Effects of Touch, Voice, and Multimodal Input, and Task Load on Multiple-UAV Monitoring Performance During Simulated Manned-Unmanned Teaming in a Military Helicopter

Samuel J. Levulis, Texas Tech University, Lubbock, USA,
Patricia R. DeLucia, Rice University, Houston, Texas, USA, and
So Young Kim, NASA Jet Propulsion Laboratory, Pasadena, California, USA

Objective: We evaluated three interface input methods for a simulated manned-unmanned teaming (MUM-T) supervisory control system designed for Air Mission Commanders (AMCs) in Black Hawk helicopters.

Background: A key component of the U.S. Army's vision for unmanned aerial vehicles (UAVs) is to integrate UAVs into manned missions, called MUM-T (Department of Defense, 2010). One application of MUM-T is to provide the AMC of a team of Black Hawk helicopters control of multiple UAVs, offering advanced reconnaissance and real-time intelligence of flight routes and landing zones.

Method: Participants supervised a (simulated) team of two helicopters and three UAVs while traveling toward a landing zone to deploy ground troops. Participants classified aerial photographs collected by UAVs, monitored instrument warnings, and responded to radio communications. We manipulated interface input modality (touch, voice, multimodal) and task load (number of photographs).

Results: Compared with voice, touch and multimodal control resulted in better performance on all tasks and resulted in lower subjective workload and greater subjective situation awareness, *p*s < .05. Participants with higher spatial ability classified more aerial photographs (r = .75) and exhibited shorter response times to instrument warnings (r = −.58) than participants with lower spatial ability.

Conclusion: Touchscreen and multimodal control were superior to voice control in a supervisory control task that involved monitoring visual displays and communicating on radio channels.

Application: Although voice control is often considered a more natural and less physically demanding input method, caution is needed when designing visual displays for users sharing common communication channels.

Keywords: supervisory control, uninhabited aerial vehicles, touchscreens, speech user interfaces, multimodal displays

Address correspondence to Patricia R. DeLucia, Department of Psychology, MS 25, Rice University, Houston, TX 77251-1892, USA; e-mail: pat.delucia@rice.edu.

HUMAN FACTORS
Vol. 60, No. 8, December 2018, pp. 1117–1129
DOI: 10.1177/0018720818788995

INTRODUCTION

A key component of the U.S. Army's vision for unmanned aerial vehicles (UAVs) in future operations is manned-unmanned teaming (MUM-T), which involves the integration of manned and unmanned vehicles, sensors, and weapons (Department of Defense, 2010). One application of MUM-T that offers substantial value is to give the Air Mission Commander (AMC) of a team of Black Hawk helicopters control of UAVs to reduce current command-and-control latencies and provide the AMC with greater situation awareness (SA) of the tactical situation.

Designing a system that enables multi-UAV control without exceeding the AMC's performance limitations presents a myriad of challenges, including the selection of an interface control method that facilitates natural and efficient interaction. Conventional interfaces in Black Hawks are controlled with buttons and knobs (e.g., buttons on the bezels of multifunction displays) or cursor control. Two emerging technologies with the potential to enhance an AMC's effectiveness on the flight deck are touchscreens and voice recognition software, and thus were selected for the current study.

Theoretical Motivation

Compared with conventional flight deck displays, touchscreens offer engineering advantages, such as reduced weight and greater modifiability (Cockburn et al., 2017). Touchscreens also may provide greater location compatibility (Wickens & Carswell, 2012) than conventional displays because users can interact directly with the display elements themselves, rather than use buttons or knobs physically displaced from the display.

Compared with voice control, touchscreen control offers more effective use in high-noise environments and less interference with radio or intracrew communication tasks. Moreover, touchscreens could offer greater stimulus-central processing-response (SCR) compatibility than voice control. SCR compatibility is the tendency for performance to benefit when the stimulus and response modes for a task correspond with the central processing codes (spatial vs. verbal/linguistic) that the task utilizes (Proctor & Vu, 2016; Wickens, Vidulich, & Sandry-Garza, 1984). Because many of the AMC's current and envisioned tasks rely heavily on spatial codes of processing (e.g., maintaining awareness of the positions of vehicles in the flight, analyzing UAV sensor feed imagery), performance may benefit more from touch responses than voice responses.

Voice control also offers advantages, such as the ability to control an interface from a distance, the potential to replace a series of manual inputs with a single "voice macro," and the capability of hands-free and heads-up control (Calhoun & Draper, 2006). Moreover, due to greater SCR compatibility, voice control should be advantageous for tasks that involve printed and spoken information, or performing linguistic operations, mental arithmetic, or rehearsal (Wickens et al., 1984).

Voice control could also reduce the costs associated with information access effort, which should increase the probability of an operator sampling information from a given display element (Wickens & McCarley, 2008). Specifically, voice control could reduce scanning requirements (e.g., allowing the AMC to call up a UAV by name rather than search the display for its icon) and head and trunk movements (e.g., allowing the AMC to vocally acknowledge an auditory warning rather than turn from one display to another to provide manual input).

Additionally, voice control eliminates the physical demands associated with regularly moving the upper limbs to interact with a display. Humans have a natural tendency to minimize information access effort costs, although top-down factors such as training and expertise can counteract this propensity (Wickens & McCarley, 2008). However, voice control could present greater cognitive costs than touchscreen control due to demands of recalling and articulating voice commands, and such costs may reduce the probability of a user accessing information with this modality (Wickens & McCarley, 2008).

