., happiness and surprise). However, older adults were significantly worse than younger adults at identifying negative facial emotions (i.e., sadness, anger, and fear). The results of this study indicated that there is an age-related difference for the decoding of negative facial expressions, but not positive facial expressions (Orgeta & Phillips, 2007). The ease of recognition for certain emotional expressions is a phenomenon pertinent to this research area. As Orgeta and Phillips (2007) showed, older adults may have a positivity bias that allows them to overcome any cognitive decrements that interrupt other emotional decoding, thus decoding positive facial expressions as accurately as younger adults. Other research has supporting data showing that positive expressions (e.g., happiness) are processed more quickly, supported by faster N170 latencies (Batty & Taylor, 2003). Perhaps this quick processing attributes to the robustness of the happy facial expression compared to other expressions.

A previous study manipulated the factors of chronological age and the participant's working self-concept to determine if the positivity effect could in fact be

evoked in younger adults, and likewise the negativity effect in older adults (Lynchard & Radvansky, 2012). During the experiment the participant would complete a possible selves orienting task. The older adults completed the younger possible selves orienting task, while the younger adults completed the older possible selves orienting task. Essentially, this made the participant's working self-concept the opposite of their chronological age. The results showed a reversal of stereotypical age-related emotional information processing. Younger adults displayed a positivity effect, which is thought to be a unique attribute of older adults. Similarly, older adults displayed a negativity effect, which is thought to be unique to younger adults (Lynchard & Radvansky, 2012). This study showed that more than just chronological age plays a role in the socioemotional selectivity theory. Humans are subject to emotional information processing biases based on less concrete variables such as their working self-concept.

Decoding facial expressions is a cross-cultural behavior that is a critical part of human life. There are six basic emotions that transcend culture. These are: anger, happiness, fear, surprise, disgust, and sadness (Ekman & Friesen, 1975). These emotions can be represented with facial expressions (Lee, 2006; Batty, 2003). Because these facial expressions are not confined to specific cultures, it puts no restraints on the ability of different people groups to successfully decode these facial expressions. It appears that increasing age is a factor that may cause differences in aspects of facial expression decoding, while cultural background seems to be of no hindrance. The unique quality that facial expressions have in their prevalence and familiarity in human culture makes them a

good candidate for an ambient display. This quality of facial expressions allows the heuristic of matching the system to the real world to be met.

Limitations of Previous Literature

The previous literature has provided a foundation for knowledge about facial expressions, but there are limitations to these studies. The Hess (1997) study presented emotional facial expressions in a single-task format. The participants viewed the image and rated it on the emotionality and intensity that they perceived. This methodology does not clarify whether facial emotion decoding is truly resource/attention-free as neuropsychological studies suggest. A dual-task experiment should be implemented to properly measure attention usage. In order to gain this data; measures of response time, accuracy, and subjective workload should be used. The Hess (1997) study also measured decoding accuracy for each facial expression image through the presentation of several emotion scales at once. The participant was presented with seven emotional labels, which they manipulated to show the intensity of emotion for the previous picture. Instead of presenting seven individual scales, it seems to be less complicated to present one scale or to have a quick input device (e.g., keyboard number keys) after the image is viewed.

The Hess (1997) study presented facial expression intensity in increments of 20 % intensity. This intensity scale may not provide enough precision or a complete spectrum of facial expression decoding data. The Orgeta and Phillips (2007) study also presented only four intensity levels. The number of intensity levels may need to be increased (i.e., create smaller increments of percentage changes between each stimuli) to capture a more accurate representation of participants' ability to decode facial expression. Another

limitation in the Orgeta and Phillips (2007) study was the facial images were presented in increasing order as the participant advanced through the experiment. This method may have led to participants forming an anticipation bias that the next facial image was going to be more expressive.

Previous research has also provided evidence that age-related effects may cause differences in the ability for humans to properly decode facial expressions. It has been shown that older adults are worse at identifying negative facial expressions (i.e., sadness, anger, and fear). Older adults struggled significantly versus younger adults in properly recognizing the negative emotions at intensity levels of 50 %, 75 %, and 100 %. It appears that older adults have a higher recognition threshold for certain negative emotions than younger adults. Basically, older adults do not pick up on negative facial stimuli as easily as younger adults and need more intense facial expressions to determine the appropriate emotional state (Orgeta & Phillips, 2007). In order to determine if theories such as the socioemotional selectivity theory pertain to Chernoff face recognition, there needs to be an independent variable of age with levels of younger and older adults.

The variable of gender of the facial expression stimuli could be considered a confounding variable. Hess (1997) used two male and two female actors to create facial expressions for their study. Results of this study showed that the gender of the stimuli (i.e., actors) did influence participant rating accuracy. For the expressions of happy and sad, there was an interaction of the gender of the stimuli x intensity of the expression

(Hess, 1997). Because of this reported interaction, it would be beneficial to use non-gender specific stimuli to eliminate this confounding variable.

Previous studies have looked at users' ability to properly decode facial expression type (Ekman & Friesen, 1975), intensity (Tsurusawa, Goto, Mitsudome, Nakashima, & Tobimatsu, 2007; Hess 1997), and the effectiveness of Chernoff faces (Chernoff 1973; Tsurusawa, Goto, Mitsudome, Nakashima, & Tobimatsu, 2007; Morris, Ebert, & Rheingans, 2000). The purpose of the current study is to examine the users' ability to accurately decode a quantitative value from Chernoff facial expressions.

Overview of the Current Study

In order to determine the attention usage by the participants, a dual-task methodology was used. Our study used the dual-task paradigm to measure the attention-demanding characteristics of facial displays. The Hess (1997) study measured participant's decoding accuracy with several scales after each trial. This method may create confusion for the participant, and not accurately record participant decoding time. The interface should allow for quick and simple input of the facial expression intensity from the participant. The current study used only one measurement scale (direct key entry) after each trial to eliminate any confusion for the participants about what the scales are measuring and give a better approximation about how quickly the participant can decode the facial expression. In the Orgeta and Phillips (2007) study the facial expressions were shown in increasing order. This technique was not replicated in the current study. Instead, a randomized sequence of facial expression stimuli was used to control for any biases that could be formed due to participant expectations. The Chernoff

face stimuli were manipulated differently compared to previous research (Chernoff, 1973; Tsurusawa, Goto, Mitsudome, Nakashima, & Tobimatsu, 2007; Morris, Ebert, & Rheingans, 2000). Only the mouth was manipulated in order to gain understanding about the affect of this one variable on decoding. Finally, the current study used a more precise facial expression intensity scale than previous research (Hess, 1997; Orgeta & Phillips, 2007). To accomplish this, a facial expression scale presenting emotions in increments of 10 % was used. Our assumption was that by making these modifications the current study would be able to address the research question with more accuracy.

Hypotheses of the Current Study

The first hypothesis (H_1) was that there would be no age differences in facial decoding performance in the happy facial expression condition, but that there would be decoding performance differences in the sad facial expression condition. The rationale behind expecting no age difference in the happy facial expression condition is based on the socioemotional selectivity theory and research that supports positive expressions as more identifiable; referred to as the “happy face advantage” (Ekman & Friesen, 1975; Orgeta & Phillips, 2007; Calvo & Lundqvist, 2008). The rationale for the age-related difference in the sad facial expression condition is based on older adults’ difficulty in perceiving sad facial expressions (Orgeta & Phillips, 2007), and the negativity effect seen in younger adults (Lynchard & Radvansky, 2012).

The second hypothesis (H_2) was related to the rationale of hypothesis H_1 (i.e., effect of the happy face advantage), namely that even in the presence of another task, there would be no age differences in happy facial expression decoding because of its

presumed pre-attentiveness. However, we assumed that sad facial expression decoding would require attentional capacity, and thus be affected by the presence of a dual-task. If the decoding of happy facial expression is actually resource-free (Lee, 2006; Whalen, 1998; Morris, 1998), then facial decoding in the dual-task phase should be equivalent to decoding in the single-task condition. There will be similar performance scores for younger and older adults in the happy condition; regardless of phase (single or dual). This indicates that the happy facial expressions are able to mitigate the dual-task decrement that would be expected for stimuli that demand more attention, which we expect to be the sad facial expressions. Older adults' performance with sad facial expressions is expected to be worse (compared to their single-task baseline), due to their low negative emotional sensitivity (positivity bias) and the added cognitive load of the dual-task. We also expect younger adults' performance to decrease due to the additional cognitive load of the dual-task condition, which we expect will degrade any benefit of the negativity bias. Additionally, research has shown younger adults to be more quick and accurate at decoding happy expressions versus sad facial expressions (Hess, 1997; Calvo & Lundqvist, 2008).

METHODS

Participants

Eighty-three participants (42 younger adults, 41 older adults) were recruited for the current study. The younger adult age range was 18 – 21 ($M = 18.6$, $SD = .89$) and the older adult age range was 65 – 84, ($M = 72.4$, $SD = 5.19$). Younger adults were recruited from psychology courses and received class credit for participation. Older adults were

recruited from a pre-existing database of volunteers who lived in the surrounding communities. Older adults received \$25 for participation.

Design

This study was a 2 (age group: younger, older) x 2 (facial expression condition: happy, sad) x 10 (facial expression intensity: 0%-90%) x 2 (task phase: single, dual) mixed-design. Age group was a quasi-independent grouping variable. Facial expression condition was between-groups, while facial expression intensity and task phase were within-groups. The dependent variables measured were: the speed (ms) for the block task, the speed (ms) of response on the facial expression task, the amount of “misses” on the facial expression task, the amount of blocks cleared, facial expression intensity rating, and decoding accuracy (i.e., slope value) of the correspondence between the face presented and the facial expression intensity rating.

Materials

The experiment was presented on 19-inch LCD monitors and participants made responses using the keyboard. Participants were seated in office chairs about 18-24 inches from the screen in a laboratory environment. The experiment was programmed using Real Basic.

Surveys & Abilities

Participants completed a computerized cognitive abilities battery. These tests gathered information on participants’ working memory, perceptual speed, and vocabulary. Participants also completed a computerized version of the NASA-TLX survey to measure subjective workload.

Tasks

The block task was a game similar to the game Tetris (Appendix A). The block task consisted of moving multi-colored blocks. The main objective of the block task was to “clear” block rows or columns by manipulating the blocks using the arrow keys and space bar. To successfully “clear” a block row or column, the participant was required to align three blocks of the same color. This task was used in the dual-task as the primary task due to its supposed high attentional demand.

The purpose of the facial expression decoding task was to identify the level of emotion presented by a computer-generated facial expression (Appendix B). The facial expression stimuli were rendered using the statistical program R. This allowed the experimenter to have control over the faces and manipulate their facial expression intensity as desired. The facial expression stimuli were line drawings composed of black lines on a white background. This eliminated any confounding variables due to the gender, ethnicity, or age of the stimuli. There were 19 images: 9 happy stimuli (ranging from 10% expressive – 90% expressive), 9 sad stimuli (ranging from 10% expressive – 90 % expressive), and one neutral stimulus (0 % expressive), see Appendix C. The range of expressiveness was chosen from 0%-90% in an effort to make a match between the key number pad and the expression levels. The images were 170 pixels by 250 pixels.

Procedure

Participants were randomly assigned to experimental conditions (happy or sad) prior to the experiment. The participants were given an informational letter before the experiment began. The experiment consisted of three phases. The participants completed

two subsequent single-tasks (i.e., the block task and facial expression decoding task) to record baseline data on their abilities, and to become familiar with each task. To examine the attentional demands of decoding Chernoff faces, participants then engaged in the dual-task phase. Participants were instructed to focus on the block task (i.e., primary task) and consider it to be the most important task. This spatial-manipulation task was chosen due to the expectation of being cognitively taxing for the participants. Participants were told to try to complete the facial expression decoding task (i.e., secondary task) effectively, but not to sacrifice their primary task performance during the dual-task phase.

In phase 1, participants performed the block task in a single-task environment. The participant had to reach a pre-set score (based on number of blocks cleared) to complete the task. Once the participant completed this phase, the program proceeded to phase 2. In phase 2 of the experiment, participants were asked to respond to Chernoff facial expressions that were flashed on the computer screen. The participants were in one of two facial expression conditions (i.e., happy or sad) and only saw faces related to their facial expression condition.

Once phase 2 began, the Chernoff facial expression appeared in a window on the computer screen. The facial expressions were shown in a randomized order in regard to their intensity level. During the time interval that the facial expression was present, participants attempted to respond to the facial expression using the number keys. If the participant did not hit a number key before this time elapsed then a “miss” was recorded. Regardless of whether the participants had responded or missed making a response, after three to five seconds (randomized facial expression appearance time) the screen went

back to being blank until the next trial. There were 60 trials in each condition (i.e., 6 exposures to each of the stimuli for a specific condition). After the participants were exposed to all 60 stimuli the program proceeded to phase 3.

In phase 3, participants were exposed to both phases 1 and 2 simultaneously (see Appendix D). This created a dual-task situation. The task goals defined for the two single-tasks remained the same for the dual-task phase. However, participants were told to treat the block task as the primary task. This phase continued until all facial expression stimuli were presented to the participants. After the participants completed the experiment, the computer loaded the computerized NASA-TLX survey. Subsequently, the battery of computerized cognitive abilities tests was loaded for the participants to complete. Once the participants completed the cognitive abilities battery they were finished with the study and permitted to leave.

RESULTS

Participants' data were removed based on two criteria: 1) if they missed all the faces presented in phase 3 (i.e., indicating little attention paid to the secondary task), or 2) if they were 2 standard deviations below the group average for clearing blocks in phase 3 (which indicated little attention being paid to the primary task). Participants' who had marginally low performance (on either of the aforementioned criteria); subsequently had their cognitive abilities test results examined. If the participant had a cognitive ability test score 2 standard deviations below the group average (on any of the three ability tests), then their data were removed from the final analysis. This criteria resulted in the removal of nine participants: six participants due to missing all the faces presented in phase 3, one

participant who scored 2 standard deviations below the group average for clearing blocks, one participant who missed most of the faces presented in phase 3 (55 out of 60) and scored 2 standard deviations below the group average on two cognitive ability tests, and one participant was removed because they participated in the pilot testing for the current study.

The following results section is organized by task phase (i.e., single or dual). To remind the reader, phase 2 was the single-task for facial expression decoding and phase 3 was the dual-task condition. The results of the single-task facial expression decoding condition (phase 2) inform hypothesis H₁, while the dual-task facial expression decoding condition (phase 3) results are directly relevant to hypothesis H₂. In the single-task facial expression decoding condition (phase 2), the following dependent variables were analyzed: intensity key pressed, facial expression decoding accuracy, facial expression response time (ms), and the amount of face misses for the facial expression task. In the dual-task portion (phase 3), the following dependent variables were analyzed: intensity key pressed, facial expression decoding accuracy, facial expression response time (ms), the amount of face misses for the facial expression task, and computed workload from the NASA-TLX survey. An alpha level of .05 was used for all of the following statistical tests. Tests for the assumption of normality (i.e., histogram, Q-Q plot) and homoscedasticity were conducted and showed the data met the assumption for normality and homoscedasticity. For all mixed measures ANOVAs, the number of levels of the repeated measures IV (i.e., single task phase, dual task phase) was less than three, so sphericity was assumed.

Phase 2 (Single-task, Facial Expression Decoding Only)

Intensity Key Pressed

As participants were presented faces during phase 2, they were asked to give intensity ratings about each face. In order to give these intensity ratings, participants' used the keyboard number keys as the input device. The intensity key pressed ratings for a participant were averaged across all trials for phase 2. This yielded a mean intensity key pressed value that could be analyzed as a function of facial expression condition, age group, and face presented. The intensity key pressed ratings were also necessary for the calculation of decoding accuracy, which will now be explained.

Decoding Accuracy

In the facial expression decoding task, participants were asked to view facial expressions that were flashed on the computer screen (heretofore called "face presented") and to respond with an intensity rating ("intensity key pressed"). The facial expressions presented ranged from 0 (neutral) to 9 (very expressive). Decoding accuracy was operationalized as the correspondence between the face presented and participants' intensity key pressed. The regression slope of participants' correspondence was used to quantify decoding accuracy.

A hierarchical regression analysis was conducted to predict intensity key pressed as a function of age group, facial expression condition, and face presented. The predictor variables of age group and facial expression condition were dummy-coded. The predictor variables were entered in three steps, which resulted in three different models. The first step contained the following predictor variables: face presented, facial expression

condition, and age group. These predictor variables represented all of the main effects tested (model 1). The second step contained the predictor variables from model 1 with the addition of the following two-way interactions: age group x facial expression condition, face presented x age group, and face presented x facial expression condition (model 2). The third step contained all of the predictor variables from model 1 and model 2 with the addition of the following three-way interaction: face presented x age group x facial expression condition (model 3).

The three models were tested for their ability to significantly predict participants' intensity key pressed. Model 1 accounted for 44.4 % of the variance of intensity key pressed, ($R^2 = .444$, $F(3, 826) = 220.11$, $p < .001$). Model 2 accounted for 51 % of the variance of intensity key pressed, ($R^2 = .510$, $F(6, 823) = 142.62$, $p < .001$). Model 3 accounted for 51.1 % of the variance of intensity key pressed, ($R^2 = .511$, $F(7, 822) = 122.66$, $p < .001$). The addition of the two-way interactions in model 2 resulted in a R^2 change value of .065, or 6.5 %, while the addition of the three-way interaction in model 3 resulted in a R^2 change value of .001, or 0.1 %. The addition of the three-way interaction (via model 3) did not add a significant amount of predictive power to the model.

The non-significance of the hypothesized three-way interaction of face presented x age group x facial expression condition ($b = -.11$, $t(822) = -1.39$, $p = .165$), caused slope comparisons to be confined to the two-way interactions in model 2. The two-way interaction terms in the hierarchical regression were a method to test for a significant difference between the regression line slopes. Therefore, when a two-way interaction was found to be significant, it was showing the two regression slopes to be significantly

different. First, main effects and interactions for intensity key pressed will be addressed, followed by interactions related to decoding accuracy.

Main Effects and Interactions for Intensity Key Pressed

There was a significant main effect of face presented on participants' intensity key pressed, ($b = .53$, $t(826) = 25.27$, $p < .001$), which meant participants were generally able to discriminate the various levels of face presented. As the actual face presented stimuli increased from 0 % to 90 %, there was a .53 unit increase for intensity key pressed by the participants. There was a significant main effect of facial expression condition, ($b = .57$, $t(826) = 4.67$, $p < .001$). This main effect revealed a significant increase in mean intensity key pressed between the sad facial expression condition ($M = 4.49$, $SD = 2.15$) and the happy facial expression condition ($M = 5.06$, $SD = 2.47$). There was no main effect of age group, ($b = .01$, $t(826) = .09$, $p = .928$).

The two-way interaction of age group x facial expression condition was significant, ($b = -.64$, $t(823) = -2.82$, $p < .01$). Due to the dichotomous nature of the predictor variables (happy, sad; younger, older), the lines only contain two data points (i.e., mean values of intensity key pressed). The interaction can be conceptualized as the difference between the differences in mean values of intensity key pressed for each age group. The difference between the means (i.e., slope), for younger adults was .88, which is significantly different than the difference between the means, .25, for older adults.

Slopes were found using the following formula: $b = \frac{Y_2 - Y_1}{X_2 - X_1}$, where the mean values were used for Y and facial expression condition coding (0 = Sad, 1 = Happy) was used for X. As Figure 1 illustrates, the two-way interaction was a result of the significantly greater

increase in mean intensity key pressed in the younger adult group as a function of facial expression condition compared to older adults.

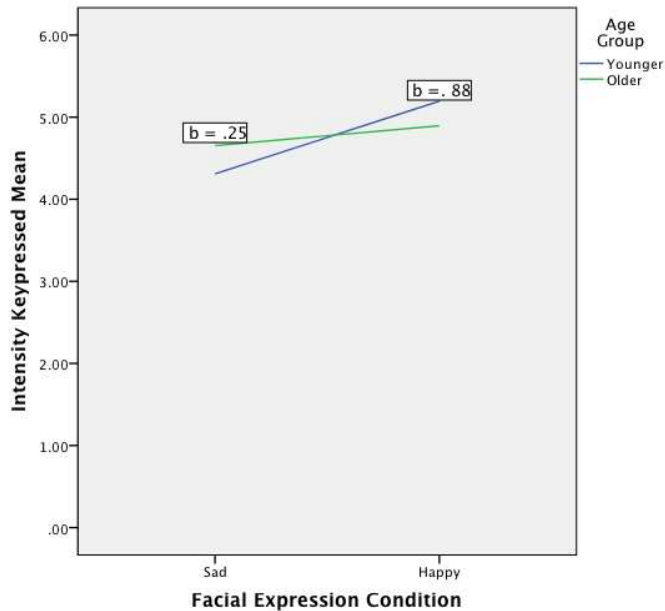


Figure 1. Mean intensity key pressed by facial expression condition for younger and older adults.

Interactions for Decoding Accuracy

The two-way interaction of face presented x age group was significant, ($b = -.18$, $t(823) = -4.46$, $p < .001$). This indicated that in general, younger adults were significantly better than older adults at accurately decoding the faces presented. Participants' facial expression decoding values were compared between the younger age group and the older age group, resulting in an observed significant decrease in slope (i.e., a younger adult slope of $b = .63$ versus an older adult slope of $b = .43$), illustrated by Figure 2.

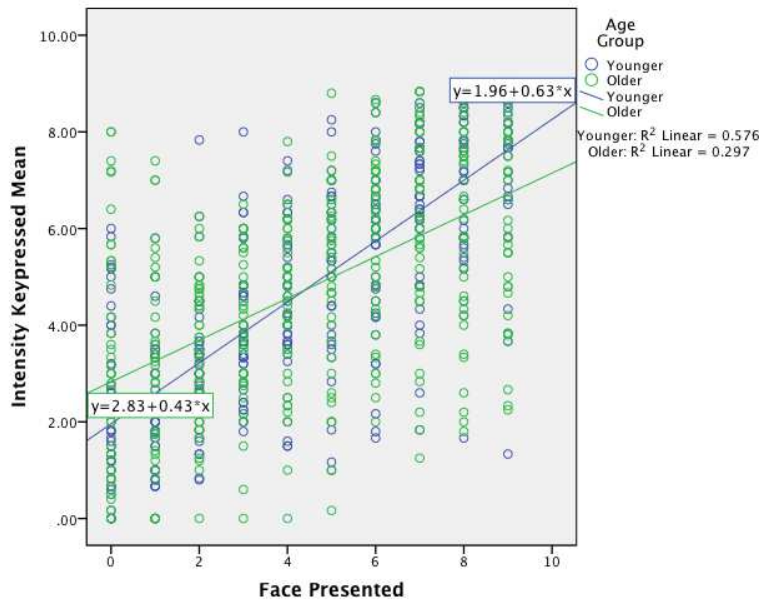


Figure 2. Mean intensity key pressed by face presented for younger and older adults.

The two-way interaction of face presented x facial expression condition was significant, ($b = .35$, $t(823) = 8.78$, $p < .001$). This indicated that all participants were generally more accurate at decoding the happy facial expression condition than the sad facial expression condition. This two-way interaction is illustrated by Figure 3. Participants' (collapsing across age group) facial expression decoding values were compared between the sad facial expression condition and happy facial expression condition, yielding a significant difference in slopes (i.e., a sad slope of $b = .35$ versus a happy slope of $b = .71$).

