

EFFECTS OF AGE ON UTILIZATION AND PERCEIVED RELIABILITY OF AN AUTOMATED DECISION-MAKING AID FOR LUGGAGE SCREENING

Jason S. McCarley,
Mississippi State University
Mississippi State, MS

Doug A. Wiegmann,
Christopher D. Wickens,
and Arthur F. Kramer
University of Illinois
Urbana-Champaign, IL

An experiment examined the effects of age on utilization and perceived reliability of an imperfectly reliable decision-making aid in a luggage x-ray screening task. Forty-five young adults and 45 elderly adults performed a simulated luggage screening task. Some subjects were provided the assistance of an automated decision aid with a hit rate of .90 and a false alarm rate of .25. Others performed the task with no aid. Signal-detection analysis revealed that automation improved sensitivity only for younger participants, suggesting a tendency for older participants to underutilize the aid's recommendations. Data also revealed unique patterns of individual differences in cue reliance among older and younger participants. Perceived reliability of the aid did not differ between age groups. Order of information presentation (with the aid's recommendation coming before or after the raw data) had little effect for either age group.

INTRODUCTION

Automated diagnostic aids offer the potential to enhance decision-making by circumventing the limited information processing capabilities and biases of human operators (Mosier & Skitka, 1996). The full extent to which these aids might actually facilitate performance in real-world tasks is unknown, however, particularly since such aids are unlikely to be 100% reliable. Research has found that operators undertrust automation that is imperfectly reliable, disagreeing with the aid more frequently than is optimal (e.g., Wiegmann, 2002). The utility of an automated aid, therefore, is determined not just by the actual reliability of the automation, but by the user's perception of the automation's reliability (Dzindolet, Pierce, Beck, & Dawe, 2002).

To date, research on human operators' utilization of decision-making aids has focused largely on performance of young adults. The design of effective automation for real-world use, however, will require an understanding of aid utilization among older adults as well. Given the physical, sensory, and cognitive declines that accompany normal aging, it is possible that elderly adults may frequently benefit more from the support of an automated aid than young adults. Alternatively, the extent to which differences in social or cognitive

factors among age groups might modulate the perceived utility or trustworthiness of an aid, it is possible that the tendency for operators to under-realize the value of an automated system might be greater for older than for younger adults.

The ordering of diagnostic information may also bias human decision making (Hogarth & Einhorn, 1992; Wickens & Hollands, 2000). However, such effects might be further modulated by age-related factors. Decision makers are often disinclined to modify existing beliefs in response to new, discrepant information. This suggests that operators who view raw data and develop an hypothesis prior to receiving an aid's conflicting diagnosis may be more likely to disagree with the aid than if they had received its advice first. Conversely, operators who are provided an aid's advice before viewing raw data may unduly bias their decisions in favor of the aid's recommendation, resulting in degraded performance if the aid is imperfectly reliable (i.e., the aid produces false alarms or misses). Utilization of an automated aid might therefore vary as a result of the order in which information is provided to operators. Furthermore, findings of age-related declines in cognitive flexibility (Sanfrey & Hastie, 2000) suggest that the effects of information order might be greater for elderly than for young adults.

The aim of the present study was to examine automation trust and reliance in young and elderly

adults performing a simulated luggage-screening task. To test the effects of information order, we varied the sequence in which operators received information from the automated aid. For some operators, the aid's recommendation preceded the raw data (luggage X-ray), whereas for other participants, the recommendation was provided only after the raw data had been inspected.

METHOD

Participants

Participants were 45 young adults (mean age = 20.8 years) and 45 older adults (mean age = 68.7 years). All participants had corrected far and near acuities of 20/30 or better.

Stimuli

Stimuli were chromatic x-ray images of airline passenger luggage, moderately to densely cluttered with a variety of everyday objects (e.g., clothes, hair dryers, pill bottles). A subset of 20% of the images, the *target-present* stimuli, also contained a digitally superimposed x-ray image of a knife. The same knife was used in all target-present images. In one half of all target-present stimuli the knife was easily detectable. In the other half, the knife was difficult to detect. Classification of targets as easy or difficult to detect was based on ratings provided by two observers in a pilot experiment.

Procedure

The participants' task on each trial was to view a stimulus image for 3 seconds, then decide whether the pictured bag contained a knife. Participants provided responses by using a mouse to point-and-click on labeled onscreen buttons. A text message providing feedback about response accuracy appeared following each trial.

Participants were divided into three groups. Two groups received assistance on each trial from an automated aid. The aid advised the observer, via a text message, as to whether or not a target was present. Hit rate for the aid (i.e., accuracy rate of the aid on target-present trials) was .9. False alarm rate for the automated aid was .25. Participants in the

pre-cue group ($n = 15$ young adults, 15 older adults) received the automated message prior to viewing the stimulus image itself. Participants in the *post-cue* group ($n = 15$ young adults, 15 older adults) received the automated message after viewing the stimulus image. Participants in the control *no-cue* group ($n = 15$ young adults, 15 older adults) received no automated recommendations. Participants in the pre-cue and post-cue groups were informed that the automation might sometimes make a bad recommendation and that they were free to disagree with the automation if they wished. Each participant performed, in random order, 160 target absent-trials, 20 easy target-present trials, and 20 difficult target-present trials. Participants from the cued experimental groups also completed a post-experimental questionnaire that asked them to estimate the hit rate and false alarm rates of the automated aid.

RESULTS

Analysis of behavioral data focused exclusively on detection of difficult targets. Sensitivity for target detection of these targets was quantified by the signal detection measure $P(A)$ (Wickens & Hollands, 2000). Planned comparisons were employed to test for differences A) between performance of cued participants and control participants, and B) between performance of pre-cued and post-cued participants.