Touchscreen and voice control have relative strengths and weaknesses. Providing users with simultaneous access to both (i.e., multimodal control) could offset the weaknesses of one modality with the strengths of the other (Chen, 2006), potentially resulting in many advantages (for a review, see Turk, 2014). For example, not only has multimodal interaction been considered more natural than unimodal interaction, research has shown that users prefer multimodal interfaces, and that multimodal interaction can be more flexible and efficient (Turk, 2014). Consistent with multiple resource theory (Wickens, 2008), multimodal displays may also improve user performance because activity can be divided across separate response modalities.

Purpose

Few studies have systematically evaluated how the choice of interface input method affects operator performance with UAV control systems (see Calhoun, Ruff, Behymer, & Rothwell, 2017; Draper, Calhoun, Ruff, Williamson, & Barry, 2003; Taylor et al., 2015). None have evaluated systems designed for an AMC located in a Black Hawk helicopter, which is important because it is critical to consider the specific application when designing interfaces for multi-UAV control (Calhoun & Draper, 2015). The purpose of this study was to evaluate three interface input methods (touch, voice, multimodal) for a MUM-T supervisory control system designed for an AMC located in a Black Hawk helicopter.

Participants in our study supervised a (simulated) team of two helicopters and three UAVs as they traveled toward a landing zone to deploy ground troops. Participants classified aerial photographs collected by their UAVs, monitored instrument warnings, and responded to radio communications. The aims of the study are to: (1) Quantify the effects of the interface input method on AMC task performance, (2) identify how the effectiveness of each input method

varies with task difficulty, (3) determine whether the input methods interfere with verbal communication tasks, and (4) examine whether individuals with greater spatial ability and attentional control exhibit greater task performance, as reported in prior literature on unmanned systems (see Cahillane, Baber, & Morin, 2012).

In line with these aims, we examined the following hypotheses:

H1: Due to its superior flexibility, multimodal input will result in greater task performance (more classified photographs, greater sensitivity classifying photographs, and shorter response times [RTs] to warnings) than touch or voice input alone.

H2: Because the requirement for maximal efficiency is more critical, advantages of multimodal input will be more pronounced with increased task difficulty, indicated by greater differences in task performance relative to the other input modalities in high task load conditions.

H3: Due to greater overlap in codes of processing and modalities of response, voice input will interfere with verbal communication tasks more than touch or multimodal input, indicated by longer RTs and lower accuracy for communication task queries.

H4: Consistent with previous studies on operator performance with unmanned systems, participants with greater attentional control and spatial ability will exhibit greater task performance.

METHOD

Participants

Eighteen students and staff members from Texas Tech University (17 male) between 18 and 26 years old ($M = 19.72$, $SD = 2.22$) participated. All participants reported normal vision, hearing, and motor control via a prescreening questionnaire. All were fluent English speakers with American accents; two were working toward a private pilot license, and two were members of the university's Air Force Reserve Officer Training Corps program. Participants were required to have averaged at least 5 hours of action video game play (e.g., action adventure, first-person shooter) per week across the previous 12 months, because action video game players have shown superior performance on tasks relevant to multi-UAV control (e.g., Chen & Barnes, 2012a, 2012b; McKinley, McIntire, & Funke, 2011). Self-reported action video game experience across the previous 12 months ranged from 5 to 23 hrs/week ($M = 12.58$ hrs, $SD = 5.02$ hrs). Participants received $25 compensation, with the additional incentive of $30 gift cards given to the two highest-performing participants. This research complied with the American Psychological Association Code of Ethics and was approved by the institutional review board at Texas Tech University. Informed consent was obtained from each participant.

Apparatus

The study was conducted with a Dell Optiplex 390 computer with integrated Intel HD Graphics 2000, and two Dell touchscreen LED monitors with 1920 × 1080 resolution and refresh rates of 60 frames/s. Displays (see Figure 1) simulated monitoring and control screens used by an AMC in the front-left seat of a UH-60 Black Hawk. The primary monitor had a 27-in. diagonal viewing area. The secondary monitor had a 23-in. diagonal viewing area, but the software only filled the top-left 10.88 in. (horizontal) × 7.13 in. (vertical); the remaining portion was black and inactive.

Participants wore a GE monaural hands-free headset paired with Intel RealSense version 6.0.21 voice recognition software. The headset's microphone was set one inch from the participant's mouth. The observed reliability of the voice recognition software was 94.85% (see Supplement A for details). Communication task queries were played with VLC Media Player version 2.2.1. Participants wore SMI mobile eye-tracking glasses for another study; these data are not reported here.

Displays & Controls

The selected components of the AMC's tasks were determined on the basis of software constraints (i.e., what was possible to make functional at the current point in the development lifecycle), and discussions with subject-matter experts (SMEs) with experience as AMCs in the U.S. Army.

Figure 1. The experimental display configuration.

Primary monitor. Displays consisted of a top-down moving map, map controls, a primary flight display, and a power pod instrument (showing helicopter engine performance). For this study, the map controls and primary flight display were static and nonfunctional. The moving map contained five icons: Two circular icons represented the participant's helicopter and a lead helicopter. Three hexagonal icons (labelled 1, 11, and 23) represented Gray Eagle UAVs under the AMC's command.