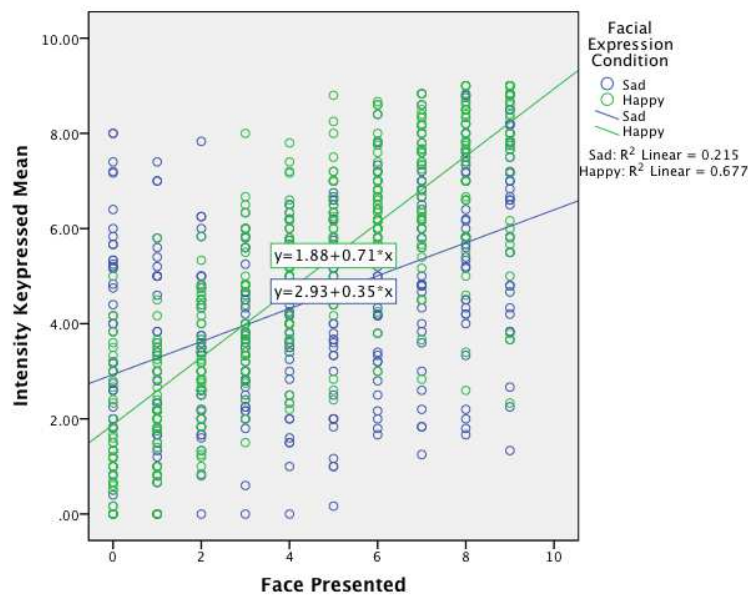


Figure 3. Mean intensity key pressed by face presented for sad and happy facial expression conditions.

The three-way interaction for face presented x age group x facial expression condition was not significant ($b = -.11$, $t(822) = -1.39$, $p = .17$). This means that facial expression decoding accuracy did not differ as a function of age group and facial expression condition. This does not support hypothesis H₁, which predicted no age differences in decoding accuracy in the happy facial expression condition, while predicting an age difference in the sad facial expression condition.

Intensity Key Pressed Response Time

The speed at which participants made responses could be interpreted as the level of attentional demand required of the stimuli. The purpose of measuring intensity key pressed response time was to examine whether attentional demand changed as a function of facial expression condition, age group, or an interaction of facial expression condition x age group. The response time for a participant was operationalized as the time in

milliseconds (ms) it took a participant to depress a number key when presented with a facial expression. The facial expression would appear randomly throughout phase 2 (every 3-5 seconds) to avoid a predictable appearance interval. However, the face appeared or was shown for the same amount of time for every trial (2 seconds for younger adults, 2.5 seconds for older adults). Response time data was discussed in terms of seconds for ease of understanding.

A 2 (age group) x 2 (facial expression condition) ANOVA was conducted to analyze participants' response time data. A significant main effect was found for age group ($F(1, 81) = 317.80, p < .001$). Younger adults' response time ($M = 1.27$ s, $SD = .11$ s) was significantly faster than older adults' response time ($M = 1.9$ s, $SD = .20$ s). There was no main effect for facial expression condition ($F(1, 81) = .342, p = .56$), and no significant interaction for age group x facial expression condition ($F(1, 81) = .03, p = .86$). Regardless of facial expression condition, younger adults had significantly faster response times than older adults; illustrated by Figure 4.

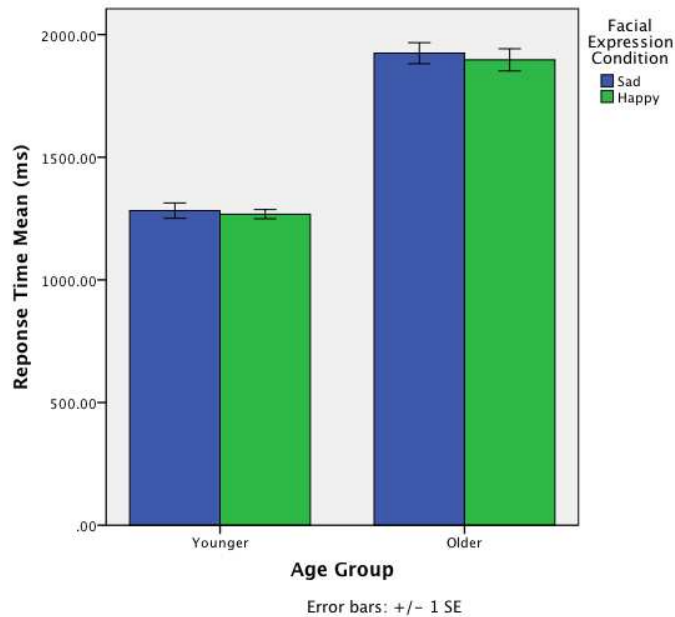


Figure 4. Mean response time (ms) by age group for sad and happy facial expression conditions.

Face Misses

The extent that participants “missed” identifying faces in the allotted time could be used to understand the attention demanding characteristics of the faces. We anticipated pre-attentive faces to be less “missed” compared to faces that required more attention. Face misses were operationalized as situations where the participant did not respond, or failed to press the number key (i.e., intensity key pressed) within the allotted time interval. When participants “missed” a facial expression it was recorded, and misses were summed and averaged for participants’ experimental session.

A 2 (age group) x 2 (facial expression condition) ANOVA was conducted to analyze participants’ amount of misses. A significant main effect was found for facial expression condition ($F(1, 81) = 5.9, p = .02$). Participants in the sad facial expression condition had significantly more misses ($M = 8.53, SD = 5.48$) than participants in the happy facial expression condition ($M = 6.05, SD = 3.6$). There was no main effect of age

group ($F(1, 81) = 2.68, p = .11$), and no interaction for age group x facial expression condition ($F(1, 81) = 3.66, p = .06$). Figure 5 highlights the main effect of facial expression condition and the marginally significant interaction between age group x facial expression condition.

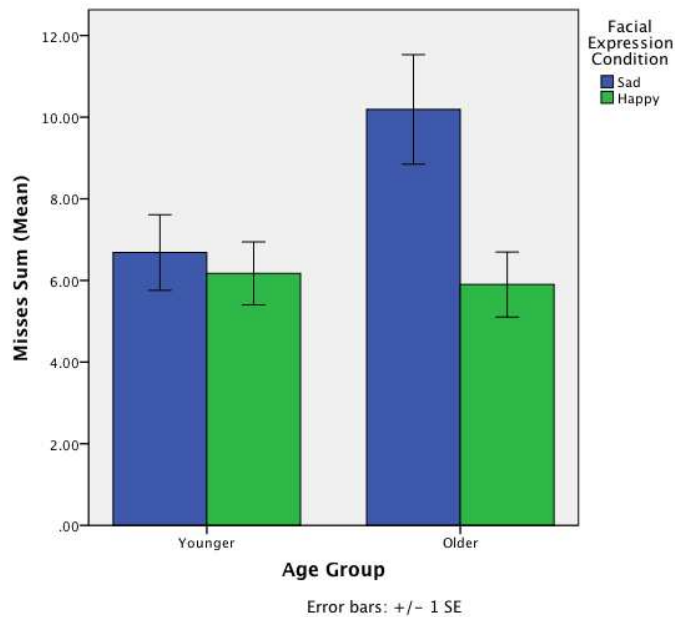


Figure 5. Mean number of face misses by age group for sad and happy facial expression conditions.

In sum, the results of the analysis of task phase 2 show that the variables of face presented, facial expression condition, and age group had a significant effect on participants' performance. The significant main effect of face presented on participants' intensity key pressed showed a positive linear trend for intensity key pressed as the variable of face presented increased. The significant main effect of facial expression condition on intensity key pressed revealed a significant increase in mean intensity key pressed when comparing between the sad facial expression condition and the happy facial expression condition. The significant main effect of age group on response time showed

younger adults' response time was significantly faster than older adults' response time. The significant main effect of facial expression condition on face misses showed participants in the sad facial expression condition had significantly more misses than participants in the happy facial expression condition. The significant two-way interaction of age group x facial expression condition showed a significantly higher intensity key pressed for younger adults compared to older adults, when comparing between the sad and happy facial expression condition. The significant two-way interaction of face presented x facial expression condition showed participants in the happy facial expression condition had significantly higher decoding accuracy than those in the sad facial expression condition. However, the lack of a three-way interaction suggested that the happy face advantage for decoding was not significant for older adults. The significant two-way interaction of face presented x age group showed younger adults had a significantly higher decoding accuracy than older adults.

Examination of the aforementioned data was from task phase 2 (single-task phase) where presumably, all attention was devoted to the facial expression decoding task. To examine the attentional demands of facial decoding, performance in the facial expression decoding task was examined in the context of a dual-task environment (phase 3).

Task Phase 3 (Dual-task, Block Task and Facial Expression Decoding)

In task phase 3, participants were given a primary task (block game) and a secondary task (facial expression decoding). This dual-task paradigm allowed participant performance data from phase 2 to be compared to phase 3 (i.e., attention divided

situation). The purpose of the following analyses was to determine the extent to which facial expression decoding was disrupted (i.e., dual-task cost) by the block task.

In phase 3, intensity key pressed and decoding accuracy were operationalized as described in phase 2. However, the new independent variable of task phase provided a method to compare performance variables as a function of single or dual-task.

A hierarchical regression analysis was conducted to predict intensity key pressed as a function of age group, facial expression condition, face presented, and task phase. The predictor variables of age group, facial expression condition, and task phase were dummy-coded. The predictor variables were entered in four steps, which resulted in four different models. The first step contained the following predictor variables: face presented, facial expression condition, age group, and task phase. These predictor variables represented all of the main effects tested (model 1). The second step contained the predictor variables from model 1 with the addition of the following two-way interactions: age group x facial expression condition, face presented x age group, face presented x facial expression condition, face presented x task phase, task phase x age group, and task phase x facial expression condition (model 2). The third step contained all of the predictor variables from model 1 and model 2 with the addition of the following three-way interactions: face presented x age group x facial expression condition, task phase x age group x facial expression condition, face presented x task phase x age group, and face presented x task phase x facial expression condition (model 3). The fourth step contained all of the predictor variables from model 1, model 2, and model 3, with the

addition of the following four-way interaction: face presented x task phase x facial expression condition x age group (model 4).

The models were tested for their ability to significantly predict participants' intensity key pressed. Model 1 accounted for 43.6 % of the variance of intensity key pressed, ($R^2 = .436$, $F(4, 1552) = 299.92$, $p < .001$). Model 2 accounted for 49.3 % of the variance of intensity key pressed, ($R^2 = .493$, $F(10, 1546) = 150.34$, $p < .001$). Model 3 accounted for 49.6 % of the variance of intensity key pressed, ($R^2 = .496$, $F(14, 1542) = 108.33$, $p < .001$). Model 4 accounted for 49.6 % of the variance of intensity key pressed, ($R^2 = .496$, $F(15, 1541) = 101.21$, $p < .001$). The addition of the two-way interactions in model 2 resulted in an R^2 change value of .057, or 5.7 %, while the addition of the three-way interaction in model 3 resulted in a R^2 change value of .003, or 0.3 %. The addition of the four-way interaction resulted in no significant R^2 change compared to model 3.

As expected, (due to the low R^2 change value from model 2 to model 3), the hierarchical regression showed non-significant values for all of the task phase related three-way interactions: task phase x age group x facial expression condition ($b = .08$, $t(1542) = .21$, $p = .83$), face presented x task phase x age group ($b = -.02$, $t(1542) = -.35$, $p = .72$), and face presented x task phase x facial expression condition ($b = -.05$, $t(1542) = -.85$, $p = .40$). This meant no two-way interactions significantly changed across the predictor variable of task phase (e.g., face presented x facial expression condition did not change due to task phase). It was determined that model 4 did not yield a significant four-way interaction, ($b = -.14$, $t(1541) = -1.1$, $p = .269$). Due to the non-significant results of the three-way and four-way interaction terms, the following analyses concentrate on

model 1 and model 2. Slope comparisons will be confined to only two-way interactions related to model 2. The analyses of model 1 and model 2 give a simplified overview (i.e., less complex interactions) of the effect of task phase on participant performance.

Main Effects and Interactions for Intensity Key Pressed

There was no main effect of task phase on participants' intensity key pressed, ($b = .09$, $t(1552) = .927$, $p = .354$). As participants' moved from single to dual-task there was no significant difference for intensity key pressed values. The non-significant main effect of task phase can be thought of as a manipulation check, indicating that participants did not give the facial expression stimuli significantly different mean intensity ratings in the single-task phase versus the dual-task phase.

There was no significant two-way interaction for facial expression condition x task phase, ($b = .18$, $t(1546) = .99$, $p = .32$). Facial expression condition did not have a significant effect on the difference between the differences of means (i.e., slope) for intensity key pressed, when comparing across task phase.

A significant two-way interaction was found for age group x task phase, ($b = .39$, $t(1546) = 2.17$, $p = .03$), illustrated by Figure 6. Task phase had a significant effect on the difference between the differences of means (i.e., slope) for intensity key pressed, when comparing across age group. Slopes were found using the following formula: $b = \frac{Y_2 - Y_1}{X_2 - X_1}$, where the mean intensity key pressed values were used for Y and age group coding (0 = Single, 1 = Dual) was used for X. The slope for younger adults ($b = -.05$) was significantly different from the slope for older adults ($b = .27$). The change in mean

intensity key pressed, as a function of task phase for older adults, was significantly greater than younger adults.

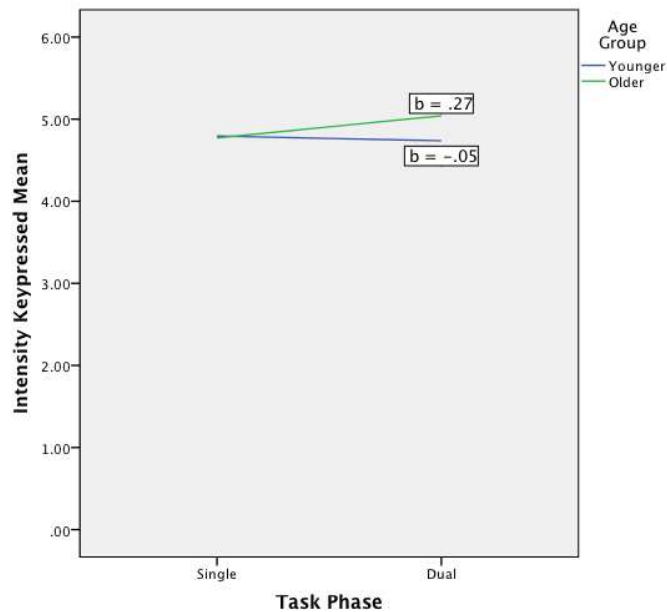


Figure 6. Mean intensity key pressed by task phase for younger and older adults.

Interactions for Decoding Accuracy

There was no significant two-way interaction of face presented x task phase, ($b = .04$, $t(1546) = 1.17$, $p = .24$). Participants' decoding accuracy (when collapsing across age group and facial expression condition) was not significantly affected by the task phase of the experiment. The slope values for each task phase did not significantly differ.

No significant three-way interactions were observed as a function of task phase. The three-way interaction of task phase x age group x facial expression condition was not significant ($b = .08$, $t(1542) = .21$, $p = .83$), the three-way interaction of task phase x face presented x age group was not significant ($b = -.02$, $t(1542) = -.35$, $p = .72$), and the

three-way interaction of task phase x face presented x facial expression condition was not significant ($b = -.05$, $t(1542) = -.85$, $p = .40$). The non-significance of these three-way interactions indicated that no two-way interactions significantly differed across task phase. The significant two-way interaction of face presented x age group shown in the single-task phase, remained significant ($b = -.20$, $t(720) = -4.14$, $p < .001$) in the dual-task phase, illustrated by Figure 7. This meant the significant interaction between face presented x age group (i.e., younger adults had significantly higher decoding accuracy than older adults) in the single-task, was replicated in the dual-task. The two-way interaction of face presented x facial expression condition shown in the single-task phase, remained significant ($b = .30$, $t(720) = 6.13$, $p < .001$) in the dual-task phase, illustrated by Figure 8. This meant the significant interaction between face presented x facial expression condition (i.e., happy condition was significantly higher for decoding accuracy than sad condition) in the single-task was replicated in the dual-task. Essentially, this showed there was no dual-task cost for these two-way interactions.

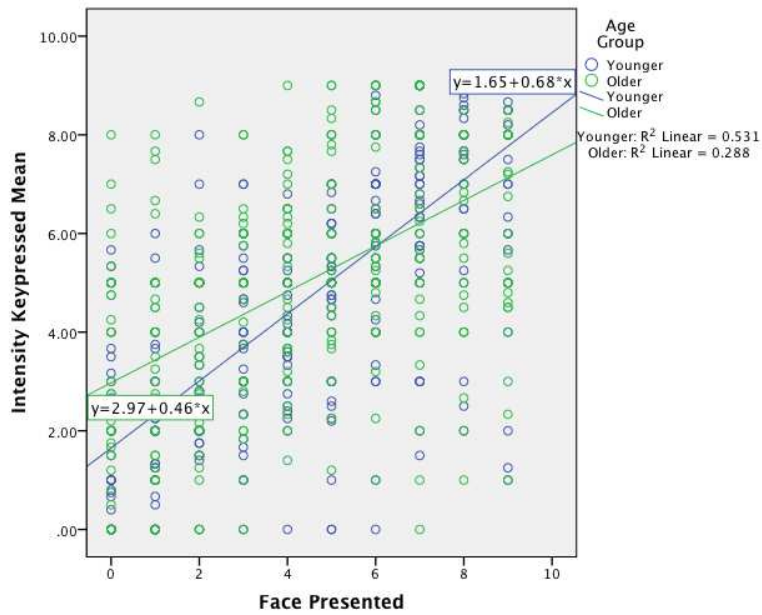


Figure 7. Mean intensity key pressed by face presented for younger and older adults (dual-task).

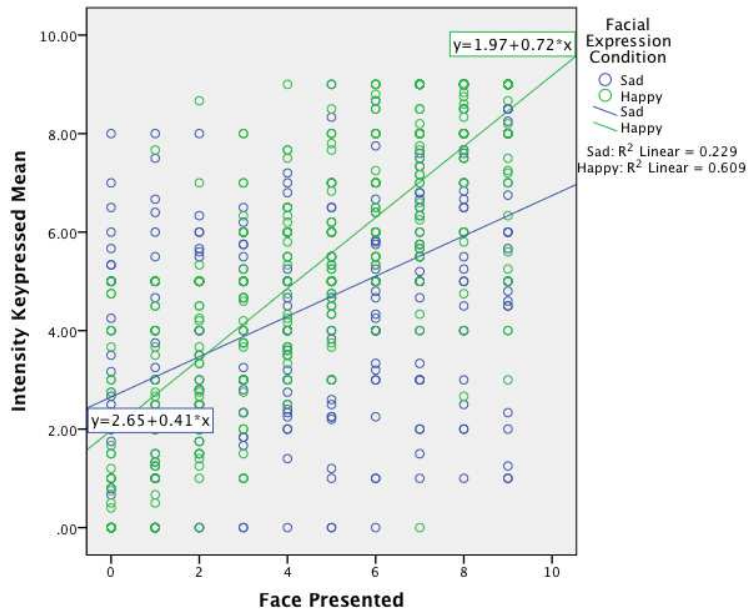


Figure 8. Mean intensity key pressed by face presented for sad and happy facial expression condition (dual-task).

The four-way interaction of face presented x task phase x facial expression condition x age group was not significant, ($b = -.14$, $t(1541) = -1.11$, $p = .27$). This finding showed that no three-way interactions significantly differed across task phase. This showed a lack of dual-task cost for the interaction of face presented x facial expression condition x age group. In the single-task happy facial expression condition, the significant two-way interaction for face presented x age group ($b = -.23$, $t(426) = -5.03$, $p < .001$) remained significant in the dual-task happy facial expression condition, ($b = -.32$, $t(384) = -5.58$, $p < .001$), illustrated by Figures 9 and 10. This meant the significant interaction between face presented x age group (i.e., younger adults had significantly higher decoding accuracy than older adults) in the single-task happy facial expression condition, was replicated in the dual-task happy facial expression condition.

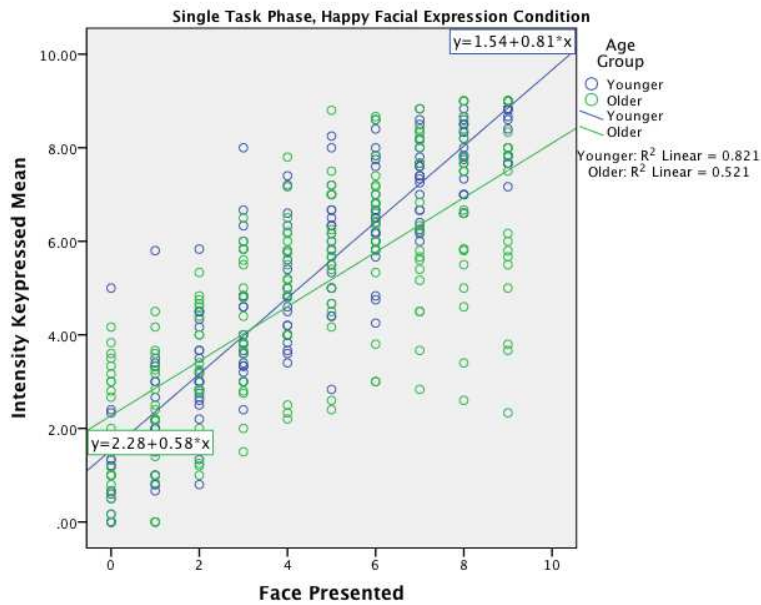


Figure 9. Mean intensity key pressed by face presented for younger and older adults (single-task, happy facial expression condition).

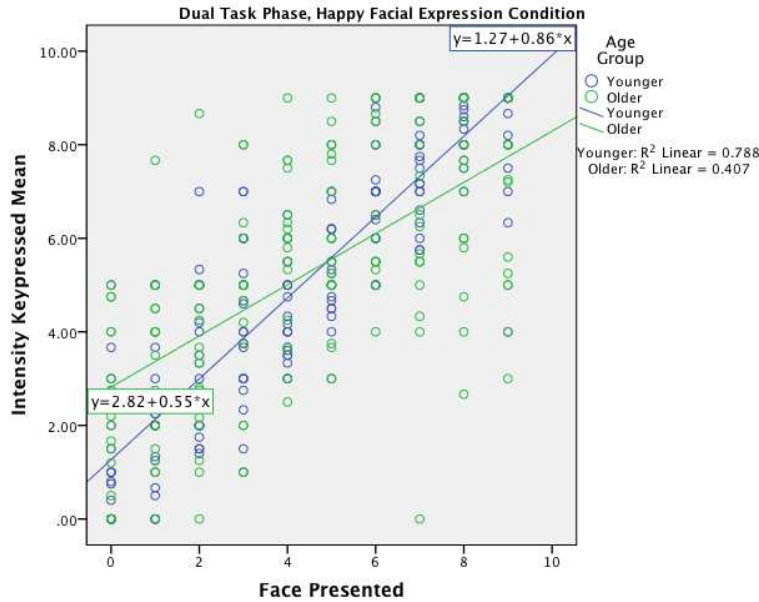


Figure 10. Mean intensity key pressed by face presented for younger and older adults (dual-task, happy facial expression condition).

In the single-task sad facial expression condition, the non-significant two-way interaction for face presented x age group ($b = -.12$, $t(396) = -1.82$, $p = .07$) remained non-significant in the dual-task happy facial expression condition ($b = -.07$, $t(335) = -.86$, $p = .39$), illustrated by Figures 11 and 12. This meant the non-significant interaction between face presented x age group (i.e., younger adults had similar decoding accuracy as older adults) in the single-task sad facial expression condition, was replicated in the dual-task sad facial expression condition.