Analysis of young adults' behavioral data revealed that sensitivity was similar for pre-cued ($M = .78$) and post-cued groups ($M = 0.80$), $F < 1$, but was higher overall for cued than for the non-cued participants ($M = 0.71$), $F(1, 42) = 9.38$, $p < .01$. Hit rates were significantly higher for cued than for uncued participants' ($M = 62\%$ vs. $M = 48\%$), $F(1, 42) = 9.29$, $p < .01$, while false alarm rates were similar across groups ($M = 4\%$ vs. $M = 5\%$), $F < 1$. In contrast, examination of older adults' behavioral data not only failed to reveal any difference in sensitivity between pre-cued ($M = .62$) and post-cued ($M = .60$) groups, $F < 1$, but also failed to indicate any difference in sensitivity between cued and non-cued ($M = .60$) groups, $F < 1$. Hit rates were statistically identical between cued and uncued groups ($M = 29\%$ vs. $M = 27\%$, $p > .75$), $F < 1$, as were false alarm rates ($M = 5\%$ vs. $M = 7\%$), $F < 1$.

To assess the degree to which individual participants relied on the automated aid, we calculated the difference between participants' false alarm rates on inaccurately-cued and accurately-cued target absent trials (i.e., noise trials). Note that larger scores would indicate a greater tendency to agree with the aid on noise trials. We correlated this "reliance index" for noise trials with hit rates for accurately cued target-present trials. For younger participants, the resultant correlation was negative, $r = -.376$, $p = .04$. Younger participants whose hit rates were highest tended to rely on the automation less, thus producing fewer false alarms. In contrast, the correlation between cue reliance on noise trials and hit rate for older observers was significantly positive, $r = .58$, $p = .001$. Thus older participants who did well in detecting targets were generally those who relied on the aid, and thus produced higher false alarms on noise trials.

Cued groups' mean post-experimental estimate of the automation's hit rate was 65% for young adult participants and 68% for elderly adult participants. Mean estimate of false alarm rate was 26% for young adult participants and 35% for elderly adult participants. These estimates were not reliably different between age-groups, $ps > .15$.

DISCUSSION

The present study found clear differences in young and elderly adults' utilization of an automated decision aid in a simulated luggage screening task. For young adult participants, the assistance of an automated decision aid with a hit rate of 90% and false alarm rate of 25% reliably increased operator sensitivity relative to baseline, resulting in a significant increase in target detection rates relative to that of an uncued control group. Young participants however did not completely utilize the aid, achieving a hit rate (62%) that was significantly lower than was ultimately obtainable given the 90% hit rate of the aid. Apparently, the occasional false alarms made by the aid undermined participants' trust in the aid's ability to detect targets when they were present, resulting in periodic disagreements with aid's diagnosis. Indeed, young participants accurately estimated the false alarm rate of the aid, but dramatically underestimated its hit rate. This finding is consistent with previous

research indicating that when automation is less than perfectly reliable, operators tend to underestimate its true reliability (Wickens & Hollands, 2000; Wiegmann, 2002).

For older adults, the assistance of the same aid produced no significant increase in sensitivity, and no reliable increase in hit rate relative to the control group. Data thus suggest a tendency for older participants to disregard the automated aid altogether. This effect, however, does not appear to be due solely to older participants distorting the true reliability of the aid. Albeit they too tended to underestimate the aid's hit rate. However, such estimates did not differ from those of the young participants. Therefore, some other factors appear to have influenced older adults' utilization of the aid in addition to automation trust. More research is needed, however, to identify these factors so that appropriate automation utilization can be facilitated in older adults, who apparently are in greater need of such assistance given the much poorer performance of the older adults in general.

Different patterns of individual responses to automation failures across younger and older participants were also observed. Younger participants' data evidenced a negative correlation between cue reliance and target detection rate, suggesting that participants who were better able to detect the target tended to rely on the automated aid less than did participants who had greater difficulty detecting targets. In contrast, the correlation between cue reliance and target detection rate was significantly positive in older participants data, suggesting that cue reliance was not modulated by individual participants' level of ability. Hence, older participants performance tended to be based on a bias to either agree or disagree with the aid rather than on their actual sensitivity, which on average was generally poorer than younger participants.

Agreement rates were similar in the pre- and post-cue conditions for both age groups, suggesting that order of information presentation did little to bias decision making in favor of or against the recommendations of the aid. Unlike previous studies in this area (e.g., Wiegmann, 2002), the present study mimicked a real-world scenario with a relatively low rate of target present trials, as well as the inclusion of easy target trials. Such conditions may serve to counteract anchoring effects.

Additional research is also needed to explore this issue.

ACKNOWLEDGMENTS

This research was supported in part by a grant from the Center on Aging and Cognition: Health, Education and Training (CACHET) and the Federal Aviation Administration. CACHET is an Edward R. Roybal Center funded by the National Institute on Aging (NIA). The CACHET contract monitor was Denise Park. The FAA contract monitor was Joshua Rubinstein. The views expressed are those of the authors and do not necessarily reflect those of the CACHET or FAA.

REFERENCES

- 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.
- Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24, 1-55.
- 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: Lawrence Erlbaum.
- Sanfey, A. G., & Hastie, R. (2000). Judgment and decision making across the life span: A tutorial review of psychological research. In D. Park & N. Schwarz (Eds.), *Cognitive aging*. Philadelphia, PA: Psychology Press.
- Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology and human performance* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- Wiegmann, D. A. (2002). Agreeing with automated diagnostic aids: A study of users' concurrence strategies. *Human Factors*.