The scenarios simulated a portion of an air assault (troop insertion) mission in which UAVs performed route reconnaissance and monitored for threats along the flight plan. The UAVs tracked the helicopters along one of six randomly selected flight plans during each trial and intermittently detected "targets of opportunity" (TOOs), which were people located on the ground along the flight path. When a UAV detected a TOO, it captured an aerial photograph of it to be inspected by the AMC, and the UAV's icon became outlined in yellow and began to flash. The TOOs could be hostiles (individuals in a lowered stance holding a rocket launcher on their shoulder) or neutrals (individuals standing upright without a rocket launcher).

At four randomly selected times within each trial, the power pod moved into a nonnormal state and turned red, and the word "Warning" appeared. An auditory alarm repeated in the participant's headset until the warning was acknowledged. The alarm was a 1-s complex tone with two whoops per second (peak frequency = 516 Hz; average intensity = 71.25 dB).

Secondary monitor. Displays included aerial photographs of TOOs taken by the UAVs. The photograph appeared when the participant touched the UAV's icon on the primary monitor.

Radio communications. In cognitive task analyses conducted with AMCs, we determined that intrateam, auditory-vocal communications play a pivotal role in their development of SA and will likely still exist in future operations. To simulate these communication demands, participants completed three types of communication tasks partly based on Chen and Joyner (2009): A spatial-reasoning task (e.g., "If the tank is to your left and the Humvee is to your right, what direction is the tank to the Humvee?"), a short-term memory task (e.g., "Please remember the following words: winter, lemon, penny. . . . Please repeat the words."), and a call sign task, in which participants responded when they heard their call sign but not when they heard other call signs. Each trial contained 12 communication task queries. The order and times of queries were randomized with the constraint that queries were not presented within 15 s of one another.

Controls. One of three interface input methods (hereafter referred to as modalities) were used to control the displays in a given block of trials: touch, voice, or multimodal (touch and voice). Within the touch and voice modalities there were eight control commands (see Table 1); all 16 were available during multimodal trials.

Procedure

After providing informed consent, participants completed the Attentional Control Scale (ACS; Derryberry & Reed, 2002) and revised Object Perspective Test (rOPT; Hegarty & Waller, 2004) to assess for individual differences that might moderate task performance and potentially inform operator selection for future UAV supervisory control systems. The ACS is a 20-item questionnaire that assesses individuals' abilities to flexibly focus and shift attention. The rOPT is a spatial, perspective-taking test (for details, see Hegarty & Waller, 2004). Previous studies showed that ACS and rOPT scores were related to operator performance with unmanned

TABLE 1: Commands Used to Control the Interface With the Three Interface Input Modalities

	Modality		
Command	Touch	Voice	Multimodal
Pull up sensor feed for:			
Gray Eagle 1	Touch Gray Eagle 1's icon	Say "Gray Eagle 1"	Touch Gray Eagle 1's icon or Say "Gray Eagle 1"
Gray Eagle 11	Touch Gray Eagle 11's icon	Say "Gray Eagle 11"	Touch Gray Eagle 11's icon or Say "Gray Eagle 11"
Gray Eagle 23	Touch Gray Eagle 23's icon	Say "Gray Eagle 23"	Touch Gray Eagle 23's icon or Say "Gray Eagle 23"
Classify photograph as:			
Neutral	Touch *Neutral* button	Say "classify neutral"	Touch *Neutral* button or Say "classify neutral"
Hostile	Touch *Hostile* button	Say "classify hostile"	Touch *Hostile* button or Say "classify hostile"
Show next photograph the displayed Gray Eagle has taken	Swipe to the left on currently displayed photograph	Say "show next"	Swipe to the left on currently displayed photograph or Say "show next"
Show previous photograph the displayed Gray Eagle has taken	Swipe to the right on currently displayed photograph	Say "show previous"	Swipe to the right on currently displayed photograph or Say "show previous"
Acknowledge power pod warning	Touch *Warning* button above power pod	Say "acknowledge warning"	Touch *Warning* button above power pod or Say "acknowledge warning"

systems (e.g., Baber et al., 2011; Chen & Barnes, 2012b). After completing the ACS and rOPT, participants watched a 20-minute presentation detailing the displays and tasks.

Participants completed three blocks of trials, one for each modality. Prior to completing the experimental trials for a given modality, participants were trained on the commands and completed a 2-minute practice trial. All participants reported being fully comfortable with the commands prior to beginning the experimental trials in each block.

For each modality, participants completed two 8.5-minute experimental trials that differed in task load—operationalized as the number of detected TOOs. We conducted pilot testing to determine a task load that was moderately difficult (72 TOOs) and another that was very difficult but possible to complete (92 TOOs). This manipulation provided a test of our hypothesis that the relative effectiveness of multimodal input would be more pronounced with increased task difficulty.

After each trial, participants completed the National Aeronautics and Space Administration Raw Task Load Index (NASA-RTLX; Byers, Bittner, & Hill, 1989) and the 3D Situation Awareness Rating Technique (3D-SART; Jones, 2000) to assess subjective mental workload and

SA, respectively. After each block of trials, participants completed the Post-Study System Usability Questionnaire (PSSUQ; Lewis, 1995) to assess perceived system usability with the given modality.

Participants were told that they were responsible for overseeing two helicopters and three UAVs as they traveled toward a landing zone to deploy ground troops, and that they were responsible for completing three tasks: Classify aerial photographs collected by their UAVs (as neutral or hostile), acknowledge the power pod warning whenever it became activated, and respond to radio communications. Participants were instructed to complete all tasks as quickly and as accurately as possible. Participants were permitted to adopt the control strategy they felt was most efficient with a given modality. That is, we did not require them to use one or two hands in touch trials, and we did not attempt to dictate the proportions of touch and voice commands they used in multimodal trials. The entire experiment lasted 2.25 hrs.