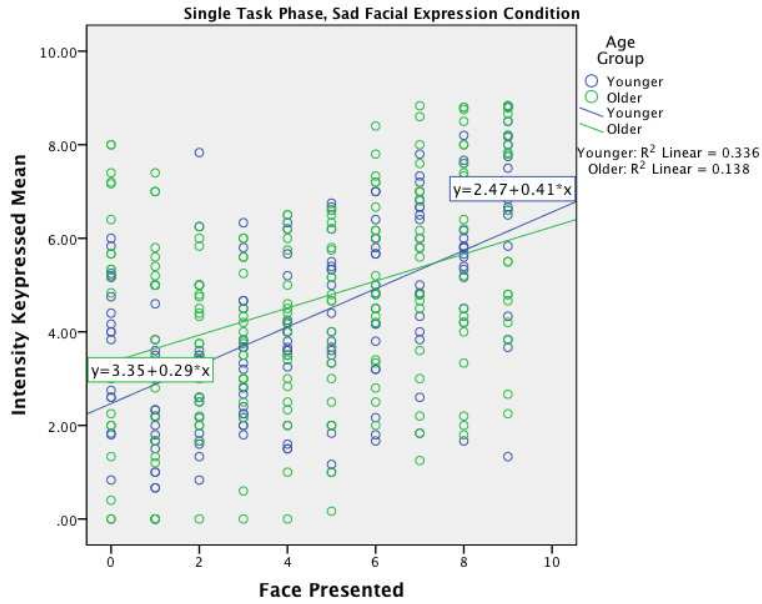


Figure 11. Mean intensity key pressed by face presented for younger and older adults (single-task, sad facial expression condition).

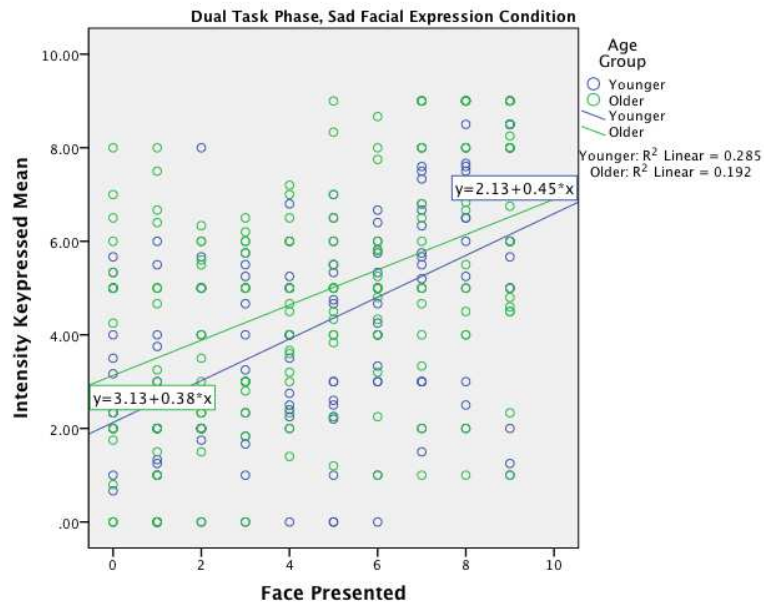


Figure 12. Mean intensity key pressed by face presented for younger and older adults (dual-task, sad facial expression condition).

Intensity Key Pressed Response Time

A mixed measures ANOVA was conducted on the response time data for facial expression decoding. There was a significant main effect of task phase on response time ($F(1, 79) = 34.34, p < .001$), illustrated by Figure 13. Response time for task phase 2 ($M = 1.59$ s, $SD = .36$ s) was significantly faster than reaction time for task phase 3 ($M = 1.72$ s, $SD = .38$ s). There were no significant interactions for task phase x age group, task phase x facial expression condition, or task phase x age group x facial expression condition. There was a significant main effect for age group on response time ($F(1, 79) = 345.50, p < .001$). Response time for younger adults ($M = 1.34$ s, $SD = .24$ s) was significantly faster than for older adults ($M = 1.98$ s, $SD = .24$ s), illustrated by Figure 14. The main effect for facial expression condition was not significant, nor was the interaction of age group x facial expression condition.

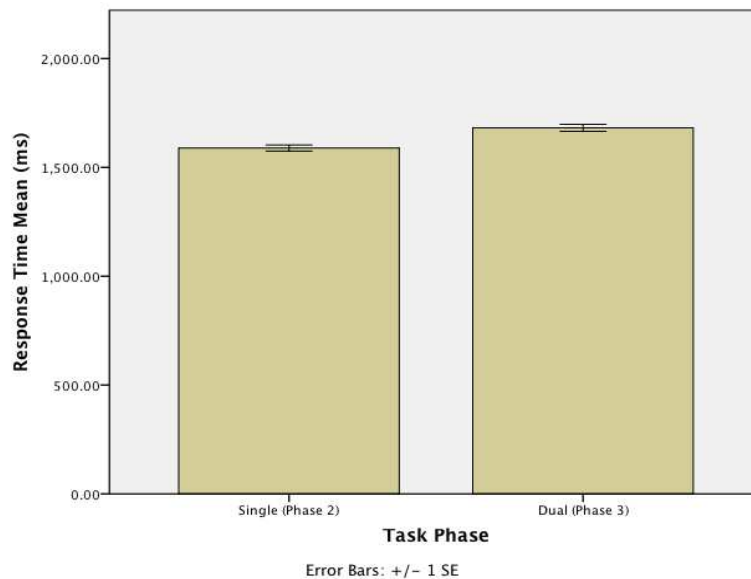


Figure 13. Mean response time (ms) by task phase.

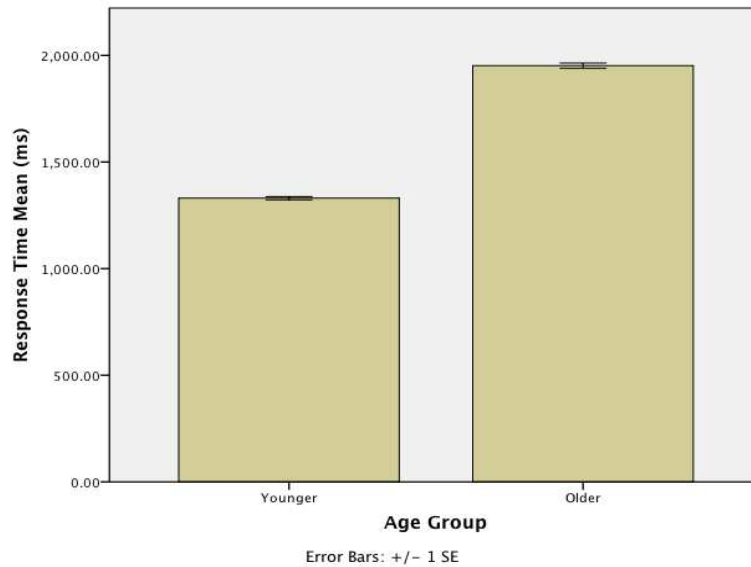


Figure 14. Mean response time (ms) by age group.

Face Misses

A mixed measures ANOVA was conducted on the amount of face misses between the single and dual-task phase. A significant main effect was found for task phase ($F(1, 79) = 276.68, p < .001$), such that participants had fewer misses in the single-task ($M = 7.24, SD = 4.74$) compared to the dual-task ($M = 33.55, SD = 14.10$), illustrated by Figure 15. There were no significant interactions for task phase x facial expression condition, task phase x age group, or task phase x facial expression condition x age group. There was no significant main effect for facial expression condition or age group. There was also no significant interaction for facial expression condition x age group.

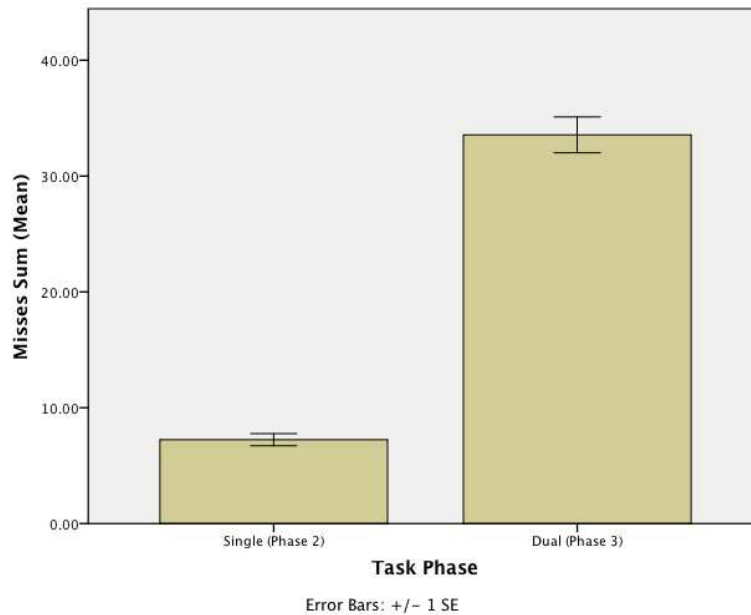


Figure 15. Mean number of face misses by task phase.

Blocks Cleared

A 2 (age group) x 2 (facial expression condition) ANOVA was conducted on the number of blocks cleared in the dual-task phase. There was a significant main effect for age group ($F(1,79) = 160.29, p < .001$), such that younger adults cleared significantly more blocks ($M = 46.95, SD = 10.37$) than older adults ($M = 20.07, SD = 8.61$), illustrated by Figure 16. There was no significant main effect of facial expression condition or significant interaction of age group x facial expression condition.

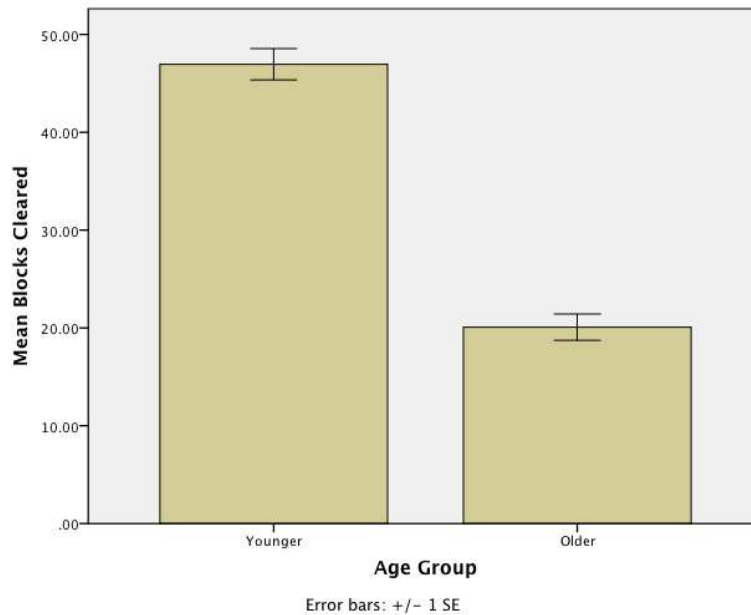


Figure 16. Mean blocks cleared by age group.

NASA-TLX Survey

The NASA-TLX subjective workload survey was given to all participants in order to assess the amount of perceived workload they experienced during the dual-task phase of the experiment. Data was only collected after the dual task phase, so a comparison across task phase could not be analyzed. A 2 (age group) x 2 (facial expression condition) ANOVA was run to determine if the independent variables of age group and facial expression condition had a significant effect on computed workload. There was no significant main effect for age group ($F(1, 78) = .17, p = .68$), for facial expression condition ($F(1, 78) = 2.41, p = .13$), or for the interaction of age group x condition ($F(1, 78) = 1.64, p = .21$). Neither age group nor facial expression condition significantly affected participants' subjective workload, illustrated by Figure 17.

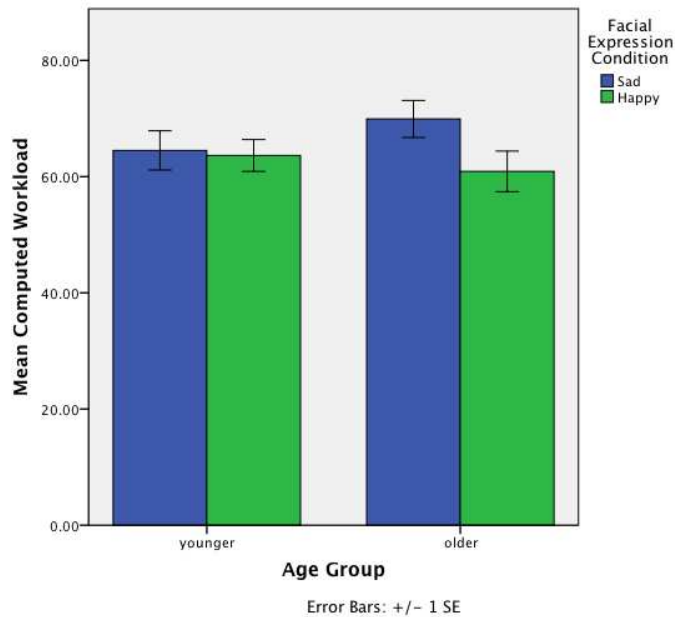


Figure 17. Mean computed workload by age group for sad and happy facial expression conditions.

In sum, the results of the analysis of task phase 3 show that facial expression decoding accuracy did not significantly differ as a function of task phase, but the measures of intensity key pressed, response time, and face misses did show a dual-task cost. There was a main effect of task phase on response time for all participants, which showed faster response times in phase 2 compared to phase 3. A main effect of age group showed older adults to be significantly slower in response time compared to younger adults. There was also a main effect of task phase on the amount of faces that were missed, which showed more faces were missed in phase 3 than phase 2, however this did not differ by age group or facial expression condition. The two-way interaction of age group x task phase was significant and showed mean intensity key pressed significantly increased for older adults across task phase compared to younger adults.

DISCUSSION

The goal of the current study was to investigate whether Chernoff face stimuli could serve as ambient (i.e., relatively resource-free) indicators of quantitative information, using a dual-task paradigm. It was hypothesized (H_1) that a significant three-way interaction would occur between face presented x age group x facial expression condition for decoding performance in the single-task phase. Both age groups were expected to have similar decoding accuracy (i.e., similar regression slopes) in the happy facial expression condition, but non-similar slopes in the sad facial expression condition. This age-related difference in decoding accuracy as a function of facial expressions being happy or sad, was based on literature indicating positive facial expression provided a decoding advantage (Bartneck & Reichenbach, 2005; Calvo & Lundqvist, 2008; Rellecke, 2011), and literature that suggested older adults could decode positive facial expressions as accurately as younger adults (Orgeta & Phillips, 2007).

Hypothesis 1: A Three-Way Interaction of Age Group, Facial Expression Condition, and Face Presented

Hypothesis 1 was not fully supported. The current experiment revealed that the interaction between face presented x age group x facial expression condition for decoding performance in the single-task phase was not significant. However, it was found that the relationship between younger and older adults' decoding accuracy did significantly change due to facial expression condition. There was an age-related difference in decoding accuracy in the happy face condition. Younger adults' significantly higher decoding accuracy in the happy face condition was unexpected due to the

“happy face advantage” that was anticipated for older adults (Ekman & Friesen, 1975; Orgeta & Phillips, 2007; Calvo & Lundqvist, 2008). There was not an age-related difference in decoding accuracy in the sad face condition. The absence of an age-related difference in decoding accuracy in the sad facial expression condition was also unexpected. The similarity of decoding accuracy performance between younger and older adults in the sad face condition was not hypothesized, and may be evidence of the lack of a negativity effect for younger adults, which was based on previous research (Lynchard & Radvansky, 2012).

Participants’ (collapsed across age group) had higher decoding accuracy when they were presented with happy facial expressions. This finding supports a general “happy face advantage” across age group and suggests that when compared to sad Chernoff facial expressions, happy Chernoff facial expressions are more advantageous for decoding. In terms of using a Chernoff face for the display of quantitative information; the use of happy facial expression was shown to be an overall more decodable stimuli. This finding corroborates with previous research that also provides evidence of more accurate happy face decoding (Hess, 1997). While this finding doesn’t fully support hypothesis 1, it does add support to the general hypothesis that happy Chernoff faces would be decoded the most accurately compared to sad Chernoff faces.

Younger adults had significantly faster response times compared to older adults, regardless of the facial expression condition. This was not expected and did not support the hypothesis that happy facial expression would allow older adults to maintain a similar response time as younger adults in the happy facial expression condition (i.e., happy face

advantage). Previous research showing the capacity of quick decoding for happy facial expressions (Calvo & Lundqvist, 2008) was paired with the socioemotional selectivity theory (Carstensen, Issacowitz, & Charles, 1999) to reach the concept of older adults decoding happy facial expression with quickness. Since response time was interpreted as a measure of attentional demand on the participant, it was inferred that older adults' incurred a higher attentional demand when performing the facial decoding task. The non-main effect of facial expression condition showed that happy and sad facial expressions were responded to with similar response times within age groups. This was expected for younger adults (i.e., no decrement in response time due to facial expression condition), but not for older adults. The non-significant difference for older adults' response times in terms of facial expression condition indicates no response time advantage for either facial expression.

The main effect of facial expression condition on faces missed indicated participants in the sad facial expression condition missed significantly more faces than participants in the happy facial expression condition. This supports the general idea that happy faces are more quickly (i.e., perhaps pre-attentively) decoded than sad faces. This finding partially supports hypothesis 1. It was expected for older adults to miss significantly more sad facial expressions, but younger adults were expected to see no change in faces missed across facial expression condition. The main effect of facial expression condition showed that sad Chernoff faces were missed significantly more regardless of age group. However, this preliminary finding indicating a pre-attentive or

resource-free quality of happy faces was more thoroughly investigated in phase 3, where additional attentional demand was placed on the participants.

The finding of participants' significantly higher decoding accuracy for happy facial expressions can be paired with participants' lower amount of misses for happy facial expressions. This forms a case that happy facial expressions are generally more easily decodable than sad facial expressions, which is consistent with previous research (Hess, 1997; Bartneck & Reichenbach, 2005; Calvo and Lundqvist, 2008). The results yielded from the testing of H_1 gave evidence that happy facial expressions have a significant advantage for decoding, in situations of low attentional demand. However, it is important to remember that older adults performed significantly lower than younger adults in terms of decoding accuracy (when collapsed across facial expression condition) and response time. This suggests that older adults had difficulty decoding the Chernoff facial expressions. Because of this finding, Chernoff facial expressions ability to transcend age group as a type of ambient display is suspect.

An aspect of the current study that may have contributed to the absence of an older adult happy face advantage (in phase 2) was the amount of intensity levels for the variable of face presented. Unlike previous studies (Hess, 1997; Orgeta & Phillips, 2007), faces in the current study changed incrementally by 10 % on a scale from 0 % - 90 %. Thus, we may have increased the amount of discrimination required of our participants. It was shown in previous research that 10 % intensity level steps were too small to be discriminated, and participants were not as accurate in their decoding (Bartneck & Reichenbach, 2005).

The manipulation of only one facial feature may not have been optimal for facial expression decoding in adults. A plausible explanation for older adults' lower decoding accuracy was the simplistic level of face manipulation used on the Chernoff faces (i.e., only the mouth was manipulated). Perceiving slight changes in mouth curvature of the Chernoff faces may have been too difficult a task for older adults. A previous study suggested that children (ages 11-12) were more successful at recognizing changes in single features (e.g., mouth, eyebrows) than adults (ages 20-45) (Tsurusawa, Goto, Mitsudome, Nakashima, & Tobimatsu, 2007). This was due to the lack of development of holistic facial expression decoding in children. The current study generalizes this finding to older adults due to their observed lower slope value in facial decoding accuracy. Potentially, the ability for people to discern slight manipulations of a single facial feature is negatively associated with age. The concept of a "pseudo-Chernoff face", which manipulated only one facial feature, was shown to be difficult for older adults to decode. Although the percentage information conveyed by the Chernoff face was univariate in nature, it may be more helpful to manipulate multiple facial features to communicate such information. The holistic manipulation of a face (i.e., mouth, eyes, eyebrows, etc.) could provide a better decoding accuracy for both younger and older adults. The idea presented by Montello and Gray (2005) of communicating data univariately seems to have been misapplied to facial expression in the current study. Unintentionally, we may have created a more difficult decoding task by manipulating only one facial characteristic.

Hypothesis 2: A Four-Way Interaction of Age Group, Facial Expression Condition, Face Presented, and Task Phase

It was hypothesized (H₂) that participants' performance across age groups in the dual-task condition would not significantly decline when in the happy facial expression condition, while a dual-task cost would be observed in the sad facial expression condition. This expected finding was linked to the happy face advantage used as a basis for hypothesis 1 (Ekman & Friesen, 1975; Orgeta & Phillips, 2007; Calvo & Lundqvist, 2008).

The four-way interaction associated with hypothesis 2 was not supported, and confirmed that the three-way interaction of face presented x age group x facial expression condition did not significantly differ across task phase. Decoding accuracy in the dual-task phase was statistically similar to the single task phase. Every interaction that involved decoding accuracy as a function of task phase yielded non-significant results. This was an unexpected finding and presents a question as to why there was no dual-task cost.

The main effect of task phase and main effect of age group on response time suggests that the dual-task phase was contributing to a decrease in performance. Therefore, the prediction that happy facial expressions do not produce a significant increase in response time was not supported. The happy face stimuli used in our study were not immune to dual-task cost. As previous research has stated, (Morris, 1998; Whalen, 1998) the potential advantage of using a face as an ambient display is the face's ability to not add any cognitive load on the user, specifically in an attentional demanding

situation. Response time data has shown Chernoff facial expressions do not meet this requirement, and hence may not be good ambient displays. The main effect for age group suggested that older adults were significantly slower at decoding facial expressions. The slower response time for older adults was also seen in the single task phase.

The amount of misses a participant incurred was significantly different based on task phase. Participants recorded significantly more misses on average (by a factor of 4) in the dual-task condition than the single-task condition. Just as response time indicated a dual-task cost, so do the amount of misses observed for participants. This finding does not fully support hypothesis 2. Since misses significantly increased for both happy and sad facial expressions, there was no apparent happy face advantage. The significant main effect for facial expression condition shown in phase 2 (i.e., sad faces yielded more misses) was not shown in phase 3.

Participants' number of blocks cleared for the block game (in the dual-task phase) was significantly different based on age group. Younger adults cleared more blocks than older adults when completing the dual-task. This finding suggests that younger adults were able to complete the primary block task at a higher level than older adults. There was no significant main effect of facial expression condition, which showed participants did not significantly differ in number of blocks cleared based on which facial expression condition they were placed.

One potential answer to the question of no dual-cost for decoding accuracy is that the primary task in the dual-task phase was not engaging enough. The relationships for the two-way interactions observed in phase 2 may not have significantly changed in

phase 3 because participants' were not being exposed to a high attentional demanding situation (i.e., relative to phase 2). However, the data from response time and amount of face misses provide evidence that the dual-task condition was causing dual-task cost among participants. The lack of dual-cost for decoding accuracy may be explained by the significant difference observed between decoding accuracy as a function of age group in phase 2. Younger adults had a significantly higher decoding accuracy (collapsing across facial expression condition) than older adults in the single-task phase (phase 2). However, younger and older adults may have experienced a floor effect in decoding accuracy that prevented the expected significant decrease in decoding accuracy (in the sad facial expression condition) from phase 2 to phase 3. This indicates that participants' significantly lower decoding accuracy for sad Chernoff facial expressions might not be directly due to the additional attentional demand of phase 3, but is due to the general difficulty of decoding the sad Chernoff facial expressions. Similar to the single task phase, the facial expression stimuli may not have conveyed emotion clearly enough (possibly due to the manipulation of only one facial feature) to result in the expected three-way interaction across task phase.

One possibility for the consistent slower response times for older adults, as previously mentioned, is related to the stimuli. The stimuli were potentially more difficult for the older adults to decode. This detracts from the universal usability (i.e., usable for all age groups) of Chernoff faces as a method for communicating information. A second possibility is that the input of decoding facial expression was more physically taxing for the older adults. Using the number pad may have been a difficult input for older adults

who have joint disorders (e.g., arthritis) or other physical ailments. A more novel input mode (e.g., speech) may provide a way to avoid the confounding variable of input mechanism.

When looking at the response time and face misses data, there is an underlying concept pertaining to Chernoff faces that may explain the dual-task cost. Previous research claimed that Chernoff faces were not processed in parallel and were more difficult to decode (Morris, Ebert, Rheingans, 2000). The concept that Chernoff faces are not pre-attentive and are processed serially adds support to the dual-task cost seen in the current study.

The age-related effect found for the number of blocks cleared gave evidence that younger adults became better adapted to the dual-task phase than older adults. The proficiency shown by younger adults in the block task could help explain why there was a younger adult advantage for decoding accuracy in the dual-task phase. Older adults' significantly lower decoding accuracy in the dual-task could be attributed to the difficulty of the block task. The cognitive demands of the block task may have caused older adults to experience a significant performance decrement when compared to younger adults, in both the number of blocks cleared and decoding accuracy. Due to the lack of an effect of facial expression condition, it can be inferred that the happy face advantage shown in the dual-task was not due to participants' inappropriate allocation of attention in the dual-task. Essentially, participants' higher decoding accuracy in the happy face condition was not due to their neglect of the primary task.

In sum, the results gained from the comparison of performance measures across task phase indicated attention-demanding environments degrade the decoding of Chernoff faces. While decoding accuracy performance did not show a dual-task cost, response time and amount of face misses revealed a significant dual-task cost. Based on decoding accuracy performance, happy facial expression appear to be more beneficial than sad facial expression in an attention-demanding environment. Even though the happy facial expression condition shows significantly higher decoding accuracy, it is not immune to dual-task cost in terms of response time and the amount of misses incurred. Younger adults experienced less decrement in overall performance compared to older adults in the dual-task. Results from the number of blocks cleared by participants in the dual-task phase showed younger adults out performed older adults on the primary task. The block game appeared to be more cognitively demanding for older adults, which may have led to lower decoding accuracy. The dual-task cost seen for response time and face misses indicated that Chernoff facial expressions create a significant demand on users' attention. Therefore, Chernoff faces do not have an observed benefit for communicating information in a resource-free manner.

There were a few limitations to this study that could be improved upon in future research. The facial expressions stimuli could have been manipulated to take advantage of more facial features when conveying expression. Future studies could measure decoding performance for Chernoff faces with variations of manipulated facial characteristics (e.g., manipulation of mouth and eyes, versus manipulation of mouth, eyes, and eyebrows). Another limitation was only having participants complete a NASA-

TLX survey after the dual task phase. It would be beneficial to have participants complete the NASA-TLX survey after the single-task as well. This would allow for comparison of subjective workload between task phases in an effort to gain another measure of dual-task cost. A trust rating measure was not included in the current study, but could be in a future study as a measure of subjective trust concerning the facial expressions. It would be interesting to observe how a participants' trust is affected by the independent variables of: age, facial expression intensity, and facial expression condition. Understanding which faces receive significantly different trust ratings would add an interesting element to a future study. Another improvement for the current study involves the placement of the Chernoff face in the computer program. The peripheral position of the Chernoff face may have put participants at a disadvantage for decoding. A future study may place the facial expression in a more centralized location. A final improvement could be to add more facial expression conditions. Previous literature has expressed an "anger superiority" effect (Ohman, Lundqvist, Esteves, 2001), which could be investigated using Chernoff facial expressions.

CONCLUSION

The results of this study suggest that Chernoff faces communicate facial expression more effectively when happy facial expressions are used. However, older adults have more difficulty in decoding Chernoff facial expressions. There is also a dual-task cost for the decoding of Chernoff faces in terms of increased response time and a higher amount of faces missed. The ability for Chernoff faces to act as effective ambient

displays was not supported by this study, but more research on Chernoff faces should be conducted to further explore their usefulness in communicating information.

APPENDICIES

APPENDIX A

Screenshot of Block Game Task (Phase 1)

Instructions for Task 1

First, you will get some practice on a game.

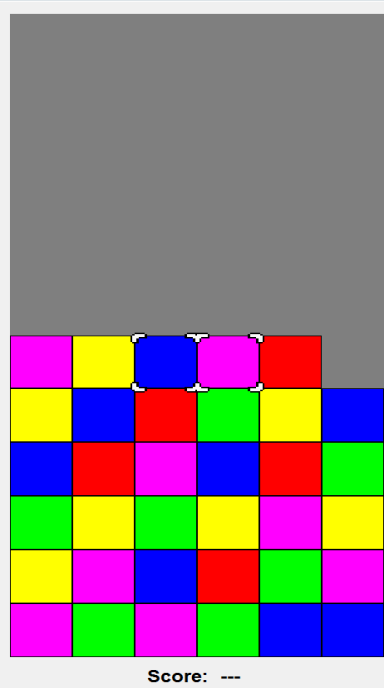
In this game, you must match at least three blocks vertically or horizontally of the same color. But you can only switch any two blocks horizontally.

Use the cursor keys (up, down, left, right) to move your selector.

Press the space bar to switch blocks.

Please work as quickly as you can to increase your score. This part of the study will end automatically.

Click "Start practice" to begin.



Score: ---

APPENDIX B

Screenshot of Facial Expression Decoding Task (Phase 2)

Instructions for Task 2

That is the end of your game practice. Do you have any questions?

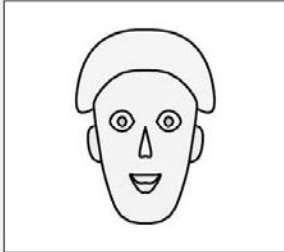
Now we will move to your other task practice. On the right side of the screen you will see a face appear in the white box.

When you see the face you should identify the level of emotional expression on the presented face.

You will use the keys from 0 to 9 to indicate no expression (key 0) to high expression (key 9) and any in-between.

You should use your own judgment--there are no right or wrong answers.

When you are ready to begin, please click "Start practice".



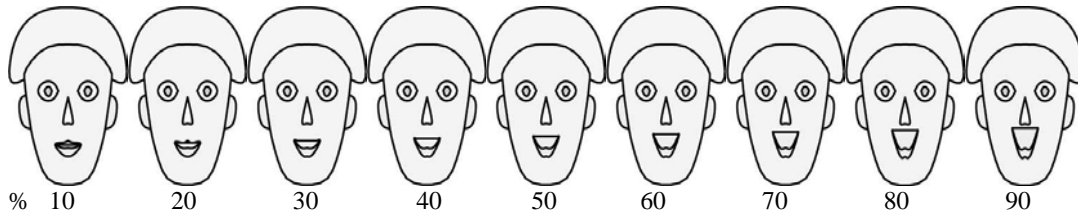
APPENDIX C

Chernoff Facial Expression Stimuli Organized by Expression and Intensity

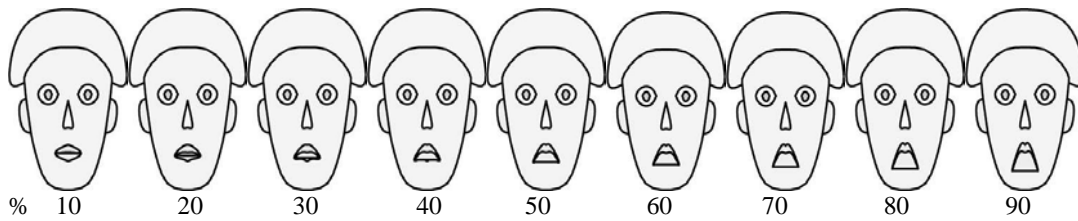
Neutral Facial Expression



Happy Facial Expressions



Sad Facial Expressions



APPENDIX D

Screenshot of Block Task and Facial Expression Decoding Task (Phase 3)

Instructions for Task 3

Now, you will do both tasks at the same time. That is, you will have the blocks game and the face identification task occurring at the same time.

Like before, you will control the blocks game by using the cursor keys (up, down, left, right) and the space bar to switch any two blocks horizontally.

You will also identify the level of emotion expressed on the presented face in the far right. Like before, you will use the keys from 0 to 9 to indicate no expression (key 0) to high expression (key 9) and any in-between.

Doing these two tasks at the same time is very challenging. Your main focus should be the blocks game. You should try to maximize your score as quickly as possible.

Any reserve attention you have available should be used for the face identification task.

Do you have any questions?
Please ask the experimenter now.

If you are ready, please click "Start experiment".

Score: 3

REFERENCES

- Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdale. *Nature*, ProQuest Nursing & Allied Health Source, 669.
- Antifakos, S., Kern, N., Schiele, B., & Schwaninger, A. (2005). Towards improving trust in context-aware systems by displaying system confidence. Proceedings from Mobile HCI '05: *The 7th International Conference on Human Computer Interaction with Mobile Devices & Services*. Salzburg, Austria.
- Bartneck, C., & Reichenbach, J. (2005). Subtle emotional expressions of synthetic characters. *International Journal Human-Computer Studies*, 62, 179-192.
- Batty, M., & Taylor, M. J. (2003). Early processing of the six basic facial emotional expressions. *Cognitive Brain Research*, 17, 613-620.
- Bubb-Lewis, C., & Scerbo, M. (1997). Getting to know you: Human computer communication in adaptive automation. In M. Mouloua & J. M. Koonce (Eds.), *Human-automation interaction: Research and practice* (pp. 92-99). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Calvo, M.G., & Lundqvist, D. (2008). Facial expressions of emotion (KDEF): Identification under different display-duration conditions. *Behavior Research Methods*, 40(1), 109-115.
- Carstensen, L.L., Issacowitz, D.M., & Charles, S.T. (1999). Taking time seriously: A theory of socioemotional selectivity. *American Psychologist*, 54(3), 165-181.
- Chernoff, H. (1973). The use of faces to represent points in k dimensional space graphically. *Journal of the American Statistical Association*, 68(342), 361-368.
- Ekman, P. & Friesen, W.V. (1975). *Unmasking the face: A guide to recognizing emotions from facial cues*. Oxford, England: Prentice-Hall.
- Hess, U., Blairy, S., & Kleck, R. E. (1997). The intensity of emotional facial expressions and decoding accuracy. *Journal of Nonverbal Behavior*, 21(4), 241-257.
- Kabulov, B.T. (1992). A method for constructing Chernoff faces oriented toward interval estimates of the parameters. *Soviet Journal of Computers and System Sciences*, 30(3), 94-97.

- Lee, J. D. (2006). Affect, attention, and automation. In A. Kramer, D. Wiegmann & A. Kirlik (Eds.), *Attention: From theory to practice*. New York: Oxford University Press.
- Lee, M.D., Reilly, R.E., & Butavicius, M.A. (2003). An empirical evaluation of Chernoff faces, star glyphs, and spatial visualizations for binary data. *Proceedings from APVis '03: Asia Pacific Symposium on Information Visualization, 24*, 1-10.
- Lynchard, N.A., & Radvansky, G.A. (2012). Age-related perspectives and emotion processing. *Psychology and Aging*. Advance online publication. doi: 10.1037/a0027368.
- Mankoff, J., Dey, A. K., Hsieh, G., Kientz, J., Lederer, S., Ames, M. (2003). Heuristic evaluation of ambient displays. Proceedings from SIGCHI '03: *Conference on Human Factors in Computing Systems*. Ft. Lauderdale, FL, USA.
- Montello, D.R., & Gray, M.V. (2005). Miscommunicating with isolines: Design principles for thematic maps. *Cartographic Perspectives*, 10-19.
- Morris, C. J., Ebert, D. S. & Rheingans, P. (2000). An experimental analysis of the effectiveness of features in chernoff faces. Proceedings from SPIE '00: *The 28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making*. Washington, D. C., USA.
- Morris, J. S., Friston, K. J., Buchel, C., Firth, C. D., Young, A.W., Calder, A. J., & Dolan, R. J. (1998). A neuromodulatory role for the human amygdala in processing emotional facial expressions. *Brain*, 121, 47-57.
- Nelson, E.S. (2007). The face symbol: Research issues and cartographic potential. *Cartographica*, 42(1), 53-64.
- Ohman, A., Lundqvist, D., & Esteves, F. (2001). The face in the crowd revisited: A threat advantage with schematic stimuli. *Journal of Personality and Social Psychology*, 80(3), 381-396.
- Orgeta, V., & Phillips, L. H. (2007). Effects of age and emotional intensity on the recognition of facial emotion. *Experimental Aging Research*, 34(1), 63-79.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230-253.
- Poupyrev, I., Maruyama, S., & Rekimoto, J. (2002). Ambient touch: Designing tactile interfaces for handheld devices. Proceedings from UIST '02: *The 15th Annual ACM Symposium on User Interface Software and Technology*. Paris, France.

- Rellecke, J., Palazova, M., Sommer, W., & Schacht, A. (2011). On the automaticity of emotion processing in words and faces: Event-related brain potentials evidence from a superficial task. *Brain and Cognition, 77*, 23-32.
- Sawhney, N., & Schmandt, C. (2000). Nomadic radio: Speech and audio interaction for contextual messaging in nomadic environments. *ACM Transactions on Computer Human Interaction, 7*(3), 353-383.
- Tsurusawa, R., Goto, Y., Mitsudome, A., Nakashima, T., Tobimatsu, S. (2008). Different perceptual sensitivities for Chernoff's face between children and adults. *Neuroscience Research, 60*(2), 176-183.
- Weiser, M., & Brown, J. S. (1995). Designing calm technology. Retrieved from <http://www.ubiq.com/weiser/calmtech/calmtech.htm>.
- Whalen, P. J., Rauch, S. L., Etcoff, N. L., McInerney, S. C., Lee, M. B., & Jenike, M. A. (1998). Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *The Journal of Neuroscience, 18*(1), 411-418.
- Wisneski, C. (1999). The design of personal ambient displays. (Unpublished master's thesis). MIT, Boston, MA.

APPENDIX 8: Branyon, J. (in progress). Investigating older adults' trust, causal attributions, and perception of capabilities in robots as a function of robot appearance, task, and reliability.

Investigating Older Adults' Trust, Causal Attributions, and Perception of Capabilities in Robots
as a Function of Robot Appearance, Task, and Reliability

A Thesis Proposal by:

Jessica J. Branyon

Clemson University Graduate School

Committee Members:

Dr. Richard Pak (Chair)

Dr. Patrick Rosopa

Dr. Kelly Caine

Abstract

The purpose of the current study is to examine the extent to which the appearance, task, and reliability of a robot is susceptible to stereotypic thinking. Stereotypes can influence the types of causal attributions that people make about the performance of others. Just as causal attributions may affect an individual's perception of other people, it may similarly affect perceptions of technology. Stereotypes can also influence perceived capabilities of others. That is, in situations where stereotypes are activated, an individual's perceived capabilities are typically diminished. The tendency to adjust perceptions of capabilities of others may translate into levels of trust placed in the individual's abilities. A cross-sectional factorial survey using video vignettes will be utilized to assess young adults' and older adults' attitudes toward a robot's behavior and appearance. We hypothesize that a robot's older appearance will result in lower levels of trust, more dispositional attributions, and lower perceptions of capabilities while high reliability should positively impact trust.

Investigating older adults' trust, causal attributions, and perception of capabilities in robots as a function of robot appearance, task, and reliability

When interacting with technology, people focus on human-like qualities of the technology more than the asocial nature of the interaction (Reeves & Nass, 1996; Nass & Moon, 2000) attributing human-like qualities such as personality, mindfulness, and social characteristics. The attribution of human-like qualities makes technology susceptible to stereotyping based on appearance and etiquette (e.g., Nass & Lee, 2001; Parasuraman & Miller, 2004; Eyssel & Kuchenbrandt, 2012). For example, when a male or female anthropomorphic computerized aid was included in a trivia task, participants were more likely to trust the male aid's suggestions and ranked the female aid as less competent (Lee, 2008).

The purpose of the current study is to examine the extent to which the appearance, task, and reliability of a robot is be susceptible to stereotypic thinking. The theoretical relevance is that the results of this study will inform the limits of stereotypic thinking by investigating whether stereotypes are applied to robots. The practical relevance is that the current study may inform the design of robots to enhance human-robot interaction, particularly for older adults who tend to be less accepting of technological aids than other age groups (Czaja et al., 2006).

Stereotypes and Aging

In order to make efficient social judgments about others, individuals rely on the use of heuristics. One example heuristic involves placing an individual into a pre-determined schema (i.e., a stereotype). Stereotypes are cognitive shortcuts that result in impressions of others (e.g., Ashmore & Del Boca, 1981). Therefore, older adults may be more likely than younger adults to apply stereotypes when they do not have other sources of information available to them (i.e., under situations of ambiguity).

Stereotypes are more likely to be activated in domains that are inconsistent with prescriptive societal gender or age roles (e.g., Kuchenbrandt, Häring, Eichberg, Eyssel, & André, 2014). For example, individuals perceived a female-voiced computer to be more informative about romantic relationships than the male-voiced computer (Nass, Moon, & Green, 1997). Although gender stereotypes have been studied using anthropomorphic technological aid paradigms, aging stereotypes have been investigated to a lesser degree within this context. Pak, McLaughlin, & Bass (2014) examined whether the physical appearance of an anthropomorphic aid would activate stereotypic thinking and affect individuals' trust in the aid. Using a factorial design, Pak et al. manipulated the technological aid's gender and age (younger, older) as well as participants' perceptions of the reliability of the automation. Participants were told that the automation was either 45%, 70%, or 95% reliable. However, the automation always provided a correct answer during testing. The task in this study was a health behaviors test regarding participants' knowledge about diabetes. Before beginning the task, participants were told that the automated aid was a Smartphone application recommended by a doctor designed to help people make the best decisions about diabetes. As the participants answered each question, the decision aid smart phone app would appear on the screen and the agent would recommend a correct answer. All of the agents were dressed as doctors. Participants rated their subjective trust in the automation and whether they would actually use the advice of the application on a 1-7 Likert scale.

Pak, McLaughlin, & Bass (2014) found that both younger and older adult participants trusted the older anthropomorphic aids more than the younger aids, the male aids more than the female aids, and more reliable applications than less reliable applications. However, stereotypic thinking was activated when perceptions of reliability were low or ambiguous. When the app had

low reliability, the younger female aid was trusted less than younger male agents. Also, under medium reliability, the older female aid was trusted less than the older male aid. These results suggest that trust in automation can be influenced by physical appearance (i.e., gender and perceived age) of the technology. These results also further support the notion that technology is, like humans, also susceptible to stereotyping.

Physical appearance is known to play a large role in the activation of aging stereotypes. The link between physical characteristics and stereotypes has been well established in the social cognition literature (Brewer & Lui, 1984; Hummert, 1994; Hummert, Garstka, & Shaner, 1997). Within this context, facial features are considered to be the main source of information used in order to activate stereotypes. Hummert et al. (1997) found that negative age stereotypes were associated with the perception of advanced age through facial photographs. Overall, these findings suggest that physical cues are major indicators within the context of social judgments.

Stereotypes about older adults, although pervasively negative, can be multidimensional in the right context. People hold both positive and negative stereotypes about older adults (Hummert, 1993). When adults of all ages completed a trait card-sorting task where they were asked to generate traits they associated with older adults, Hummert and colleagues (1994) found approximately 10 different aging stereotypes, including positive ones like the “golden ager” who leads an active and engaged lifestyle. Although many stereotypes are held in common by people of all ages, aging stereotypes tend to become increasingly differentiated as people grow older (Hummert, 1993; Hummert et al., 1994).

Stereotypes and other social beliefs can influence the way in which individuals process information in order to form social judgments, including the types of causal attributions that people make about the performance of others (Fiske & Taylor, 1991). When trying to determine

the causality of an event, people tend to use two types of information: internal or dispositional qualities of the individuals involved in an outcome and the influences of the situation itself (Gilbert, 1993; Krull, 1993; Krull & Erikson, 1995). Potential biases in the attribution process can occur as a function of the valence of the situational outcome, the degree of ambiguity of the situation (or of the information given about causal factors), and the controllability of the situation (Blanchard-Fields, 1994). Blanchard-Fields suggested that, in general, older adults are most likely to make dispositional attributions when the outcome of a situation was negative and the actor's role in the outcome was ambiguous. When personal beliefs about another individual or situation are violated, older adults are also more likely to make dispositional attributions of blame rather than situational (Blanchard-Fields, 1996; Blanchard-Fields, Hertzog, & Horhota, 2012). Just as causal attributions, or the extent to which behavior is attributed to situational or dispositional causes, may affect an individual's perception of other people, it may also similarly affect perceptions of technology. For example, blaming technology for unreliable performance is likely to induce less trust (Moray, Hiskes, Lee, and Muir, 1995; Madhavan, Wiegmann, & Lacson, 2006). Attribution of fault has been studied in the automation and has been referred to as automation bias (Mosier & Sitka, 1996). Automation bias has been defined "as a heuristic replacement for vigilant information seeking and processing" (Mosier & Sitka, p. 202) which results in increased omission errors and commission errors.

Expectations of performance outcomes are influenced by stereotypes. Adults of all ages expect memory performance to decline with age (Lineweaver and Hertzog, 1998). Similarly, older adults' abilities are perceived negatively in domains involving memory (Kite & Johnson, 1988; Kite, Stockdale, Whitley & Johnson, 2005) and physical well-being (Davis & Friedrich, 2010). In memory taxing situations, older adults are perceived as being less credible and less

accurate (Muller-Johnson, Toglia, Sweeney, & Ceci, 2007). The tendency to adjust perceptions of capabilities of others based on appearance may translate into levels of trust placed in the individual's abilities.

Trust in Automation

Trust in technological agents is important because it affects an individual's willingness to accept robot's input, instructions, or suggestions (Lussier, Gallien, & Guiochet, 2007). For example, Muir and Moray (1996) found a strong positive relationship between adults' level of trust in an automated system and the extent to which they allocated control to the automated system. Interestingly, Muir (1987) suggests that people's trust in technology is affected by factors that are also the basis of interpersonal trust. Trust in automation is thought to develop overtime (Maes, 1994) suggesting that trust is influenced by past experiences with the technology. For example, Merritt and Ilgen (2008) describe dispositional trust as the trust placed in a person or automation during a first encounter before any interaction has been made while history based trust reflects the prior experience a person has with another person or automation.

Performance based factors have a large influence in perceived trust in HRI (Brule, Dotsch, Bijlstra, Wigboldus, & Haselager, 2014). In fact, a recent meta-analysis suggests that a robot's task performance was the most important factor in adults' trust in robots (Hancock et al., 2011). That is, if the robot performs reliably, the human will exhibit greater trust towards the robot. The same meta-analysis found that behavior, proximity, and size of the robot also affect trust to a lesser extent. However, human-automation trust literature suggests that appearance can have reliable effects on trust (Pak Fink, Price, Bass, & Sturre, 2012). Indeed, studies in the social literature have found that people often judge an individual's levels of trustworthiness based on facial appearance (Oosterhof & Todorov, 2008) and that trust judgments can be formed after

only a brief exposure (100 ms) to a face (Willis & Todorov, 2006). It is also important for the robot's appearance to be compatible with its function at face value. Goetz, Kiesler, & Powers (2003) found that people are more likely to accept a robot when its appearance matches its perceived capabilities. This is thought to be the case because when there is a high level of compatibility between appearance and functionality, users' expectations are confirmed, boosting confidence in the robot's performance. However, when appearance and capabilities are incompatible, user expectations are violated, which can result in lower levels of trust (Duffy, 2003).