Experimental Design

Our main independent variable of interest was interface input modality. We also manipulated task load within each modality (72 or 92 TOOs detected per trial). The order of modalities was counterbalanced across participants. The orders of flight plans and task loads were randomized. TOO detection times were randomized with the constraint that TOOs were not detected within the last 15 s of a trial to prevent situations in which it would be impossible for participants to classify a photograph in time.

Dependent Measures

For the TOO classification task, we measured the percentage of TOOs that participants were able to classify (correctly or incorrectly) in each trial. We also calculated signal detection theory measures (Green & Swets, 1966) of sensitivity (d') and response bias (β). For the power pod warning task, we measured RT (time between alarm onset and participant response). For the communication task, we measured RT and the percentage of correct responses.

We also analyzed subjective measures of mental workload (NASA-RTLX), SA (3D-SART), and usability (PSSUQ). Results for these measures are provided in Supplement B; compared with voice, touch and multimodal control resulted in lower subjective workload, greater subjective SA, and higher usability ratings (ps < .05). Finally, we evaluated correlations between the aforementioned dependent measures (except PSSUQ) and scores on the ACS and rOPT.

RESULTS

Performance measures were analyzed separately with 2 (task load: low, high) × 3 (modality: touch, voice, multimodal) repeated-measures ANOVAs. For significant interactions, separate one-way ANOVAs were conducted to assess for simple effects of modality for the low and high task loads. Probability values for the effect of modality reflect Greenhouse-Geisser corrections for any violations of the assumption of sphericity. Means for the effect of modality are presented in Table 2. Significant effects of modality were followed up with Tukey HSD pairwise comparisons (the probability values reported are adjusted for multiple comparisons and therefore evaluated against an alpha of .05).

TOO Classification Task

Percentage TOOs classified. There was a significant main effect of modality, $F(2, 34) = 29.17$, $p < .001$, $\eta^2_p = .63$; participants classified fewer TOOs in voice trials than in touch ($p < .001$) or multimodal ($p < .001$) trials. There was also a significant main effect of task load, $F(1, 17) = 37.97$, $p < .001$, $\eta^2_p = .69$; participants classified a greater percentage of TOOs in low task load trials ($M = 97.89\%$, $SD = 3.03\%$) than in high task load trials ($M = 90.20\%$, $SD = 6.76\%$).

There was a significant Modality × Task Load interaction, $F(2, 34) = 13.71$, $p < .001$, $\eta^2_p = .45$ (see Figure 2). For both the low, $F(2, 34) = 6.56$, $p = .004$, $\eta^2_p = 0.28$, and high, $F(2, 34) = 26.36$, $p < .001$, $\eta^2_p = 0.61$, task loads, participants classified fewer TOOs in voice trials than in touch or multimodal trials (all ps < .03). Inspection of Figure 2 shows that the limitations of voice input are exacerbated under higher task load conditions.

d'. There was a significant main effect of modality, $F(2, 34) = 4.04$, $p = .048$, $\eta^2_p = .19$; mean d' was lower for voice trials than touch

TABLE 2: Results of the TOO Classification, Power Pod Warning, and Communication Tasks

	Modality *M (SD)*		
Measure	Touch	Voice	Multimodal
TOO classification task			
Percentage TOOs classified [M,T,M*T]	98.00% (5.19%)	87.06% (7.76%)	97.07% (4.40%)
d' [M, M*T]	3.94 (0.44)	3.75 (0.49)	3.83 (0.42)
β	1.41 (0.55)	1.39 (0.59)	1.36 (0.58)
Power pod warning task			
RT [M]	1.91 s (0.29 s)	5.36 s (1.88 s)	2.60 s (1.06 s)
Communication task			
RT	1.09 s (0.63 s)	1.01 s (0.56 s)	0.89 s (0.43 s)
Percentage correct [M]	92.79% (10.36%)	87.01% (8.11%)	93.42% (7.97%)

Note. M = Significant main effect of modality (touch, voice, multimodal); T = Significant main effect of task load (low, high); M*T = Significant interaction between modality and task load. The nature of the interactions for the percentage of TOOs classified and d' are described in the text.

trials (p = .021); mean d' for multimodal trials was not significantly different from touch (p = .259) or voice (p = .443) trials. The main effect of task load was not significant (p > .10).

There was a significant Modality × Task Load interaction, $F(2, 34) = 7.39$, p = .006, η^2_p = .30, represented in Figure 3. The effect of modality was significant only in high task load trials $F(2, 34) = 6.59$, p = .010, η^2_p = .28, and mean d' was lower in voice trials than in touch (p = .004) or multimodal (p = .027) trials.

β. The main effects of modality and task load on mean β were not significant, ps > .55. Mean β values ranged from 1.31 to 1.47 in the six conditions created by crossing modality and task load.

Power Pod Warning Task

There was a significant main effect of modality on mean RT, $F(2, 34) = 34.63$, $p < .001$, η^2_p = .67; mean RT was longer for voice trials than touch (p < .001) or multimodal (p < .001) trials. This was partially due to the voice recognition system's lower reliability for the "acknowledge warning" command for some voice types at lower intensity levels (the system was highly reliable for the other commands). Due to technical limitations, participants often had to say "acknowledge warning" twice before it was recognized; this inflated RTs. An additional analysis was conducted to determine how RTs would have changed had the system been perfectly reliable. Results of this analysis are reported in Supplement A and are similar to those reported here. The effect of task load was not significant in either analysis (ps > .26).