Because studies of human robot interaction are a new field, there are many gaps in the literature especially regarding the social influences on HRI. First, although there is evidence to suggest that stereotypes can affect performance and interactions with anthropomorphized technological aids, we do not know how pre-existing age stereotypes will affect HRI. Next, it is unclear how trust might be moderated by task type and reliability. Although the automation literature suggests that reliability can influence trust, to our knowledge the relationship between robot task domain and trust has not yet been investigated. Finally, how does stereotyping technology affect perception of capabilities and the causal attributions made about performance?

The Current Study

The purpose of this study is to better understand the factors that influence older adults' trust in robots. Specifically, we are investigating whether the robots' appearance, task domain, and reliability of the robot's performance influence trust in the automation. A cross-sectional factorial survey study will be utilized using video vignettes to assess participants' attitudes towards the robots' behavior and appearance. Each vignette will include manipulations of the age of the robot, the domain of the collaborative task, and the reliability of the robot's performance.

Dependent measures will include the level of trust participants exhibit toward the robot, causal attributions regarding the robot's performance, and perceived capabilities of the robot.

It is hypothesized that manipulating a robot's appearance, level of reliability, and the task type will have an effect on the level of trust that an older adult exhibits toward a robot, the causal attributions that the individual makes about the robot's performance, and people's perceptions of the capabilities of the robot. Specifically, trust in the robot should be highest when the task is stereotypically congruent with the robot's appearance (e.g., a younger adult performing a cognitive task instead of an older adult performing a cognitive task) and its performance is reliable. This is hypothesized because appearance influences people's trust in automation (Pak, Fink, Price, Bass, & Sturre, 2012) and aging stereotypes will less likely be activated while interacting with the younger robot. The attributions about the robot's performance may be more dispositional when reliability is low and the task is incongruent with the robot's appearance. This is because older adults are more likely to make dispositional (i.e., internal) attributions of blame when an outcome of an event is perceived as negative (the unreliable condition) and when their beliefs are violated (i.e., when an older looking robot performs the cognitive and physical tasks; Blanchard-Fields, Hertzog, & Horhota, 2012). Perceived capabilities of the robot are hypothesized to depend on the robot's appearance. That is, capability ratings are expected to be higher when the younger looking robot performs the tasks, and rankings are expected to be lower when an older looking robot performs the tasks. This is expected because adults' capabilities in cognitive and physical domains are expected to decline with age (Kite, Stockdale, Whitley, & Johnson, 2005; Davis & Friedrich, 2010). Task domain will be treated as an exploratory variable. However, based on automation trust literature suggesting that trust in robot's capabilities might depend on the domain in which they are placed (e.g., industry, entertainment, social; Schaefer,

Sanders, Yordon, Billings, & Hancock, 2012), it is hypothesized that there will be a main effect of task domain such that participants will have more trust in the robot and have higher ratings of perceived capabilities when the robot performs physical tasks.

Method

Participants

50 younger adults and 50 older adults will complete the study. Younger adults will be undergraduate students who receive extra credit for participation. Older participants will normatively aging older adults recruited from the community and will receive \$15 for their participation.

Measures

Individual Difference Measures. Demographic information, vocabulary (Shipley vocabulary; Shipley, 1986), perceptual speed (digit-symbol substitution; Wechsler, 1997), and working memory (automated operation span; Unsworth, Heitz, Schrock, & Engle, 2005) will be measured. The Complacency Potential Rating Scale (CPRS; Singh, Molloy, and Parasuraman, 1993) is designed to measure complacency towards different types of automation. Participants will respond to the extent they agree with statements about automation on a scale of 1–5.

Subjective Trust. Trust will be measured by asking the participants how much they trusted the robot portrayed in the vignette. Responses will be recorded on a Likert scale from 1 (not at all) to 7 (very much). The larger the participants' ratings, the higher their subjective trust in the robot.

Causal Attributions. Causal attributions will be measured using a paradigm adapted from Blanchard-Fields, Chen, Schocke, and Hertzog (1998). Participants will be asked to indicate the degree to which either dispositional factors of the characters or situational factors

influenced the outcome of the scenario. Specifically, participants indicated the extent to which: (a) the robot was responsible for the final outcome, (b) the robot was to blame for the final outcome, (c) the final outcome was due to personal characteristics of the robot, (d) the final outcome was due to characters in the story other than the robot, (e) the final outcome was due to something other than the characters in the story, and (f) both the personal characteristics of the robot and something other than the robot contributed to the final outcome. Participants will respond using a Likert scale from 1 (very little) to 7 (very much). In order to classify the extent to which participants attributed performance to either dispositional or situational variables, we will sum the responses from a-c, which represent dispositional attributions of performance and compare them with participant's summed responses to d-f, which represent situational attributions of the final outcome. The higher the score on these two aspects, the higher the degree of either dispositional attributions or situational attributions.

Perceived Capabilities. Perceived capabilities of the robot will be measured by using a list of questions that span potential capabilities. Participants will be asked, "Based on the robot's behavior in the video you just watched, what other activities could the robot complete?" Participants will be asked about further cognitive capabilities or motor capabilities of the robot. That is, participants will rank their agreement regarding whether the robot could complete similar cognitive or physical tasks. For example, participants could be asked, "Based on the robot's performance, could it also recommend stock investment picks?" or "Based on the robot's performance, could it also vacuum a room?" Afterward, participants will be asked to write a short answer explaining what other tasks they thought the robot could do. Participants will rate the extent to which they think the robot could perform certain tasks on a 1-7 Likert scale ranging

from “Definitely No” to “Definitely Yes” with higher scores indicating increased perceptions of capabilities.

Factorial Survey. In a factorial survey, independent variables (i.e., factors or dimensions) are treated as statistically independent, making it possible to identify and separate their influences on judgments (Rossi & Anderson, 1982). In the current study, the dimensions will include the robot’s age appearance (younger, older), task domain (cognitive, physical) with two tasks per domain, and aid reliability (low, high). The levels of the dimensions will result in 12 factorial combinations or scenarios. Each scenario will be presented twice, creating 24 vignettes.

The stimuli for the robots were selected to portray a younger adult (Figure 1) and an older adult (Figure 2). Because the current study will not manipulate the gender of the robot, the facial stimuli for both the younger and older condition will be female. In order to control for potential effects for different faces, the faces selected for this study represent an age progression of the same female.

The robot used in this study will be the Baxter robot manufactured by Rethink Robotics. Baxter is designed as a manufacturing robot that can complete tasks that involve assembly and object organization (Gear & Gadgets, 2014). Adobe Photoshop CC will be used to superimpose the facial stimuli onto the robot (Figure 3).

Each video vignette will contain a slideshow of pictures portraying a human and a robot completing a collaborative task. The opening scenes will include a wide shot, introducing the positioning of the human and robot as well as the collaborative task. In order to avoid any age or gender biases of the human actor, only the actor’s arms and hands will be shown while aiding in the collaborative task. The next shot will be a close up of the robot’s trunk, arms, and face.

Finally, the human and the robot will complete the task. The final shot will include information about whether the task was performed reliably. If the task was performed reliably, the final shot will show the successfully completed task. If the task was not performed reliably, the final shot will show the final outcome being incorrectly completed or unfinished. As a manipulation check, participants will be asked to respond to the question, “Was the task portrayed in the slideshow completed successfully?” after viewing the slideshow.

During the survey, each video vignette will be presented in the center of the screen. After participants view the video, the questions and rating scales will appear in the lower half of the screen. Scenarios will be presented in a random, counterbalanced order. The survey will be programmed into the online survey program Qualtrics for administration.

Design and Procedure

The study was a 2 (participant age: younger, older) X 2 (robot age: young, old) X 2 (task domain: cognitive, physical) X 2 (robot reliability: low, high) mixed-model design, with participant age as a between-subjects variable. The within-subjects factors are manipulated in the factorial survey. The task domain dimension has two levels: cognitive and physical. These levels were selected in order to encompass the range of task domains within the HRI literature. Within those two domains, participants will view the robots doing two separate tasks. That is, the robots will complete two different cognitive tasks and two different physical tasks throughout the survey. The two cognitive tasks will include sorting recycling and sorting laundry. The two physical tasks will include moving boxes from one location to another and changing a light bulb (Figure 4).

Following participant recruitment, the experimenter will email personalized Qualtrics links to participants in order for them to complete a unique version of the factorial survey. The survey will be completed in their home so no lab visit is necessary. Participants may work through the survey at their own pace. However, they will be instructed to complete the survey in one sitting. In the survey, participants will complete a demographics form along with the vocabulary, perceptual speed, and other individual difference measures. Afterward, participants will view randomly presented vignettes and answer each question after the completion of the video. After making their trust, causal attribution, and capabilities ratings, participants will be asked to briefly explain their ratings. Participants will complete the CPRS at the conclusion of the survey. Finally, participants will be debriefed and compensated for their time.

Anticipated Results

First, outliers will be eliminated from the data. An outlier will be defined as a participant that scored more or less than 3 standard deviations from the mean on a certain measure. In order to investigate whether manipulating a robot's appearance, task, and reliability had an effect on the level of trust, causal attributions, and perception of capabilities, we will use a 2 (participant age: younger, older) X 2 (robot age: young, old) X 2 (task domain: cognitive, physical) X 2 (robot reliability: low, high) repeated measures analysis of variance (ANOVA) for subjective trust, causal attributions, and perceived capabilities. We expect to see main effects of robot appearance such that when the robot appears older, trust will be lower, causal attributions will be more dispositional, and capability of perceptions will be reduced. It is also hypothesized that there will be a significant main effect of reliability such that when reliability is low, trust and capabilities should decrease and attributions will become more dispositional. Although task domain will be treated as an exploratory variable, a main effect of task domain is hypothesized

such that trust and perceived capabilities will be highest when the robot performs physical tasks. We expect a 2 way interaction between reliability and robot age such that when reliability is low, trust in the older adult automation may be lowest (Figure 5). Next, we expect a participant age by robot appearance by reliability interaction on causal attributions such that causal attributions will be most dispositional in older adult participants when robot appearance is older and performance is unreliable (Figure 6). Finally, we expect that older adult participants will make more dispositional attributions across conditions and to have lower trust levels overall.

Discussion

This study offers a unique contribution by investigating a well-researched paradigm from the social cognition and aging literatures, stereotypes, and applying it to a novel field, HRI. If our hypotheses are supported and appearance of the robot has an effect on the levels of trust, attribution, and perceived capabilities of robots, then this data could be useful for informing future design of robotics. For example, the results of people's judgments based on task domain may suggest if certain types of anthropomorphic aids are only appropriate in certain domains. For example, it may not be appropriate to have an older looking robot in manufacturing roles that perform gross motor tasks such as heavily lifting, due to the influence its appearance may have on workers perceptions of its abilities and their trust in the system. This study can also help influence design in the sense that it further investigates which factors influence trust in automation. If the goal is to maximize human trust, then it may be beneficial to use younger looking anthropomorphism rather than older, while keeping reliability high. Overall, human-robot collaboration will become more common in the home as well as in work, thus it becomes critical to better understand how people perceive such technologies.

References

- Amadeo, R. (2014, June 15). Hands-on with Baxter, the factory robot of the future. Retrieved from <http://arstechnica.com/gadgets/2014/06/hands-on-with-baxter-the-factory-robot-of-the-future/1/>
- Ashmore, R. D., & Del Boca, F. K. (1981). Conceptual approaches to stereotypes and stereotyping. In *Cognitive processes in stereotyping and intergroup behavior*. Edited by D. L. Hamilton, 1–35. Hillsdale, NJ: Erlbaum.
- Blanchard-Fields, F. (1994). Age differences in causal attributions from an adult developmental perspective. *Journal of Gerontology, 49*, 43-51. doi:10.1093/geronj/49.2.P43
- Blanchard-Fields, F. (1996). Causal attributions across the adult life span: The influence of social schemas, life context, and domain specificity. *Applied Cognitive Psychology, 10*, 137-146.
- Blanchard-Fields, F., Chen, Y., Schocke, M., Hertzog, C. (1998). Evidence for content-specificity of causal attributions across the adult life span. *Aging, Neuropsychology, & Cognition, 5*, 241-263.
- Blanchard-Fields, F., Hertzog, C., & Horhota, M. (2012). Violate my beliefs? Then you're to blame! Belief content as an explanation for causal attribution biases. *Psychology and Aging, 27*, 324-337. doi:10.1037/a0024423
- Brewer, M. B., & Lui, L. (1984). Categorization of the elderly by the elderly. *Personality and Social Psychology Bulletin, 10*, 585-595.
- Brule, R., Dotsch, R., Bijlstra, G., Wigboldus, D. J., & Haselager, P. (2014). Do robot performance and behavioral style affect human trust?: A multi-method approach. *International Journal Of Social Robotics*, doi:10.1007/s12369-014-0231-5

- Czaja, S. J., Charness, N., Fisk, A. D., Hertzog, C., Nair, S. N., Rogers, W. A., & Sharit, J. (2006). Factors predicting the use of technology: Findings from the center for research and education on aging and technology enhancement (CREATE). *Psychology and Aging, 21*, 333-352. doi:10.1037/0882-7974.21.2.333
- Davis, N. C., & Friedrich, D. (2010). Age stereotypes in middle-aged through old-old adults. *The International Journal Of Aging & Human Development, 70*(3), 199-212. doi:10.2190/AG.70.3.b
- Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems, 42*, 177-190.
- Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal Of Social Psychology, 51*(4), 724-731. doi:10.1111/j.2044-8309.2011.02082.x
- Fiske, S., & Taylor, S. (1991). *Social Cognition*. New York: McGraw-Hill.
- Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117*, 21-38.
- Goetz, J., Kiesler, S., & Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. *The 12th IEEE International Workshop on Robot and Human Interactive Communication, Vol., IXX*, 55-60
- Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*(5), 517-527. doi:10.1177/0018720811417254

- Hummert, M. L. (1993). Age and typicality judgments of stereotypes of the elderly: Perceptions of elderly vs. young adults. *International Journal of Aging and Human Development*, *37*, 217-227.
- Hummert, M. L. (1994). Physiognomic cues and the activation of stereotypes of the elderly in interaction. *International Journal of Aging and Human Development*, *39*, 5-20.
- Hummert, M. L., Garstka, T. A., & Shaner, J. L. (1997). Stereotyping of older adults: The role of target facial cues and perceiver characteristics. *Psychology and Aging*, *12*, 107-114.
- Hummert, M. L., Garstka, T. A., Shaner, J. L., & Strahm, S. (1994). Stereotypes of the elderly held by young, middle-aged, and elderly adults. *Journal of Gerontology*, *49*, 240-249.
- Kite, M. E., & Johnson, B. T. (1988). Attitudes toward older and younger adults: A meta-analysis. *Psychology and Aging*, *3*, 233-244.
- Kite, M. E., Stockdale, G. D., Whitley, B. J., & Johnson, B. T. (2005). Attitudes Toward Younger and Older Adults: An Updated Meta-Analytic Review. *Journal Of Social Issues*, *61*(2), 241-266. doi:10.1111/j.1540-4560.2005.00404.x
- Krull, D. S. (1993). Does the grist change the mill? The effect of perceiver's goals on the process of social inference. *Personality and Social Psychology Bulletin*, *19*, 340-348.
- Krull, D. S., & Erikson, D. J. (1995). Judging situations: On the effortful process of taking dispositional information into account. *Social Cognition*, *13*, 417-438.
- Kuchenbrandt, D., Häring, M., Eichberg, J., Eyssel, F., & André, E. (2014). Keep an eye on the task! How gender typicality of tasks influence human-robot interactions. *International Journal Of Social Robotics*, *6*(3), 417-427. doi:10.1007/s12369-014-0244-0
- Lee, J. (2008). Review of a pivotal human factors article: "Humans and automation: Use, misuse, disuse, abuse." *Human Factors*, *50*(3), 404-410.

- Lineweaver, T. T., & Hertzog, C. (1998). Adults' Efficacy and Control Beliefs Regarding Memory and Aging: Separating General from Personal Beliefs. *Aging, Neuropsychology, and Cognition*, 5, 264-296.
- Madhavan, P., Wiegmann, D., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermines trust in automated aids. *Human Factors*, 48, 241-256.
- Maes, P. (1994). Agents that reduce work and information overload. *Communications of the ACM*, 37, 30-40.
- Merritt, S.M., Ilgen, D.R. (2008). Not all trust is created equal: dispositional and history-based trust in human-automation interactions. *Human Factors*, 50:194-210.
doi:10.1518/001872008X288574
- Moray, N., Hiskes, D., Lee, J., & Muir, B. M. (1995) Trust and human intervention in automated systems. Hillsdale, NJ: Erlbaum
- Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other?. In R. Parasuraman, M. Mouloua (Eds.) , *Automation and human performance: Theory and applications* (pp. 201-220). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
- Muir, B. M. (1987). Trust between human and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27, 527-539.
- Muir, B. M., & Moray, N. (1996). Trust in automation: Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39, 429-460
- Muller-Johnson, K, Togli, M.P., Sweeney, C.D., Ceci, S.J. (2007) The perceived credibility of older adults as witnesses and its relation to ageism. *Behavioral Sciences and the Law*, 25, 355-375.

- Nass, C., & Lee, K. M. (2001). Does computer-generated speech manifest personality? Experimental test of recognition, similarity-attraction, and consistence-attraction. *Journal of Experimental Psychology: Applied*, 7, 171–181.
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56, 81-103.
- Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender stereotypic responses to computers with voices. *Journal of Applied Social Psychology*, 27, 864-876.
- Oosterhof, N.N., Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, 105, 11087–11092
- Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, 55(9), 1059-1072.
- Pak, R., McLaughlin, A., & Bass, B. (2014). A multi-level analysis of the effects of age and gender stereotypes on trust in anthropomorphic technology by younger and older adults. *Ergonomics*, 57, 1277-1289. doi:10.1080/00140139.2014.928750
- Parasuraman, R. & Miller, C. (2004). “Trust and Etiquette in High-Criticality Automated Systems”. In C. Miller (Guest Ed.), special section on “*Human-Computer Etiquette*”. *Communications of the ACM*. 47, April. 51-55.
- Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 201–220). Mahwah, NJ: Erlbaum
- Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. New York: Cambridge University Press.

Rossi, P. H., & Anderson, A. B. (1982). The Factorial Survey Approach: An Introduction.

In P. H. Rossi & S. L. Nock (Eds.), *Measuring Social Judgments. The Factorial Survey Approach* (pp. 15–67). Beverly Hills, CA: SAGE Publications.

Shaefer, K. E., Sanders, T. L., Yordon, R. E., Billings, D. R., & Hancock, P. A. (2012).

Classification of robot form: Factors predicting perceived trustworthiness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56: 1548. doi:

10.1177/1071181312561308

Shipley, W. C. (1986). *Shipley Institute of Living Scale*. Los Angeles: Western Psychological Services.

Singh, I. L., Molloy, R. & Parasuraman, R. (1993). Automation-induced “complacency”:

Development of a complacency-potential scale. *International Journal of Aviation Psychology*, 3(2), 111-122.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior research methods*, 37(3), 498-505.

Wechsler, D. (1997). *Wechsler Memory Scale III*. (3rd Ed.). San Antonio, TX: The Psychological Corporation.

Willis, J., Todorov, A. (2006). First impressions: making up your mind after a 100-*Ms* exposure to a face. *Psychological Science*, 17, 592–598



Figure 1. Young-adult appearance condition



Figure 2: Older-adult appearance condition



Figure 3: Example of Baxter stimuli (older-adult appearance condition)

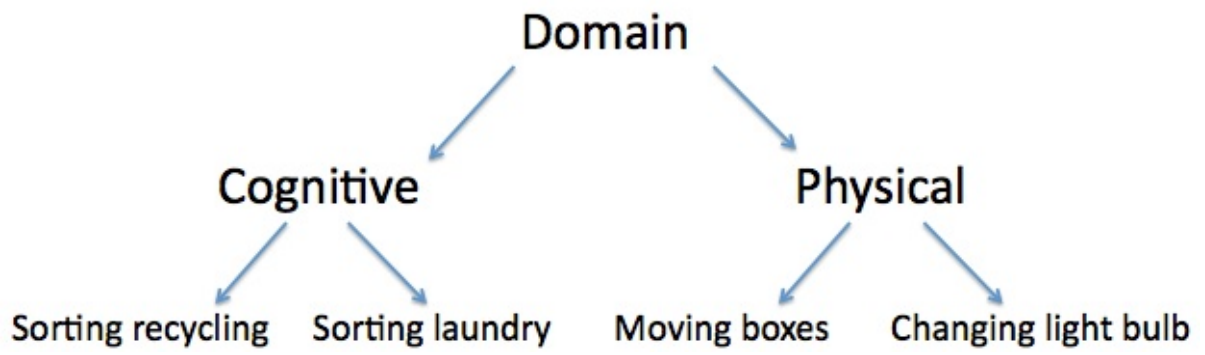


Figure 4

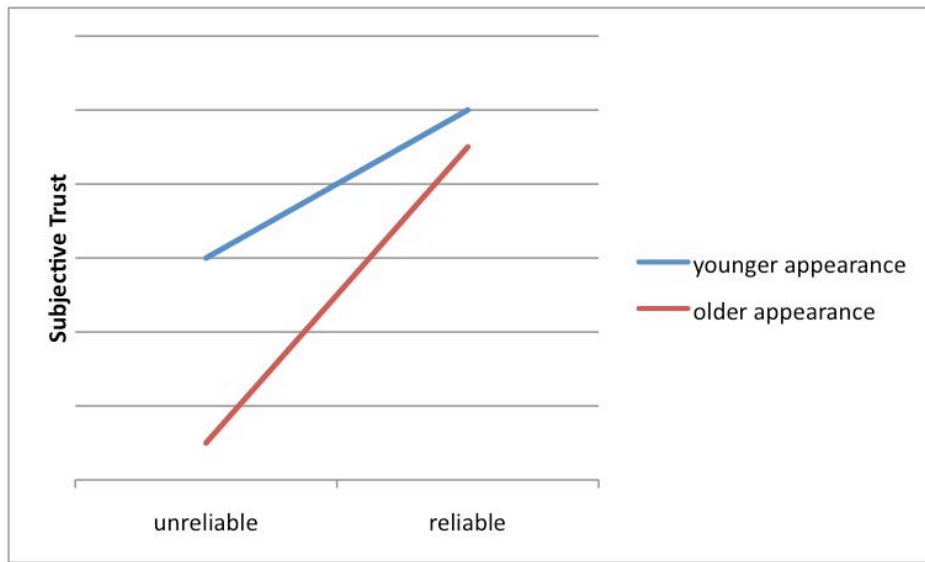


Figure 5: Reliability X robot age interaction on subjective trust

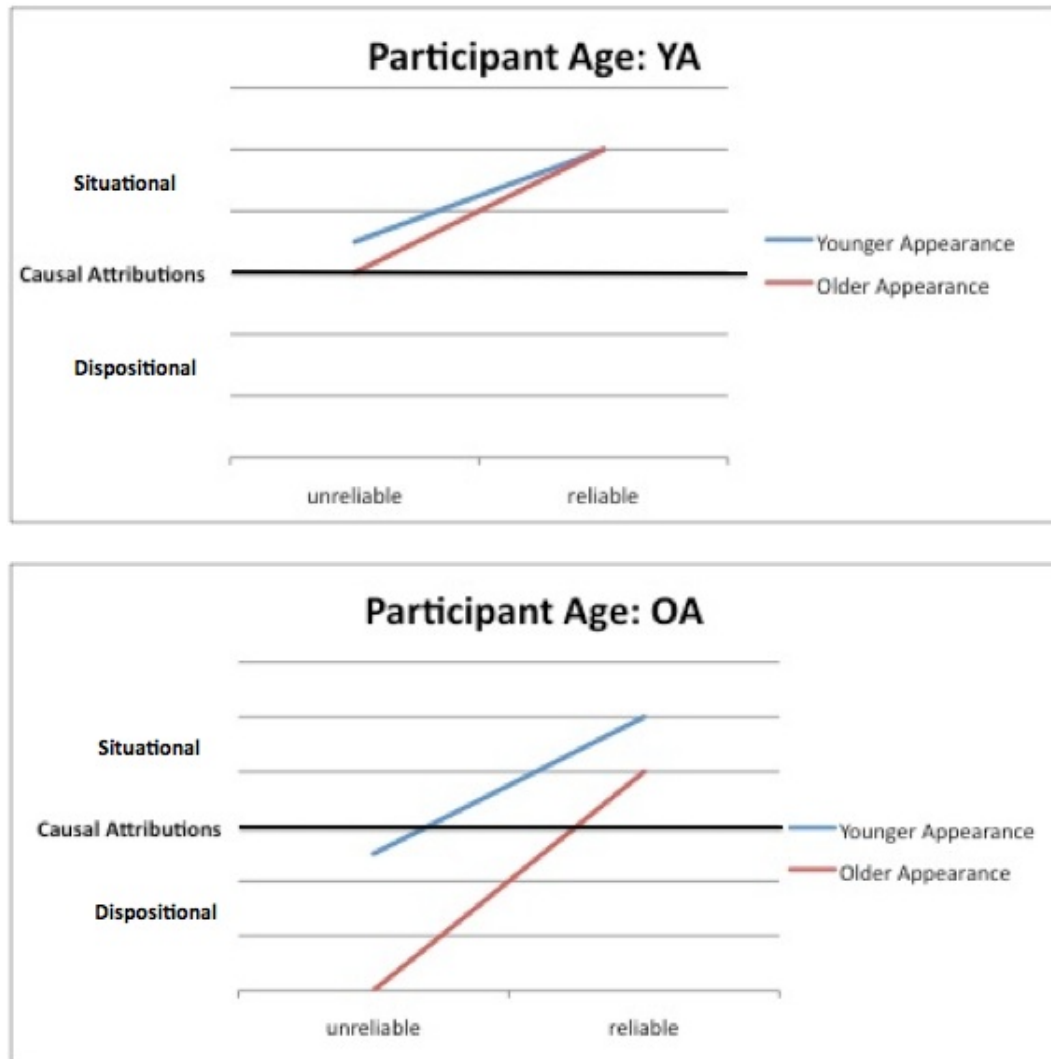


Figure 6: Participant age X robot appearance X reliability on causal attributions

APPENDIX 9: Leidheiser, W. (in progress). The Effects of Age and Working Memory Demands on Automation-Induced Complacency.

The Effects of Age and Working Memory Demands on Automation-Induced Complacency

William Leidheiser

Clemson University

Committee Members:

Dr. Richard Pak (Chair)

Dr. Kelly Caine

Dr. Patrick Rosopa

Abstract

Complacency refers to a type of automation use expressed as insufficient monitoring and verification of automated functions. Previous studies have attempted to identify the age-related factors that influence complacency during interaction with automation. However, little is known about the role of age-related differences in working memory capacity and its connection to complacent behaviors. The current study aims to examine whether working memory demand of an automated task and age-related differences in cognitive ability influence complacency. Higher degrees of automation (DOA) have been shown to reduce cognitive workload and may be used to manipulate working memory demand of a task. Thus, we hypothesize that a lower DOA (i.e. information acquisition stage with lower level) will demand more working memory than a higher DOA (i.e. decision selection stage with higher level). Older adults are expected to have a greater tendency to become complacent under a low DOA and younger adults are expected to have a greater tendency to become complacent under a high DOA.

Introduction

By the year 2050, the number of older adults (age 65 and over) in the world is estimated to reach approximately 1.5 billion (WHO, 2011). A host of automated services and devices are or will be designed to help older adults maintain independence (e.g., medication reminder apps). Despite this availability of automation and its seemingly utility to maintain independent living (Haigh & Yanco, 2002), research has shown that older adults may be more complacent with automated systems compared to younger age groups (so called automation-induced complacency).

Automation-induced complacency is the “self-satisfaction that may result in non-vigilance based on an unjustified assumption of satisfactory system state” (Billings, Lauber, Funkhouser, Lyman, & Huff, 1976). It is the state in which a user fails to notice imperfect automation. When the user poorly monitors the system and does not detect a fault, performance consequences can result (Parasuraman & Manzey, 2010). For example, an older adult with diabetes may monitor their blood glucose levels with an automated tool. If the older adult perceives the device as reliable and trusts that the blood glucose readings are accurate, they may rely on the reading even when the device starts to falter. As older adults begin to adopt automated technologies, it is important to understand the age-related factors that contribute to increased complacency and the performance costs associated with those behaviors.

Older Adults, Working Memory, and Complacency

Older adults have been found to be more complacent with automation relative to younger adults (Ho et al., 2005b). Various studies have suggested several possible explanations for older adults increased complacency. Some person-related variables range from issues such as higher levels of trust (Johnson, Sanchez, Fisk, & Rogers, 2004; Pak, Fink, Price, Bass, & Sturre, 2012),

or age-related differences in abilities (e.g., working memory; Ho et al., 2005b) while some system-related variables are reliability of the automation (Sanchez, Fisk, & Rogers, 2004; Mayer, 2008; Olson, Fisk, & Rogers, 2009), cost of error (Ezer, 2006; Ezer, Fisk, & Rogers, 2008), cost of verification (Ezer, Fisk, & Rogers, 2007; Ezer et al., 2008), expectations of system performance (Mayer, 2008), and workload (McBride, Rogers, & Fisk, 2011).

Research investigating age differences in cognitive ability as a possible explanation for changes in complacency found that in an automated task, older adults relied more on the automation, committed more errors, had greater trust in the system, and were less confident in their own abilities compared to younger adults (Ho et al., 2005b). Also, the task exerted high demand on participants' working memory, which is defined as the amount of information that can be held in the mind or kept accessible at one time (Cowan, 2004). At the conclusion of each study session, Ho et al. (2005b) had participants recite information from the task and found that greater recall accuracy was correlated with fewer automation-related errors. Based on their findings, they concluded that age-related differences in working memory might be a potential reason for age differences in complacency due to the memory dependent automated task. The researchers proposed that because the younger adults could actively store and recall task information when needed, they could more easily identify an automation failure compared to their older counterparts.

Researchers theorized there are two main factors that contribute to older adults' complacent behavior with automated technologies (Ho, Kiff, Plocher, & Haigh, 2005a). The first is that while using automation, older adults form an inaccurate mental representation of the correct values used in the decision making process due to reduced working memory capacity. Working memory has been found to be a critical determinant in mental model acquisition

(Gilbert & Rogers, 1999), where having an accurate mental model of the automation allows an individual to better understand the behavior of the system. When older adults acquire an inaccurate mental representation of the automation, they should fail to anticipate and notice the presence of system failures. The second is that due to their reduced working memory capacity, older adults are unable to judge the accuracy of automation (Ho et al., 2005a). Diminished working memory may prevent users from keeping track of an accurate summation of automation failures. If lower working memory of older adults inhibits detection of automation failures or active recall of previously encountered failures, the user will have a distorted view of system reliability. When older adults perceive automation as more reliable than it is, they should rely more and verify less (i.e. increased complacency).

In both cases, it is assumed older adults relative complacency with automation is due to a mismatch between the working memory demands of the task and working memory capacity of the person (Ho et al., 2005a). If working memory capacity plays such a central role in automation complacency, we should observe the opposite relationship as well: reduced complacency in older adults when the automation has been designed to demand relatively less working memory resources (or working memory resources are less constrained). The design of Ho et al.'s (2005b) study precludes this determination because it is unclear whether the high working memory demands of the task or the degree of automation (DOA) contributed to the difference in complacency.

In sum, several lines of research seem to point to the importance of individual and age-related differences in working memory on automation behavior, particularly complacency. The research shows that older adults are less sensitive to automation failures (McCarley, Wiegmann, Wickens, & Smith, 2002) and frequently rely on the automated system when these malfunctions

occur (Ho et al., 2005b). Older adults have greater trust in automation, even when the system is faulty to varying degrees (Mayer, 2008). They have lower working memory capacity, which decreases the ability to retain knowledge about previous automation failures and overall system reliability. When working memory demands are high (or working memory capacity is constrained), complacency seems to increase.

How Complacency is Influenced by Automation-Related Factors

Reliability

Automation reliability is the overall accuracy of the system and is an important factor of automation-induced complacency because the number of errors it produces can impact dependence on automation.

Across different levels of reliability, age is known to produce increased effects on trust in automation. For instance, several studies found that higher reliability led to higher subjective trust in the system for both age groups, but older adults had significantly higher trust than younger adults (Sanchez et al., 2004; Ho et al., 2005b). Highly reliable automation is problematic because users can become accustomed to its high level of performance and may not expect it to fail.

Research on age differences in automation use has found that older adults tend to overestimate the actual automation reliability (Olson et al., 2009). With known differences in working memory, older adults have difficulty detecting errors and perceiving overall automation performance. A combination of unnecessarily high trust in the system and a lack of working memory may produce a lack of error prone awareness consistent with complacent behavior.

Workload

The workload or demand of a task can be taxing on an individual's cognitive resources, especially when a task is performed over a long period of time. Greater complacency has been shown in a multitask environment instead of a single task or monitoring role for younger adults (Parasuraman, Molloy, & Singh, 1993). Increased task demands can burden the use of cognitive resources and can limit the ability to maintain optimal manual performance. In order to alleviate cognitive workload, the user can increase dependence on automation. If the individual has access to greater cognitive resources, they may be able to limit their dependence on automation. Since older adults have limited cognitive resources, the effect of task demand on complacency should become greater as individuals age.

Under taxing conditions, older adults have a greater tendency to monitor automation, yet fail to correctly identify automation errors (Ho et al., 2005b). Exerting more cognitive resources to complete a task may lead the user to rely on automation after task demands become too difficult to manage. There are age differences in complacency that have occurred under high workload conditions, where older adults display greater complacency than younger adults (McBride, 2010; Ho et al., 2005b). If workload only partially contributes to increases in complacency, other age-related factors must be involved as well.

Working memory capacity has been found to significantly predict younger adult performance in an automated task with varying workload (de Visser, Shaw, Mohamed-Ameen, & Parasuraman, 2010). Since working memory plays a role in predicting performance, this cognitive ability may explain some age-related differences in complacent behaviors.

Degree of Automation

Automation comes in a variety of forms, which can execute different functions for the user based on their capabilities and limitations. However, automation is not simply an all or none

concept because any individual task can feature varying degrees of automation (DOA) that take into account the use of stages and levels (Wickens, Li, Santamaria, Sebok, & Sarter, 2010).

Parasuraman, Sheridan, and Wickens (2000) identified several stages of automation that are based on an existing model of human information processing: information acquisition (stage 1), information analysis (stage 2), decision and action selection (stage 3), and action implementation (stage 4). Each stage is designed to support a different aspect of the cognitive process. For example, an individual with an unknown illness may input their symptoms into automated decision support tool to obtain a diagnosis. With a lower stage of automation, all possible illnesses related to those symptoms would be provided and the user would make a decision based on all the options listed. On the other hand, a higher stage of automation would have the decision support tool provide the user with one or several optimal choices in order to make the selection process more efficient.

Levels of automation differ from stages because they affect the role of humans and automated systems in a given task. These levels exist on a spectrum of automation, where each level between manual and fully automated changes the designation of authority for decision-making tasks. A low level of automation grants authority to the human, making the person primarily responsible for performing the task. In this case, the individual with the decision support tool would be given little to no guidance and would have to choose the best option based on the information provided. The roles are reversed under a high level of automation, where the automation has more authority to make decisions for the user and complete the task. For instance, the decision support tool might take the symptoms entered by the user and present them an ideal diagnosis.

Along each stage of automation, varying levels can be applied to achieve a lower or higher DOA. More automation or a greater DOA can be achieved with both higher levels within a stage and later stages (Manzey, Reichenbach, & Onnasch, 2012). Also, higher DOAs are associated with greater performance in addition to diminished workload (Wickens et al., 2010). Since workload is reduced under a higher DOA, the automation is taking on more of the cognitive demand for those tasks than the user. This leaves the user with more cognitive resources at higher DOAs. Thus, working memory demands should lessen as the user moves from a lower DOA towards a higher DOA.

Higher complacency can take the form of performance detriments under unreliable systems and performance gains for increasingly reliable automation. For instance, a meta-analysis found that higher DOAs lead to greater accuracy for younger adults, but only when the automation performed optimally (Onnasch, Wickens, & Manzey, 2013). However, there was a greater performance cost for imperfect automation as DOA increased. For younger adults, these findings reveal differences in performance across DOAs, which seem to indicate changes in complacent behavior. In this context of comparing performance across lower and higher DOAs, research on the older adult population has not been performed. In terms of research by Ho et al. (2005b), it is still unclear whether the high working memory demands of the task or the high DOA contributed to age-related differences in complacency.

Current Study

The current study will further examine the role of age-related differences in working memory and automation-induced complacency. If complacency is related to working memory, then altering the working memory demands of the task (or varying the person's working memory capacity) should affect overall dependence on automation. Fortunately, the working memory

demands of automation are related to how much information in the automated task is presented to the user (i.e. stage of automation) and the amount of authority allocated to the human or automation within the task (i.e. level of automation) (Parasuraman, Sheridan, & Wickens, 2000; Sheridan & Verplank, 1978). We can alter the working memory demands of the task by altering the DOA presented to the user. Thus, we should expect to observe greater age related differences in complacency at degrees that increase working memory demands for the user. Ho, Wheatley, and Scialfa (2005b) only used a high DOA (with concomitantly high working memory demands) to examine differences in complacency between younger and older adults. Therefore, we will use two DOAs that vary in working memory demand in order to investigate the effects of lower and higher based DOAs on complacency. Also, we will examine the predictive ability of working memory capacity at each DOA. We expect that working memory capacity of each age group will be relative to the working memory demand of the task. Thus, we anticipate working memory capacity to be more predictive of performance for younger adults at a low DOA and for older adults at a high DOA.

This study will utilize a low-fidelity targeting simulation, which has been used in prior research to analyze accuracy and speed of user selections during interaction with DOAs support (Rovira, McGarry, & Parasuraman, 2007). Since higher DOAs have been linked with reduced cognitive workload (Onnasch et al., 2013), we expect participants to perform better under higher DOA (i.e. decision selection stage with higher level) than lower DOA (i.e. information acquisition stage with lower level). Based on existing literature, we anticipate a main effect of age group on task accuracy and completion time, where younger adults should outperform older adults. We can infer the extent to which participants are complacent by analyzing their pattern of performance at different reliability levels. A greater difference between performance with

unreliable and reliable automation indicates higher complacency because the user is relying heavily on the system without monitoring for failures. Therefore, we will examine task accuracy for unreliable and reliable trials across DOAs and age groups. We hypothesize a lower DOA will result in a greater complacency for older adults and a higher DOA will result in greater complacency for younger adults. We anticipate this result because the high demand of a low DOA should limit older adults' ability to verify information provided by the automated system. In terms of the high DOA, lower task demands should lull younger adults into depending on the system instead of checking for errors.

Method

Participants

Thirty-six undergraduate students will be recruited for this research and given course credit for participation. Thirty-six older adults (ages 65-75) from the local area will be recruited and will be compensated \$25 for their time.

Task

The tasks for this study will be adapted from prior research that uses an automated system in the context of a low-fidelity UAV simulation (Rovira et al., 2007). The primary task for this study will be to quickly and accurately find the closest combination of friendly (green units) and enemy units (red units) in terms of distance apart on the grid (Figure 1). Automation will be presented as a table, which will display the distances and unit combinations needed by participants to complete the primary task. The secondary task will consist of checking for a specific call sign and clicking a corresponding button when it appears on screen. The call sign is comprised of a single word and number combination (e.g. Hunter-6). The program will randomly

alternate between 14 different call signs every 5 seconds as the participant completes the primary task.

Participants will complete blocks of trials, where each block will consist of a different DOA and workload level (Appendix A). The DOA manipulation will change the stage and level of the automation table used in the task. The lower DOA will use the information acquisition stage, which presents all possible friendly and enemy unit combinations in each grid, with a low level of automation that does not sort the information in any meaningful way. The higher DOA will use the decision and action selection stage, which will present the top 3 friendly and enemy unit combinations. In addition, this DOA will feature a high level of automation that will sort the information based on importance, so that the shortest distance combination is presented at the top. The workload manipulation will change the number of units presented in the grid. Low workload will present 3 friendly and 3 enemy units, while high workload will show 6 friendly and 6 enemy units. Each combination of DOA and workload will be presented twice for a total of 8 blocks and 240 trials. Participants will complete the DOA and workload manipulation pairings in a random counterbalanced order.

The overall automation reliability will be set at 80%, which is above the threshold for imperfect reliability acceptance (Wickens & Dixon, 2007). In each block of 30 trials, 24 trials will be reliable and the remaining 6 trials will be unreliable. An unreliable trial will contain inflated distance values between units or incorrect optimal suggestions within the automation support table. The first automation failure will not occur until the 10th trial, so that users can rebuild trust after each block. Also, subsequent automation failures will be distributed randomly throughout the remaining trials.

Measures

Cognitive Abilities. The following abilities will be assessed: perceptual speed (digit-symbol substitution; Wechsler, 1997), vocabulary (Shipley vocabulary; Shipley, 1986), and working memory (automated operation span (Aospan); Unsworth, Heitz, Schrock, & Engle, 2005). Instructions for the Aospan task can be found in Appendix B. These measures were chosen because they are reliable indicators of their respective abilities (e.g., Czaja et al., 2006). The cognitive ability measures were selected to confirm age differences in fluid and crystallized intelligence. Specifically, the Aospan will be used to detect age group differences and test the predictive ability of working memory capacity on performance at two DOAs. Research has shown the Aospan to be a reliable and valid indicator of working memory capacity (Unsworth et al., 2005). This version of the Ospan is preferred because the task is fully computerized, the participant can complete the task independently of the experimenter, and the experimenter can collect data from several participants simultaneously. In the Aospan task, participants will be instructed to complete simple math problems while remembering the order of individual letters that will be presented after solving each problem. Participants will need to correctly answer at least 85% of the math problems and recall as many letters as possible. The Aospan score will consist of the sum of all perfectly recalled letter sets, where higher scores indicate greater working memory capacity.

Subjective Workload. Subjective workload will be measured with the NASA-Task Load Index (NASA-TLX) (Prichard, Bizo, & Stratford, 2011). A computer version of the task will present 6 items that constitute overall workload: mental demand, physical demand, temporal demand, performance, effort and frustration. Each item is rated on a Likert scale of 0 to 20, where higher values indicate increased workload. Subjective workload will be calculated as the

average of the 6 combined items. The NASA-TLX was chosen as a manipulation check for automation stage and age differences in perceived workload.

General Trust in Automation. Trust towards everyday automation will be measured with a survey developed by Jian, Bisantz, and Drury (2000) (Appendix C). This measure is a 12-item survey that is rated on a Likert scale of 1 (not at all) to 7 (extremely). The first 5 questions are negatively framed and the last 7 are positively framed. Trust is the sum of normal and reverse coded responses, for a possible total score of 84. Higher scores on this measure indicate greater trust in the automated system. The measure will be analyzed for age-related differences in trust towards automation.

Subjective Trust. We will use a survey adapted from Lee and Moray (1992) to measure subjective trust specifically towards each DOA and working memory manipulation (Appendix D). This trust measure will pose 4 questions, rated from 0 (not at all) to 100 (extremely), about the automated aid used in each set of trials. For example, the questions will ask participants to answer how much they trusted, relied upon, or benefited from using the automated aid. The overall score will consist of the sum of average scores on questions 1, 2, and 4, where higher scores will indicate higher trust. Additionally, this questionnaire will be used to examine trust differences between age groups, workload, and DOA.

Complacency Potential. The Complacency Potential Rating Scale (CPRS) measures individual potential complacency behavior (Singh, Molloy, & Parasuraman, 1993) (Appendix E). This 20-item scale contains 4 filler items and is rated on a Likert scale of 1 (strongly disagree) to 5 (strongly agree). The CPRS score is a sum of the remaining responses, where higher values on this measure indicate an increased complacency potential. The CPRS was selected in order to

predict participant complacency within the task. Also, the measure serves to verify age differences in complacency potential.

Design

The current study is a 2 (age group: young or old) x 2 (DOA: low or high) x 2 (automation reliability: unreliable or reliable) x 2 (workload: low or high) mixed-subjects design. Age group will be a between-subjects independent variable. These groups will differ in working memory capacity because older adults have been shown to have less of this ability than younger adults. DOA, automation reliability, and workload will be within-subjects independent variables. The DOAs serve as our working memory demand manipulation.

The dependent variables will be targeting task accuracy, targeting task completion time, complacency potential, subjective trust, subjective workload, general trust in automation, and working memory capacity. *Targeting task accuracy* will be measured by the mean rate of optimal responses for each automation block. An optimal response is the identification of the closest pair of friendly and enemy units on the targeting task grid. *Targeting task time* will be measured by the average duration (in milliseconds) it takes participants to complete each trial. *Complacency potential* will be comprised of scores on the CPRS. *Subjective trust* will be measured by the sum of subjective ratings on the trust questionnaire for each combination of DOA and workload level. *Subjective workload* will consist of an average of the 6 items on the NASA-TLX and will be measured for each combination of DOA and workload level. *General trust in automation* will be measured with the corresponding scale based on ratings of trust towards everyday automated technologies. *Working memory capacity* will be measured as the sum of perfectly recalled sets of letters on the Aospan task.

Procedure

Participants will be seated at individual PC-computers and provided with informed consent. They will be instructed to complete the demographics form and the cognitive ability measures. The experimenter will then tell participants to open and observe the targeting task instructions screen. Participants will be told the following: “In this experiment, you will have two tasks. The first task will be to monitor the communications panel for the call sign Hunter-6. When you see Hunter-6, you should click the answer button. The second task will be to target enemy units with the closest friendly unit as quickly as you can. You will do this by first selecting a friendly unit from the list of buttons in the targeting input and then select an enemy target from the list of buttons and click ok. The computer aid will sometimes help you with this task by showing you the distances between friendly and enemy units. Sometimes, two sets of targets will have the same distance. In this case, you will pick the one with the shortest distance to the headquarters. Sometimes the computer aid will give you lots of information, other times it will give you much less information. The computer aid can be very reliable but it is not perfect all the time.” After these instructions, the experimenter will answer questions before the participants begin the task.

As the participants complete the tasks, the units in the grid and the values within automation table will change for each subsequent trial. Between each block of trials, participants will fill out the NASA-TLX and a brief subjective trust measure. During the experiment, a screen will appear to indicate when participants linger too long on a particular trial. If participants do not input friendly and enemy unit combinations within the set time limit, the program will automatically continue to the next trial. Younger adults will have 10 seconds to complete each trial, while older adults will have 20 seconds. Older adults will have more time for the task because of normative age-related differences in psychomotor speed (Salthouse, 1985). Time

limits were based on an analysis of incomplete trials from pilot testing the task with each age group.

Participants will proceed through each block of trials and the computer will notify them when they are finished. When they complete the automation program, participants will be presented with a general subjective measure of trust in automation and the CPRS. At the conclusion of the experiment, participants will be debriefed and provided compensation for their time.

Expected Results

To begin the analysis, outliers will be eliminated from the data. An outlier will be defined as a participant that scored greater or less than 3 standard deviations from the mean on a particular measure. In order to examine the differences in working memory demands for each DOA, we will perform regressions of working memory capacity on targeting task accuracy. Since working memory capacity has already been found to predict younger adult performance while using automation (de Visser et al., 2010), we will examine the slopes of younger and older adults at each DOA. We expect working memory capacity to be more predictive of task accuracy for younger adults at a low DOA (Figures 2-3). This result is anticipated because lower DOAs have been associated with greater cognitive workload (Wickens et al., 2010). In terms of a high DOA, we expect working memory capacity to be more predictive of task accuracy for older adults.