Communication Task

The main effects of modality and task load on RT were not significant (ps >.21). There was a significant main effect of modality on percentage correct, $F(2, 34) = 5.32$, p = .010, η^2_p = .24; accuracy was lower in voice trials than in touch (p = .029) or multimodal (p = .016) trials. The effect of task load on percentage correct was not significant (p > .57).

Multimodal Strategies

As previously noted, participants were permitted to adopt the control strategy that they felt was most efficient with each modality, and there

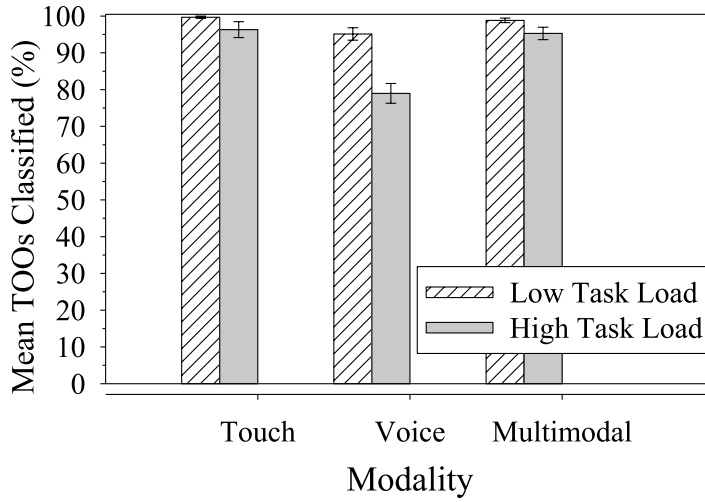

Figure 2. The effects of modality and task load on the mean percentage of TOOs classified per trial. Error bars represent ±1 standard error of the mean.

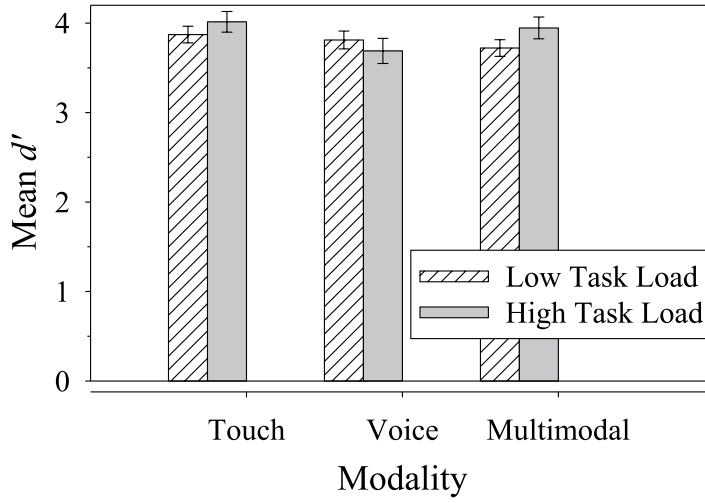

Figure 3. The effects of modality and task load on mean *d*′. Error bars represent ±1 standard error of the mean.

was considerable variability in the proportion of touch vs. voice inputs participants employed in multimodal trials. Overall, the percentage of touch inputs that participants employed in multimodal trials ranged from 14.33% to 100% (M = 73.05%, SD = 28.76%). The total numbers of touch and voice inputs used for each command (across all participants) are presented in Table 3. There was a significant positive correlation between the overall percentage of touch inputs that individual participants employed across multimodal trials and the overall percentage of TOOs classified in such trials, $r(16) = .67$, $p = .002$, indicating that touch was the more efficient modality.

Individual Difference Factors

ACS. Participants' mean ACS scores did not significantly correlate with any of the

TABLE 3: Total Touch and Voice Inputs for Each Command Across All Participants and Trials in the Multimodal Condition

Input	Gray Eagle 1	Gray Eagle 11	Gray Eagle 23	Show Previous	Show Next	Classify Neutral	Classify Hostile	Acknowledge Warning	Total
Touch	415 (68.03%)	366 (66.91%)	373 (68.69%)	76 (67.86%)	803 (73.27%)	1167 (79.39%)	1088 (78.05%)	113 (74.34%)	4401 (74.29%)
Voice	195 (31.97%)	181 (33.09%)	170 (31.31%)	36 (32.14%)	293 (26.73%)	303 (20.61%)	306 (21.95%)	39 (25.66%)	1523 (25.71%)
Total	610 (100%)	547 (100%)	543 (100%)	112 (100%)	1096 (100%)	1470 (100%)	1394 (100%)	152 (100%)	5924 (100%)

performance measures ($-.14 < r < .22$; all ps $> .38$). There was a marginally significant positive correlation between mean ACS scores and mean composite 3D-SART scores, $r(16) = .47$, $p = .051$, with those participants who reported greater attentional control tending to report greater SA. Moreover, there was a significant negative correlation between mean ACS scores and mean global NASA-RTLX scores, $r(16) = -.48$, $p = .045$; participants who reported greater attentional control tended to report experiencing lower levels of workload.

rOPT. Participants with greater spatial ability (as measured by the rOPT) tended to classify more TOOs, $r(16) = 0.75$, $p < .001$, and exhibit shorter RTs to the power pod warning, $r(16) = -.58$, $p = .011$. Scores on the rOPT were not correlated with mean d', $r(16) = .07$, $p = .768$, mean β, $r(16) = .03$, $p = .907$, or the communication task measures (RT: $r[16] = -.38$, $p = .119$; percentage correct: $r[16] = -.11$, $p = .656$). Moreover, rOPT scores were not correlated with mean global NASA-RTLX, $r(16) = -.36$, $p = .139$, or mean composite 3D-SART, $r(16) = .33$, $p = .183$, scores.