We will further investigate the effect of our manipulations on performance by conducting a 2 (age: young or old) x 2 (DOA: low or high) x 2 (workload: low or high) repeated measures analysis of variance (ANOVA) for targeting task accuracy and task time. We expect a main effect of age such that younger adults will perform the task quicker and more accurately than

older adults. We expect a main effect of DOA such that performance with the high DOA will be significantly greater than the low DOA. We anticipate a main effect of workload, where performance under low workload will be significantly greater than high workload. Graphical representations of these main effects can be found in Figure 4 and Figure 5.

In order to examine differences in complacent behavior, we will perform a 2 (age: young or old) x 2 (DOA: low or high) x 2 (automation reliability: reliable or unreliable automation) repeated measures ANOVA for targeting task accuracy. We can infer the extent to which participants are complacent by analyzing their pattern of performance at different reliability levels. A greater difference between performance with unreliable and reliable automation indicates higher complacency because the user is relying heavily on the system without monitoring for failures. From the analysis, we anticipate a 3-way interaction such that the interaction between age and DOA will change as a function of reliability (see Figure 6 and Figure 7).

We will analyze the scores on each subjective measure used in the study. We will perform a 2 (age: young or old) x 2 (DOA: low or high) repeated measures ANOVAs to analyze differences in subjective trust and workload. We expect a main effect of age, where older adults will report greater workload and trust than younger adults. We expect a main effect of DOA such that the higher DOA will produce greater subjective trust and diminished workload. Graphical representations of these main effects can be found in Figure 8 and Figure 9. Additional measures including complacency potential and general trust in automation will be analyzed with independent samples t-tests to compare scores across age groups. We expect that older adults will have greater complacency potential and greater overall trust in automation than younger adults.

Discussion

It is important to understand the factors that contribute to complacent behaviors within the human-automation interaction. For the design of automated systems, it is necessary to consider factors such as reliability and workload. Since high system reliability is common in most automated technologies today and thus makes users more susceptible to complacent behaviors, it is essential to alert the user to potential automation-related failures that can occur. In terms of task demands, keeping the task manageable for the user is critical for detecting and correcting inaccuracies.

Designers should select the appropriate DOA for the known population of users. Specifically, the design of automated tasks should consider the age of the user. Automation can be presented in many different ways and can perform a wide range of tasks for the user. Depending on the type of task, some forms may demand more working memory than others. Limiting working memory demand through automation can be beneficial to both younger and older adults. This may help to reduce the occurrence of complacent behaviors during interaction with automation.

References

- Billings, C. E., Lauber, J. K., Funkhouser, H., Lyman, G., & Huff, E. M. (1976). *Aviation Safety Reporting System* (Technical Report TM-X-3445). Moffett Field, CA: National Aeronautics and Space Administration Ames Research Center.
- Cowan, N. (2004). *Working memory capacity*. Psychology Press.
- Czaja, S. J., Charness, N., Fisk, A. D., Hertzog, C., Nair, S. N., Rogers, W. A., & Sharit, J. (2006). Factors predicting the use of technology: Findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). *Psychology and Aging, 21*(2), 333–352.
- de Visser, E., Shaw, T., Mohamed-Ameen, A., & Parasuraman, R. (2010). Modeling human-automation team performance in networked systems: Individual differences in working memory count. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 54*(14), 1087-1091.
- Ezer, N. (2006). Toward an understanding of optimal performance within a human-automation collaborative system: Effects of error and verification cost. Unpublished master's thesis, Georgia Institute of Technology, Atlanta, GA.
- Ezer, N., Fisk, A. D., & Rogers, W. A. (2007). Reliance on automation as a function of expectation of reliability, cost of verification, and age. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 51*(1), 6-10.
- Ezer, N., Fisk, A. D., & Rogers, W. A. (2008). Age-related differences in reliance behavior attributable to costs within a human-decision aid system. *Human Factors, 50*(6), 853-863.

- Gilbert, D. K., & Rogers, W. A. (1999). Age-related differences in the acquisition, utilization, and extension of a spatial mental model. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 54(4), 246-255.
- Haigh, K. Z., & Yanco, H. (2002). Automation as caregiver: A survey of issues and technologies. In *AAAI-02 Workshop on Automation as Caregiver: The Role of Intelligent Technology in Elder Care*, 39-53.
- Ho, G., Kiff, L. M., Plocher, T., & Haigh, K. Z. (2005a). A model of trust & reliance of automation technology for older users. In *AAAI-2005 Fall Symposium: "Caring Machines: AI in Eldercare"*, 45-50.
- Ho, G., Wheatley, D., & Scialfa, C. T. (2005b). Age differences in trust and reliance of a medication management system. *Interacting with Computers*, 17(6), 690-710.
- Jian, J., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1), 53-71.
- Johnson, J. D., Sanchez, J., Fisk, A. D., & Rogers, W. A. (2004). Type of automation failure: The effects on trust and reliance in automation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 48(18), 2163-2167.
- Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10), 1243-1270.
- Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, 6, 57-87.

- Mayer, A. K. (2008). The manipulation of user expectancies: Effects on reliance, compliance, and trust using an automated system. Unpublished master's thesis, Georgia Institute of Technology, Atlanta, GA.
- McBride, S. E. (2010). The effect of workload and age on compliance with and reliance on an automated system. Unpublished master's thesis, Georgia Institute of Technology, Atlanta, GA.
- McBride, S. E., Rogers, W. A., & Fisk, A. D. (2011). Understanding the effect of workload on automation use for younger and older adults. *Human Factors*, 53(6), 672-686.
- Olson, K. E., Fisk, A. D., & Rogers, W. A. (2009). Collaborative automated systems: Older adults' mental model acquisition and trust in automation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 53(22), 1704-1708.
- Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2013). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human Factors*, 56(3), 476-488.
- Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, 55(9), 1059-1072.
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381-410.
- Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced 'complacency'. *The International Journal of Aviation Psychology*, 3(1), 1-23.

- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 30(3), 286-297.
- Prichard, J. S., Bizo, L. A., & Stratford, R. J. (2011). Evaluating the effects of team-skills training on subjective workload. *Learning and Instruction*, 21(3), 429-440.
- Rovira, E., Cross, A., Leitch, E., & Bonaceto, C. (2014). Displaying contextual information reduces the costs of imperfect decision automation in rapid retasking of ISR assets. *Human Factors*, (ahead-of-print), 1-14.
- Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors*, 49(1), 76-87.
- Salthouse, T. (1985). Speed of behavior and its implications for cognition. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 400-426). New York: Van Nostrand Reinhold.
- Sanchez, J., Fisk, A. D., & Rogers, W. A. (2004). Reliability and age-related effects on trust and reliance of a decision support aid. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 48(3), 586-589.
- Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators* (Technical report). Cambridge, MA: MIT, Man Machine Systems Laboratory.
- Shipley, W. C. (1986). *Shipley Institute of Living Scale*. Los Angeles: Western Psychological Services.

- Singh, I. L., Molloy, R. & Parasuraman, R. (1993). Automation-induced “complacency”:
Development of a complacency-potential scale. *International Journal of Aviation
Psychology*, 3(2), 111-122.
- Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the
operation span task. *Behavior research methods*, 37(3), 498-505.
- Wechsler, D. (1997). Wechsler Memory Scale III. (3rd Ed.). San Antonio, TX: The
Psychological Corporation.
- Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A
synthesis of the literature. *Theoretical Issues in Ergonomics Science*, 8(3), 201-212.
- Wickens, C. D., Li, H., Santamaria, A., Sebok, A., & Sarter, N. B. (2010). Stages and levels of
automation: An integrated meta-analysis. In *Proceedings of the Human Factors and
Ergonomics Society Annual Meeting*, 54(4), 389-393).
- Wiegmann, D., McCarley, J. S., Kramer, A. F., & Wickens, C. D. (2006). Age and automation
interact to influence performance of a simulated luggage screening task. *Aviation, Space,
and Environmental Medicine*, 77(8), 825-831.
- World Health Organization, National Institute on Aging, National Institutes of Health, and U.S.
Department of Health and Human Services. (2011). *Global Health and Aging* (NIH
Publication No. 11-7737). Retrieved from
http://www.who.int/ageing/publications/global_health.pdf

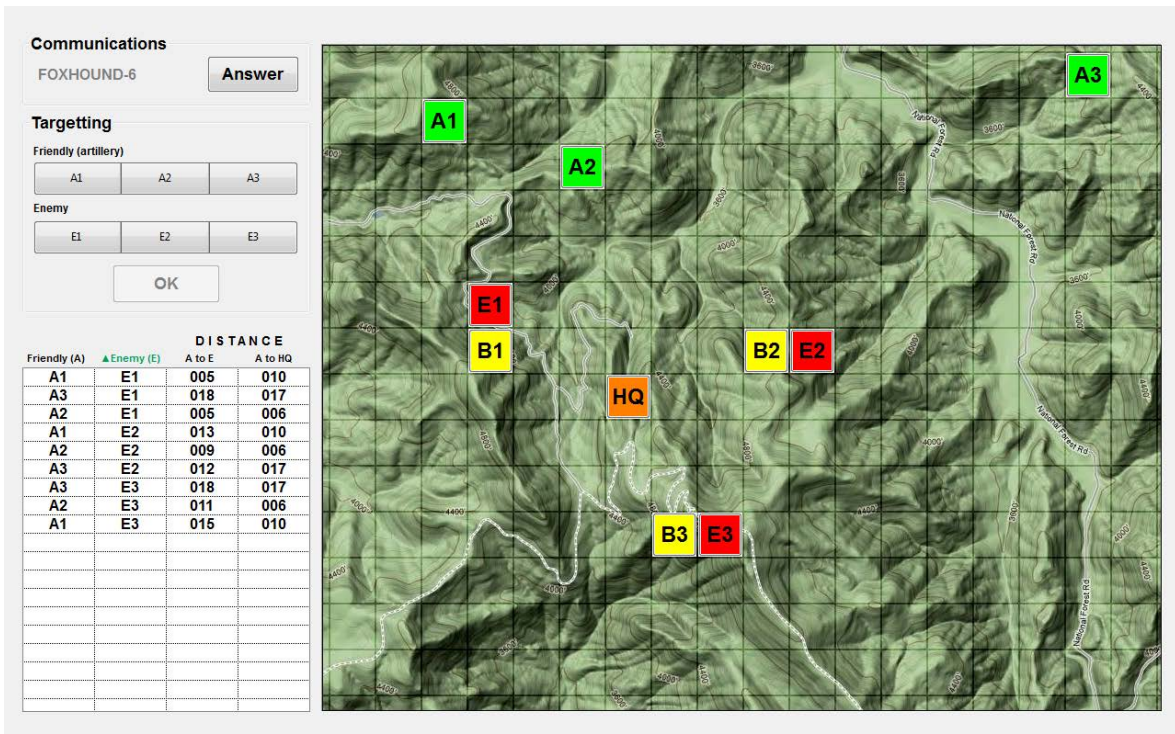


Figure 1. Screenshot of targeting task. Features communications panel (top-left), targeting input panel (top-left), automation table (bottom-left), and grid (right).

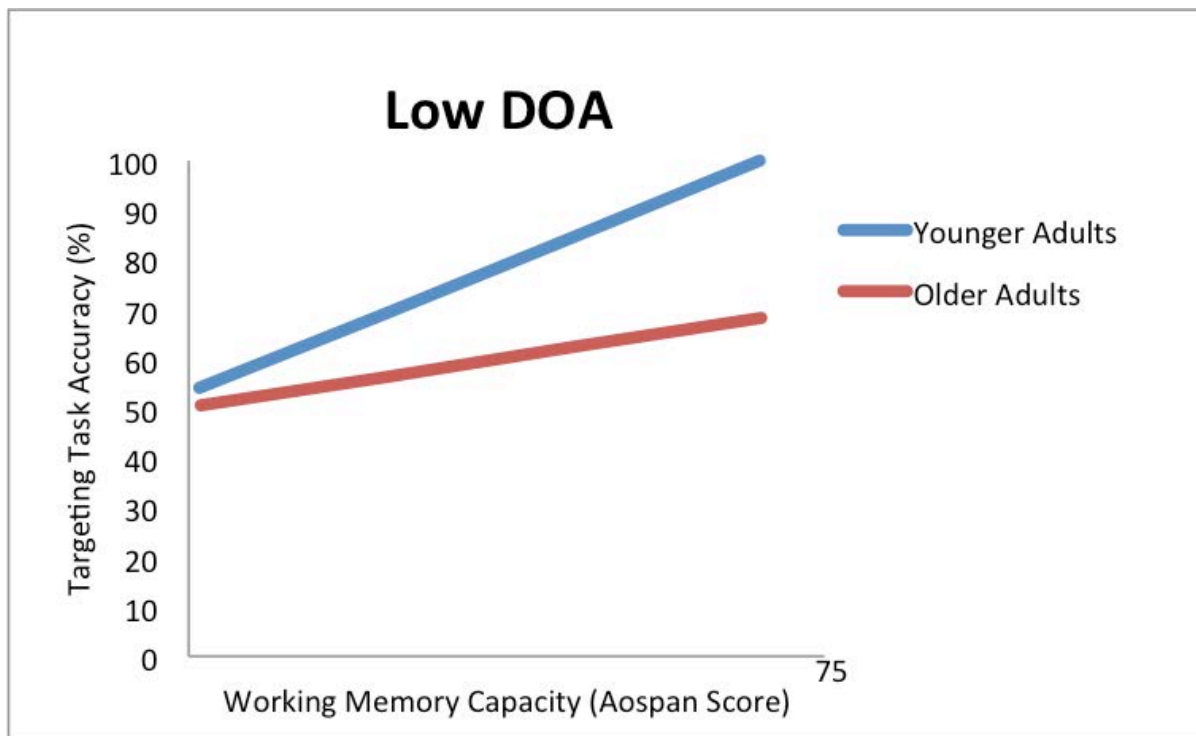


Figure 2. Linear regression between working memory capacity and targeting task accuracy (low DOA).

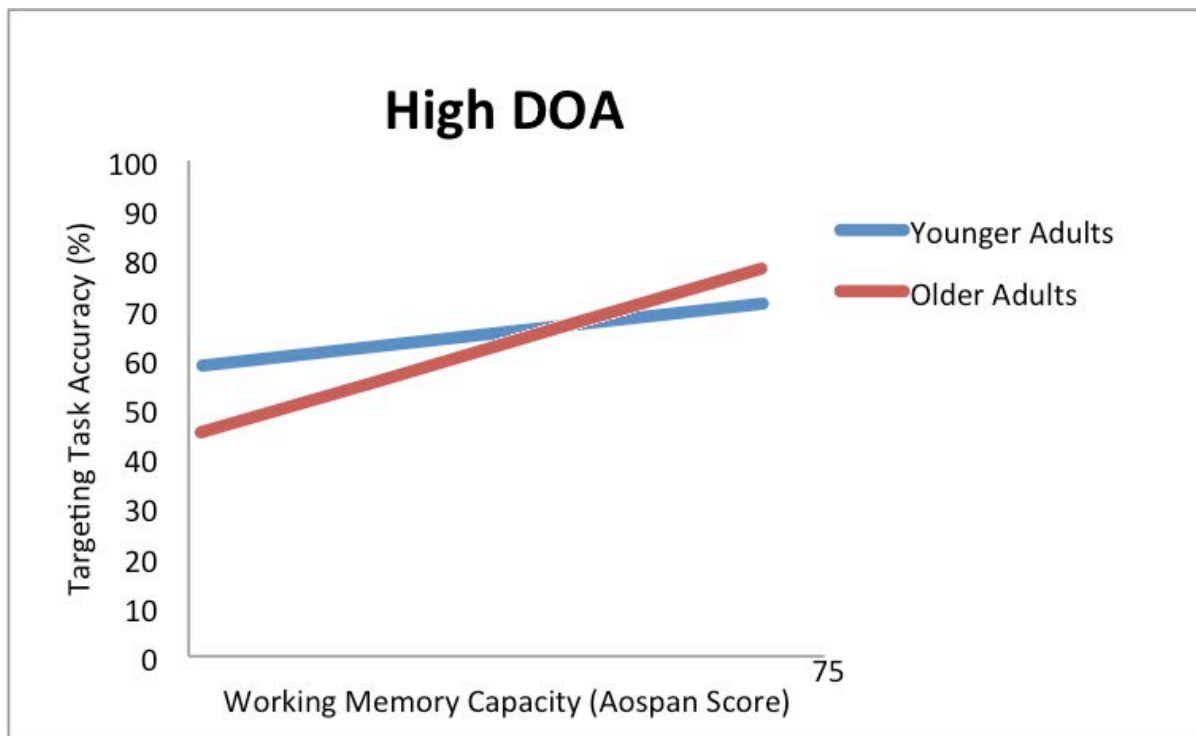


Figure 3. Linear regression between working memory capacity and targeting task accuracy (high DOA).

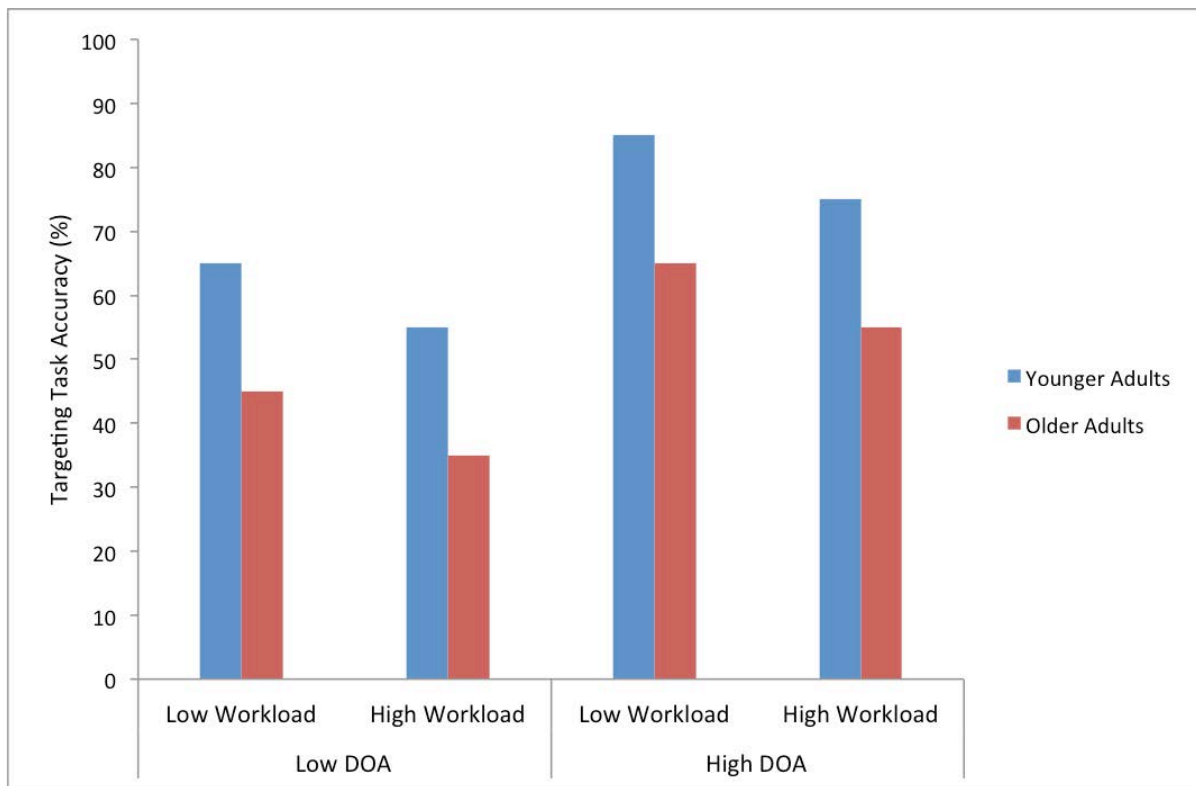


Figure 4. Graph of targeting task accuracy for each age group, DOA, and level of workload.

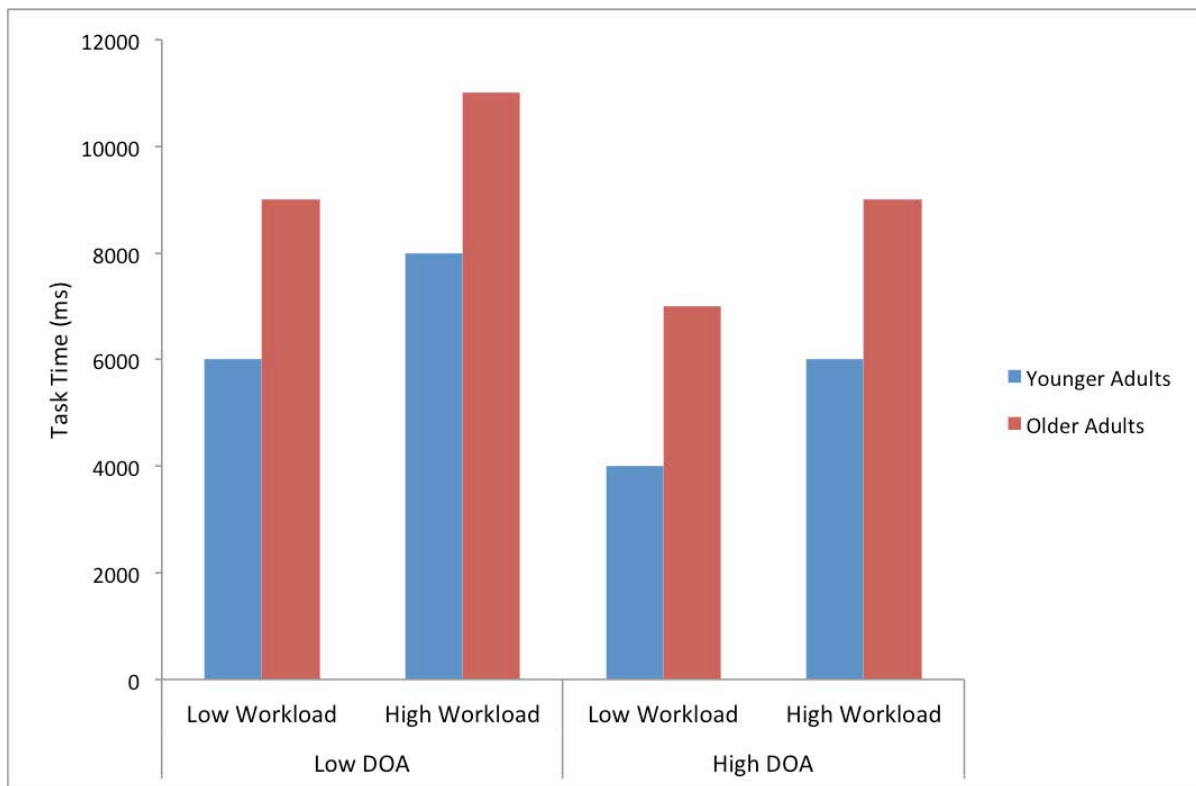


Figure 5. Graph of targeting task time for each age group, DOA, and level of workload.