DISCUSSION

The purpose of this study was to compare three interface input methods for a MUM-T supervisory control system designed for an AMC located in a Black Hawk helicopter. Our hypothesis that multimodal input would lead to greater task performance than touch or voice input was only partially supported: multimodal input was superior to voice, but did not result in significant performance advantages compared with touch alone. Participants were more efficient with touch and multimodal control (compared with voice), as evidenced by more classified TOOs and shorter RTs to the power pod warning. Participants also demonstrated greater sensitivity classifying TOOs when using touch and multimodal input in high task load conditions. The lower sensitivity during voice trials was likely due to the greater temporal demand with this input method, which may have forced participants to spend less time inspecting each photograph. The observed limitations of voice control are consistent with Baber et al. (2011), who reported that participants performed more poorly with voice control than with gamepad or multimodal (voice and gamepad) control in a simulated unmanned vehicle supervisory control task that involved controlling sensors, classifying targets, and responding to warnings.

One reason that multimodal input did not outperform touch input in the current study is that voice input resulted in greater workload and poorer performance than touch, ostensibly leading participants to rely on touch more frequently when both modalities were available. Participants used touch input for 74.29% of their interactions with the system during multimodal trials. This pattern of strategies would result in relatively comparable performance for the touch and multimodal conditions. As has been observed in other contexts, making multiple input methods available does not mean that participants will make use of all of them: For example, when participants managing unmanned vehicles in a multimodal condition were given voice and gamepad input to control a camera, classify targets, and respond to warnings, voice input was only used to issue around 10% of the

total commands, with three of 17 participants tending to use only the gamepad (Baber et al., 2011).

Our hypothesis that the superiority of multimodal input would be more pronounced with increased task difficulty was also only partially supported. The significant interactions between modality and task load observed for the percentage of TOOs classified, and sensitivity when classifying TOOs (d'), indicate that the advantage of multimodal input relative to voice input was indeed more pronounced with increased task difficulty. However, contrary to predictions, touch input performed comparably to multimodal input in both low and high task load conditions, suggesting that the added flexibility of multimodal input did not outweigh the apparent efficiency advantages afforded by touchscreen control.

Our hypothesis that voice input would interfere with the communication task more than touch or multimodal input was supported for accuracy but not RT. Participants responded correctly to fewer communication task queries when using voice input, ostensibly due to greater overlap in the resources (codes of processing and response modalities; Wickens, 2008) required by the participants' tasks. Although the requirement to perform the auditory/vocal communication task may be seen as biasing the results in favor of touch and multimodal input, discussions with SMEs indicated that this form of communication was critical to the development of SA and was likely to play a continued role in future operations. This is similar to the advantages that civil aviation pilots cite in being able to overhear "party line" information in the form of communications made between ATC and other aircraft (Midkiff & Hansman, 1993).

The superiority of touch input in the current study was partially due to the nature of the tasks. Monitoring the top-down map and inspecting the TOO photographs were both visual/spatial, rather than verbal/symbolic, tasks. Evidence suggests that visual/spatial tasks have greater SCR compatibility with manual (as opposed to verbal) responses (e.g., Wickens et al., 1984). Moreover, touch input allowed participants in our study to distribute their workload across separate information processing channels. Participants could interact with the displays using their eyes and hands while dedicating auditory-vocal resources to the communication task. This pairing of a visual/spatial/manual task (interface control) with an auditory/verbal/vocal task (communications) likely provided a greater opportunity for concurrent processing (Wickens & Carswell, 2012).

Our hypothesis that participants with greater attentional control and spatial ability would exhibit greater task performance was only partially supported. Participants' perceived attentional control (as reported on the ACS) did not correlate with any of the performance measures that we collected. However, participants with greater perceived attentional control did tend to report lower levels of subjective workload (i.e., mean global NASA-RTLX scores). This finding is consistent with Chen and Joyner (2009), who found a negative correlation between ACS and NASA-TLX scores of participants who performed a set of gunner and robotics operator tasks in a simulated mounted crew station environment.

The spatial abilities (as measured by the rOPT) of participants in the current study were correlated with their TOO classification performance (percentage of classified TOOs) and RTs to the power pod warning. These findings are consistent with studies showing that individuals with greater spatial ability exhibit greater performance on tasks involved in the operation of unmanned systems (see Cahillane et al., 2012 for a review). Our findings suggest that the rOPT could potentially be employed as a tool to predict an individual's performance with UAV supervisory control systems, and thus inform operator selection.