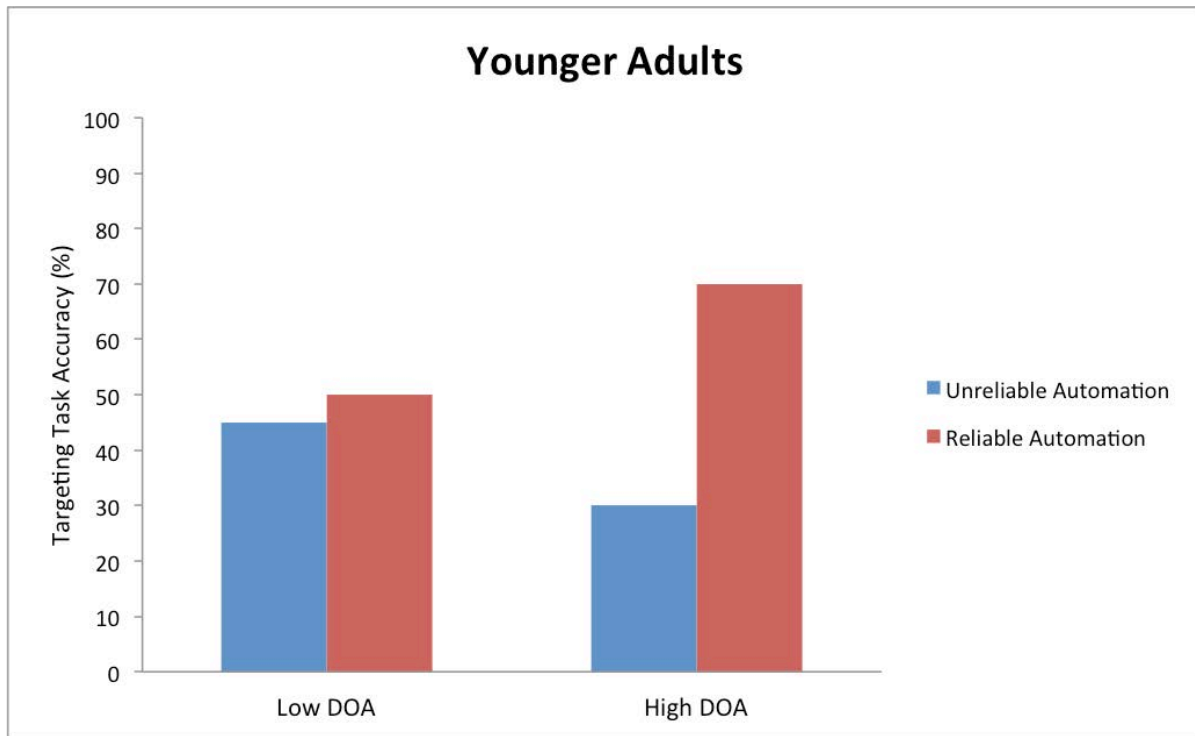


Figure 6. Graph of targeting task accuracy for younger adult participants.

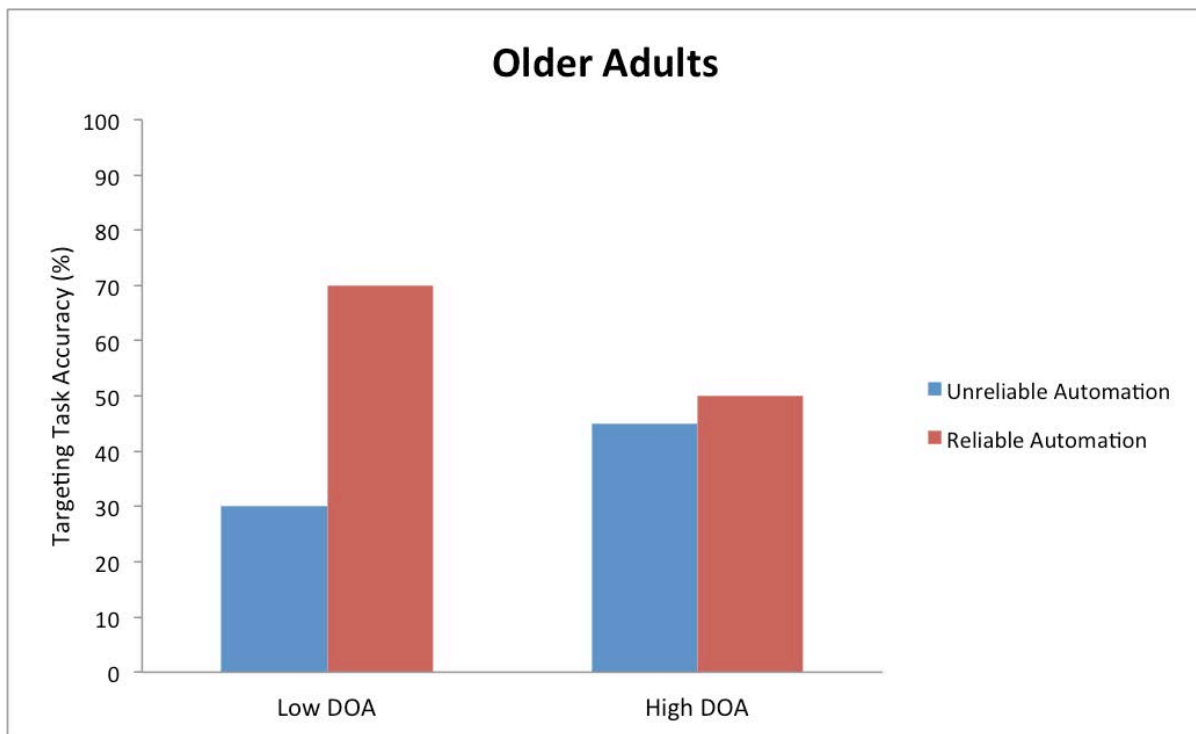


Figure 7. Graph of targeting task accuracy for older adult participants.

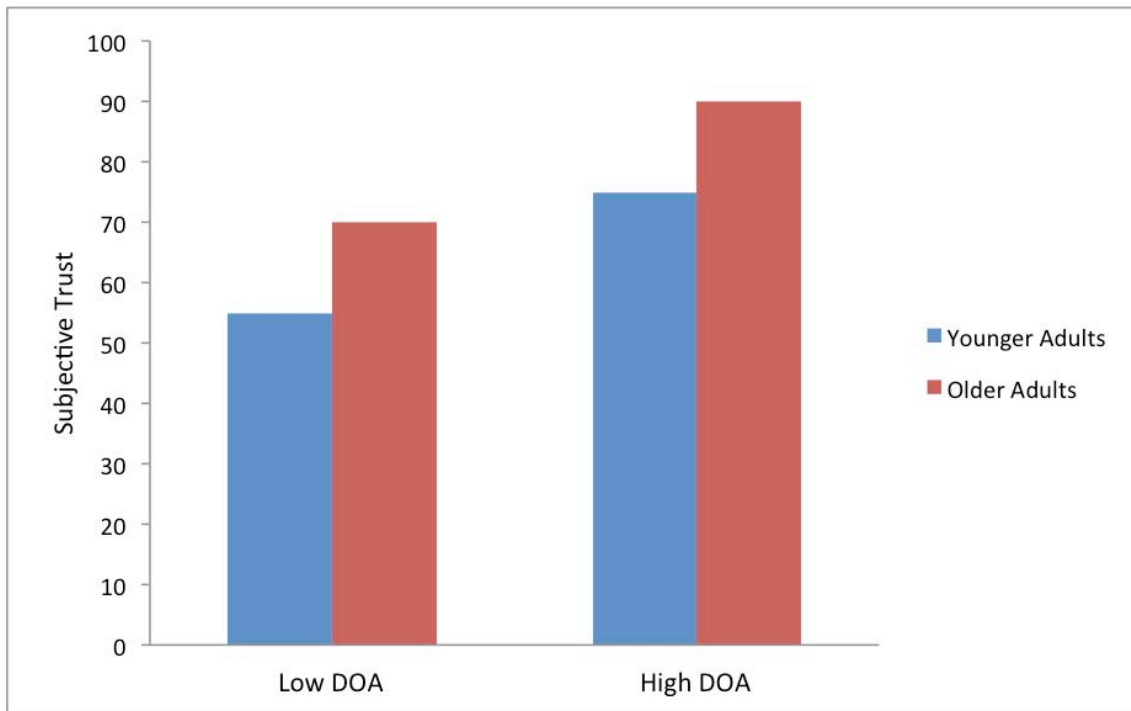


Figure 8. Graph of subjective trust reported at each DOA.

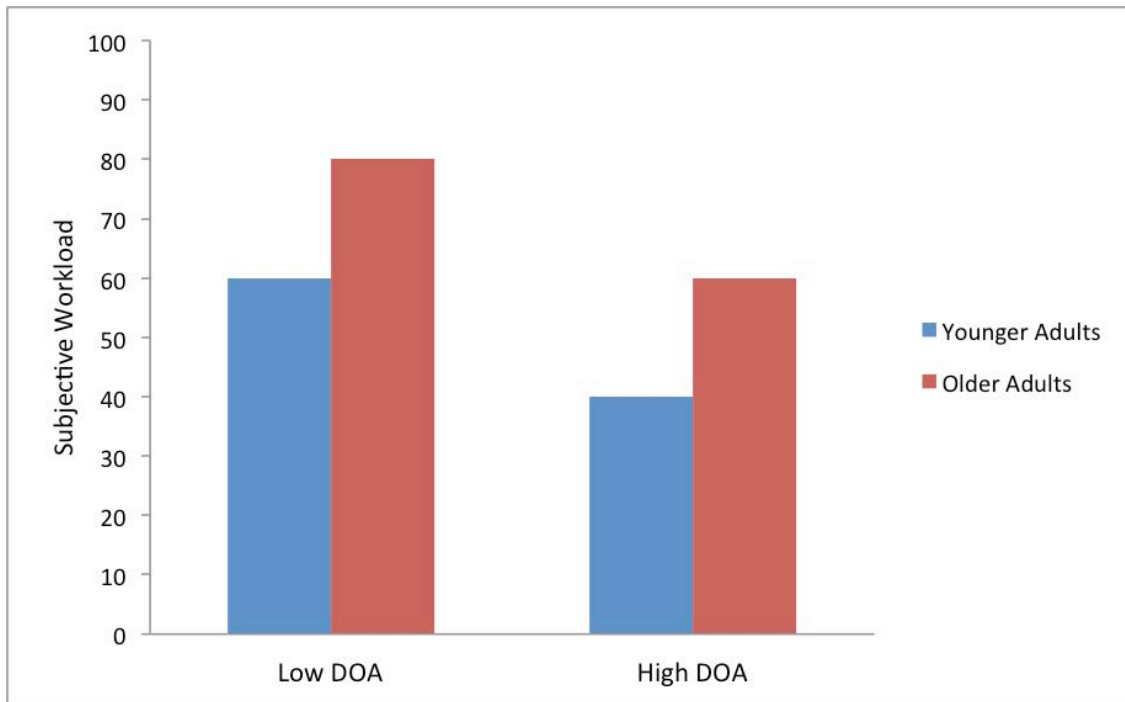


Figure 9. Graph of subjective workload reported at each DOA.

Appendix A: Examples of DOA and Workload Manipulations

Reliable, low DOA, and low workload trial example:

Communications
FOXHOUND-6 Answer

Targeting
Friendly (artillery)
A1 A2 A3

Enemy
E1 E2 E3

OK

Friendly (A)	Enemy (E)	DISTANCE	
		A to E	A to HQ
A1	E1	005	010
A3	E1	018	017
A2	E1	005	006
A1	E2	013	010
A2	E2	009	006
A3	E2	012	017
A3	E3	018	017
A2	E3	011	006
A1	E3	015	010

Reliable, low DOA, and high workload trial example:

Communications
SIGMA-6 Answer

Targeting
Friendly (artillery)
A1 A2 A3 A4 A5 A6

Enemy
E1 E2 E3 E4 E5 E6

OK

Friendly (A)	Enemy (E)	DISTANCE	
		A to E	A to HQ
A1	E1	010	004
A2	E1	004	010
A3	E1	007	009
A4	E1	010	014
A5	E1	010	008
A6	E1	013	003
A5	E2	005	008
A6	E2	008	003
A3	E2	010	009
A4	E2	015	014
A2	E2	003	010
A1	E2	005	004
A1	E3	005	004
A4	E3	007	014
A5	E3	013	008
A2	E3	007	010
A6	E3	004	003
A3	E3	004	009
A4	E4	012	014

Reliable, high DOA, and low workload trial example:

Communications
FREEDOM-6 Answer

Targeting

Friendly (artillery)

A1	A2	A3
----	----	----

Enemy

E1	E2	E3
----	----	----

OK

DISTANCE			
Friendly (A)	Enemy (E)	▲ A to E	A to HQ
A2	E3	003	005
A3	E2	003	009
A1	E3	004	008

Reliable, high DOA, and high workload trial example:

Communications
ACCORN-6 Answer

Targeting

Friendly (artillery)

A1	A2	A3	A4	A5	A6
----	----	----	----	----	----

Enemy

E1	E2	E3	E4	E5	E6
----	----	----	----	----	----

OK

DISTANCE			
Friendly (A)	Enemy (E)	▲ A to E	A to HQ
A6	E5	004	006
A1	E1	004	011
A3	E3	004	014

Appendix B: Automated Operation Span Task

Phase 1: Directions for Letter Memorization Practice Phase

- In this experiment, you will try to memorize letters you see on the screen while you also solve simple math problems.
- You will begin by practicing the letter part of the experiment.
- For the practice set, letters will appear on the screen one at a time. Try to remember each letter in the order presented.
- After 2-3 letters have been shown, you will see a screen listing 12 possible letters.
- Your job is to select each letter in the order presented. To do this, use the mouse to select each letter. The letters you select will appear at the top of the screen.
- When you have selected all of the letters, and they are in the correct order, hit the DONE box at the bottom right of the screen.
- If you make a mistake, hit the CLEAR button to start over.
- If you forget one of the letters, click the ? (question mark) button to mark the spot for the missing letter.
- Remember, it is very important to get the letters in the same order as you see them. If you forget one, use the ? button to mark the position.
- Do you have any questions so far? When you're ready, click the button below to start the letter practice.

Phase 2: Directions for Mental Math Practice Phase

- Now you will practice doing the math part of the experiment. A math problem will appear on the screen like this: $(2 * 1) + 1 = ?$

- As soon as you see the math problem, you should compute the correct answer. In the above problem, the answer 3 is correct.
- When you know the correct answer, you will click the OK button with your mouse.
- You will see a number displayed on the next screen, along with a button marked TRUE and a button marked FALSE.
- If the number on the screen is the correct answer to the math problem, click on the TRUE box with the mouse. If the number is not the correct answer, click on the FALSE box. For example, if you see the problem: $(2 * 2) + 1 = ?$ and the number on the following screen is 5 click the TRUE box, because the answer is correct. If you see the problem: $(2 * 2) + 1 = ?$ and the number on the next screen is 6 click the FALSE box, because the correct answer is 5, not 6. After you click on one of the boxes, the computer will tell you if you made the right choice,
- It is VERY important that you get the math problems correct.
- It is also important that you try and solve the problem as quickly as you can.
- Do you have any questions? When you're ready, click the mouse to try some practice problems.

Phase 3: Directions for Combined Letter Memorization and Mental Math Phase

- Now you will practice doing both parts of the experiment at the same time. In the next practice set, you will be given one of the math problems.
- Once you make your decision about the math problem, a letter will appear on the screen. Try and remember the letter.
- In the previous section where you only solved math problems, the computer computed your average time to solve the problems.

- If you take longer than your average time, the computer will automatically move you onto the next letter part, thus skipping the True or False part and will count that problem as a math error.
- Therefore, it is VERY important to solve the problems as quickly and as accurately as possible.
- After the letter goes away, another math problem will appear, and then another letter.
- At the end of each set of letters and math problems, a recall screen will appear. Use the mouse to select the letters you just saw.
- Try your best to get the letters in the correct order. It is important to work QUICKLY and ACCURATELY on the math. Make sure you know the answer to the math problem before clicking to the next screen.
- You will not be told if your answer to the math problem is correct. After the recall screen, you will be given feedback about your performance regarding both the number of letters recalled and the percent correct on the math problems.
- During the feedback, you will also see your percent correct for the math problems for the entire experiment.
- It is VERY important for you to keep this at least at 85%.
- For our purposes, we can only use data where the participant was at least 85% accurate on the math.
- Therefore, you must perform at least at 85% on the math problems WHILE doing your best to recall as many letters as possible.

Appendix C: General Rating of Trust in Automation

Below are several statements about the targeting aid that you just used (referred to as the "system").

Please rate your feelings about the aid from "not at all" to "extremely" (click one of the 7 buttons in a row for each question).

1. The system is deceptive						
1 Not at all	2	3	4	5	6	7 Extremely

2. The system behaves in an underhanded manner						
1 Not at all	2	3	4	5	6	7 Extremely

3. I am suspicious of the system's intent, action, or outputs						
1 Not at all	2	3	4	5	6	7 Extremely

4. I am wary of the system						
1 Not at all	2	3	4	5	6	7 Extremely

5. The system's actions will have a harmful or injurious outcome						
1 Not at all	2	3	4	5	6	7 Extremely

6. I am confident in the system						
1 Not at all	2	3	4	5	6	7 Extremely

7. The system provides security						
1 Not at all	2	3	4	5	6	7 Extremely

8. The system has integrity						
1 Not at all	2	3	4	5	6	7 Extremely

9. The system is dependable						
1 Not at all	2	3	4	5	6	7 Extremely

10. The system is reliable						
1 Not at all	2	3	4	5	6	7 Extremely

11. I can trust the system						
1 Not at all	2	3	4	5	6	7 Extremely

12. I am familiar with the system						
1 Not at all	2	3	4	5	6	7 Extremely

Appendix D: Subjective Trust in the Automated Aid

To what extent did you trust (i.e. believe in the accuracy of) the automation aid in this scenario?

< [Progress Bar] >

Not at all Extremely

To what extent did you rely on (i.e. actually use) the automation aid in this scenario?

< [Progress Bar] >

Not at all Extremely

To what extent were you self-confident that you could successfully perform without the automation aid in this scenario?

< [Progress Bar] >

Not at all Extremely

To what extent do you think the automation improved your performance in this scenario compared to performance without the automation?

< [Progress Bar] >

Not at all Extremely

Appendix E: Complacency Potential Rating Scale

1. Manually sorting through card catalogs is more reliable than computer-aided searches for finding items in a library.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
2. If I need to have a tumor in my body removed, I would choose to undergo computer-aided surgery using laser technology because computerized surgery is more reliable and safer than manual surgery.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
3. People save time by using automatic teller machines (ATMs) rather than a bank teller in making transactions.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
4. I do not trust automated devices such as ATMs and computerized airline reservations systems.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
5. People who work frequently with automated devices have lower job satisfaction because they feel less involved in their job and those who work manually.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
6. I feel safer depositing my money at an ATM then with a human teller.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
7. I have to record an important TV program for a class assignment. To ensure that the correct program is recorded, I would use the automatic programming facility on my recording device rather than manual taping.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
8. People whose jobs require them to work with automated systems are lonelier than people who do not work with such devices.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
9. Automated systems used in modern aircraft, such as the automatic landing system, have made their journey safer.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
10. ATMs provide a safeguard against the inappropriate use of an individual's bank account by dishonest people.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
11. Automated devices used in aviation and banking have made work easier for both employees and customers.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
12. I often use automated devices.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
13. People who work with automated devices have greater job satisfaction because they feel more involved than those who work manually.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
14. Automated devices in medicine save time and money in the diagnosis and treatment of disease.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
15. Even though the automatic cruise control in my car is set to a speed below the speed limit, I worry when I pass police radar speed-trap in case the automatic control is not working properly.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
16. Bank transactions have become safer with the introduction of computer technology for the transfer of funds.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
17. I would rather purchase an item using a computer that have to deal with the sales representative on the phone because my order is more likely to be correct using the computer.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
18. Work has become more difficult with the increase of automation in aviation and banking.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
19. I do not like to use ATMs because I feel that they are sometimes unreliable.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
20. I think that automated devices used in medicine, such as CAT scans and ultrasound, provide very reliable medical diagnosis.	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree

1.

1. Report Type

Final Report

Primary Contact E-mail

Contact email if there is a problem with the report.

richpak@clemson.edu

Primary Contact Phone Number

Contact phone number if there is a problem with the report

864-656-1584

Organization / Institution name

Clemson University

Grant/Contract Title

The full title of the funded effort.

Anthropomorphic Interfaces on Automation Trust, Dependence, and Performance in younger and Older Adults

Grant/Contract Number

AFOSR assigned control number. It must begin with "FA9550" or "F49620" or "FA2386".

FA9550-12-1-0385

Principal Investigator Name

The full name of the principal investigator on the grant or contract.

Chong Hyon Richard Pak

Program Manager

The AFOSR Program Manager currently assigned to the award

Benjamin Knott

Reporting Period Start Date

07/15/2012

Reporting Period End Date

07/14/2015

Abstract

This proposal sought to better understand the psychological component of human-automation interaction with a focus on understanding what makes automation seem "trustable". Specifically, we will investigate the role of anthropomorphic automation on operator's trust, dependence, and performance with automation. Evidence from the literature and our own recently collected data suggests that the design of automation can affect how operators perceive the automation and their likelihood of using it. We seek to investigate the conditions under which anthropomorphized automation, or automation that appears to possess human-like characteristics, affects the calibration of trust between the operator and the system. A secondary goal is to understand how anthropomorphic automation effects are moderated by the age of the operator. Older users have different reactions to automation (some research shows over-trust while other research shows under-trust).

Distribution Statement

This is block 12 on the SF298 form.

Distribution A - Approved for Public Release

Explanation for Distribution Statement

DISTRIBUTION A: Distribution approved for public release.

If this is not approved for public release, please provide a short explanation. E.g., contains proprietary information.

SF298 Form

Please attach your [SF298](#) form. A blank SF298 can be found [here](#). Please do not password protect or secure the PDF. The maximum file size for an SF298 is 50MB.

[AFD-070820-035.pdf](#)

Upload the Report Document. File must be a PDF. Please do not password protect or secure the PDF. The maximum file size for the Report Document is 50MB.

[pak final report.pdf](#)

Upload a Report Document, if any. The maximum file size for the Report Document is 50MB.

Archival Publications (published) during reporting period:

Publications:

Pak, R., McLaughlin, A. C., & Bass, B. (2014). A Multi-level Analysis of the Effects of Age and Gender Stereotypes on Trust in Anthropomorphic Technology by Younger and Older Adults. *Ergonomics*.

Rovira, E., Pak, R., & McLaughlin, A. C. (under review). Low Memory, Mo' Problems: Effects of individual differences on types and levels of automation. *Human Factors*.

Conference Proceedings

Leidheiser, W., & Pak, R. (2014). The Effects of Age and Working Memory Demands on Automation-Induced Complacency. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58(1), 1919–1923. doi:10.1177/1541931214581401

Bass, B. M., Goodwin, M., Brennan, K., Pak, R., & McLaughlin, A. C. (2013). Effects of age and gender stereotypes on trust in an anthropomorphic decision aid. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 57(1), 1575-1579.

Bass, B. M., & Pak, R. (2012). Faces as Ambient Displays: Assessing the attention-demanding characteristics of facial expressions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56(1), 2142–2146.

Changes in research objectives (if any):

Change in AFOSR Program Manager, if any:

Awarding program manager (2012): Joseph Lyons

Current program manager (2015): Benjamin Knott

Extensions granted or milestones slipped, if any:

AFOSR LRIR Number

LRIR Title

Reporting Period

Laboratory Task Manager

Program Officer

Research Objectives

Technical Summary

Funding Summary by Cost Category (by FY, \$K)

	Starting FY	FY+1	FY+2
Salary			
Equipment/Facilities			
Supplies			
Total			

Report Document

Report Document - Text Analysis

Report Document - Text Analysis

Appendix Documents

2. Thank You

E-mail user

Oct 14, 2015 10:17:59 Success: Email Sent to: richpak@clemsn.edu