Limitations

This study has important limitations that potentially reduce its generalizability to the operational setting. First, environmental factors such as vibration, glare, and cockpit noise could reduce the effectiveness of touchscreen displays and voice recognition systems (Cockburn et al., 2017; Noyes & Haas, 2010), and these were not simulated in our study. Additionally, our system did not use features that make voice control especially useful, such as the consolidation

of multiple touch inputs into a single "voice macro" (Draper et al., 2003). Finally, because our evaluation involved the comparison of a particular touchscreen display to a particular voice recognition system/vocabulary, our results do not necessarily generalize to all such implementations of these technologies (see Supplement A for additional discussion of this topic).

CONCLUSIONS & IMPLICATIONS

The purpose of this study was to evaluate the effects of touch, voice, and multimodal input for a MUM-T supervisory control system designed for an AMC located in a Black Hawk helicopter. We found that, compared with voice, touch and multimodal input resulted in greater task performance, lower subjective workload, and greater subjective SA. Moreover, participants with higher spatial ability exhibited better performance classifying aerial photographs and monitoring instrument warnings.

The strategies that participants in our study employed have implications for theories of human information processing. Across multimodal trials, participants used voice input for 32.12% of their commands to call up a UAV's sensor feed but only 21.26% of TOO classifications. This disparity may have been driven partially by differences in the SCR compatibility of various pairings of task and response type (Wickens et al., 1984). The TOO classification task had a significant visuospatial component, requiring rapid visual search of a complex scene, and consequently may have been better paired with a manual response. Pulling up a UAV's sensor feed also had a visual component (detecting the flashing UAV icon), but the numerical labels on the icons may have invoked symbolic coding of the UAV's name, lending the task to relatively superior pairing with a vocal response. An alternative (and potentially complementary) interpretation of the tendency for participants to use voice more often for selecting UAVs than classifying TOOs is that they perceived this distribution of tasks to modalities as leading to the fastest outcomes. Because TOO classification required fine detail perception, it may have been seen by some participants as more efficient to orient oneself toward the secondary display to inspect photographs and simply make glances to the primary display to monitor for newly detected TOOs and use voice commands to call up the appropriate UAV's sensor feed. These participants may have seen the use of voice input as faster than rotating their trunk and moving one of their arms to select the given UAV's icon. This would be consistent with the soft constraints hypothesis, which proposes that during interactive behavior people allocate perceptual, cognitive, and motor resources on the basis of which relative allocation will result in the fastest response (Gray, Sims, Fu, & Schoelles, 2006).

Our results also have practical implications for voice-controlled systems. For example, there has been a recent emphasis on integrating voice control into the aircraft cockpit (Thomas, Biswas, & Langdon, 2015). Although voice control is often considered a more natural and less physically demanding input method, our results suggest that caution is needed when designing interfaces for users who share common communication channels (e.g., ATC frequencies in today's airliner cockpit). Designers should carefully consider the potential performance tradeoffs when selecting from candidate input methods during system development.

Additionally, our results provide further support for Oviatt's (1999) number one myth about multimodal interaction: "If you build a multimodal system, users will interact multimodally." Although we expected participants to use a relatively equal mix of touch and voice input in multimodal trials, four participants (22%) used exclusively touch and another six (33%) used over 75% touch control. These preferences were seemingly driven by the greater efficiency of touch input, but they may also have been influenced by sample demographics; for example, participants were relatively young and quite comfortable with touchscreen technology. In a postexperiment questionnaire, participants reported being significantly more comfortable with touchscreen displays than voice recognition software ($p < .001$). With the growing ubiquity of voice recognition capability in products such as smartphones and smart speakers, this disparity may very well shrink in the near future.

ACKNOWLEDGMENTS

This work was supported by a grant from General Electric. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of General Electric. We thank Alex Carroll, Maria Teresa Manrique, Sundar Murugappan, and Kris Thibault for their help with software design and development, Adam Braly for his technical support with apparatus assembly, and Erika De Los Santos, Kevin Silva, July Sosebee, Leslie Turner, Maranda Vasquez, and Hannah Woodlee for their help with data coding and analysis.

KEY POINTS

- We evaluated the effects of touch, voice, and multimodal control for a manned-unmanned teaming supervisory control system to be used by an Air Mission Commander located in a Black Hawk helicopter.
- Participants supervised a (simulated) team of two helicopters and three unmanned aerial vehicles and were responsible for classifying aerial photographs, monitoring instrument warnings, and responding to radio communications.
- Compared with voice, touch and multimodal control resulted in greater task performance, lower subjective workload, and greater subjective situation awareness.
- Participants with higher spatial ability exhibited better performance classifying aerial photographs and monitoring instrument warnings.

SUPPLEMENTAL MATERIAL

Supplemental material for this article is available with the manuscript on the *HF* website.

ORCID ID

Patricia R. DeLucia https://orcid.org/0000-0002-1735-9154

REFERENCES

Baber, C., Morin, C., Parekh, M., Cahillane, M., & Houghton, R. J. (2011). Multimodal control of sensors on multiple simulated unmanned vehicles. *Ergonomics*, *54*, 792–805.

Byers, J. C., Bittner, A. C., Jr., & Hill, S. G. (1989). Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary? In A. Mital (Ed.), *Advances in industrial ergonomics and safety I* (pp. 481–485). Philadelphia, PA: Taylor and Francis.

Cahillane, M., Baber, C., & Morin, C. (2012). Human factors in UAV. In P. Angelov (Ed.), *Sense and avoid in UAS: Research and applications* (pp. 119–142). West Sussex, UK: Wiley.

Calhoun, G. L., & Draper, M. H. (2006). Multi-sensory interfaces for remotely operated vehicles. In N. J. Cooke, H. L. Pringle, H. K. Pedersen, & O. Connor (Eds.), *Advances in human performance and cognitive engineering research* (Vol. 7, pp.149–163). Oxford, U.K.: Elsevier.

Calhoun, G. L., & Draper, M. H. (2015). Display and control concepts for multi-UAV applications. In K. P. Valavanis & G. J. Vachtsevanos (Eds.), *Handbook of unmanned aerial vehicles* (pp. 2443–2473). The Netherlands: Springer.

Calhoun, G. L., Ruff, H. A., Behymer, K. J., & Rothwell, C. D. (2017). Evaluation of interface modality for control of multiple unmanned vehicles. *International Conference on Engineering Psychology and Cognitive Ergonomics*, *10276*, 15–34.

Chen, F. (2006). *Designing human interface in speech technology*. New York, NY: Springer.

Chen, J. Y., & Barnes, M. J. (2012a). Supervisory control of multiple robots in dynamic tasking environments. *Ergonomics*, *55*, 1043–1058.

Chen, J. Y., & Barnes, M. J. (2012b). Supervisory control of multiple robots: Effects of imperfect automation and individual differences. *Human Factors*, *54*, 157–174.

Chen, J. Y., & Joyner, C. T. (2009). Concurrent performance of gunner's and robotics operator's tasks in a multitasking environment. *Military Psychology*, *21*, 98–113.

Cockburn, A., Gutwin, C., Palanque, P., Deleris, Y., Trask, C., Coveney, A., . . . MacLean, K. (2017). Turbulent touch: Touchscreen input for cockpit flight displays. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 6742–6753. doi:10.1145/3025453.3025584

Department of Defense. (2010). *Eyes of the army: U.S. Army roadmap for unmanned aircraft systems 2010–2035*. Fort Rucker, AL: U.S. Army UAS Center of Excellence.

Derryberry, D., & Reed, M. A. (2002). Anxiety-related attentional biases and their regulation by attentional control. *Journal of Abnormal Psychology*, *111*, 225–236.

Draper, M., Calhoun, G., Ruff, H., Williamson, D., & Barry, T. (2003). Manual versus speech input for unmanned aerial vehicle control station operations. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *47*, 109–113.

Gray, W. D., Sims, C. R., Fu, W. T., & Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review*, *113*, 461–482.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, *32*, 175–191.

Jones, D. G. (2000). Subjective measures of situation awareness. In M. R. Endsley & D. J. Garland (Eds.), *Situation awareness: Analysis and measurement* (pp. 101–114). Mahwah, NJ: Erlbaum.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, *7*, 57–78.

McKinley, R. A., McIntire, L. K., & Funke, M. A. (2011). Operator selection for unmanned aerial systems: Comparing video game players and pilots. *Aviation, Space, and Environmental Medicine*, *82*, 635–642.

Midkiff, A. H., & Hansman, R. J., Jr. (1993). Identification of important "party line" informational elements and the implications for situational awareness in the datalink environment. *Air Traffic Control Quarterly*, *1*, 5–30.

Noyes, J. M., & Haas, E. (2010). Military applications: Human factors aspects of speech-based systems. In F. Chen & K. Jokinen (Eds.), *Speech technology: Theory and applications* (pp. 251–269). New York, NY: Springer.

Oviatt, S. (1999). Ten myths of multimodal interaction. *Communications of the AMC*, *42*, 74–81.

Proctor, R. W., & Vu, K. P. L. (2016). Principles for designing interfaces compatible with human information processing. *International Journal of Human-Computer Interaction*, *32*, 2–22.

Taylor, G., Purman, B., Schermerhorn, P., Garcia-Sampedro, G., Hubal, R., Crabtree, K., . . . Spriggs, S. (2015). Multi-modal interaction for UAS control. *Proceedings of SPIE 9468, Unmanned Systems Technology 17*, 946802. doi:10.1117/12.2180020.

Thomas, P., Biswas, P., & Langdon, P. (2015). State-of-the-art and future concepts for interaction in aircraft cockpits. *Proceedings of Universal Access in Human-Computer Interaction: Access to Interaction*, *9*, 538–549.

Turk, M. (2014). Multimodal interaction: A review. *Pattern Recognition Letters*, *36*, 189–195.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, *50*, 449–455.

Wickens, C. D., & Carswell, C. M. (2012). Information processing. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (4th ed., pp. 117–161). Hoboken, NJ: Wiley.

Wickens, C. D., & McCarley, J. S. (2008). *Applied attention theory*. Boca Raton, FL: CRC Press.

Wickens, C. D., Vidulich, M., & Sandry-Garza, D. (1984). Principles of S-C-R compatibility with spatial and verbal tasks: The role of display-control location and voice-interactive display-control interfacing. *Human Factors*, *26*, 533–543.

Samuel J. Levulis earned his PhD in Experimental Psychology from the Human Factors Psychology Program at Texas Tech University in May, 2018.

Patricia R. DeLucia is a professor in the Psychology Department at Rice University. She is a Fellow of the American Psychological Association, Association for Psychological Science, Human Factors and Ergonomics Society, and Psychonomic Society. She completed her PhD in Experimental Psychology at Columbia University in 1989.

So Young Kim is a senior user experience lead at the NASA Jet Propulsion Laboratory. She completed her PhD in aerospace engineering from Georgia Institute of Technology in 2011.

Date received: June 20, 2017
Date accepted: June 13, 2018