EMBRY-RIDDLE Aeronautical University
SCHOLARLY COMMONS

International Journal of Aviation, Aeronautics, and Aerospace

Volume 6 | Issue 1 Article 7

2019

Speech Interfaces and Pilot Performance: A Meta-Analysis

Kenneth A. Ward
Embry-Riddle Aeronautical University, wardk1@my.erau.edu

Follow this and additional works at: https://commons.erau.edu/ijaaa

 Part of the Artificial Intelligence and Robotics Commons, Aviation Commons, Cognitive Psychology Commons, and the Graphics and Human Computer Interfaces Commons

Scholarly Commons Citation

Ward, K. A. (2019). Speech Interfaces and Pilot Performance: A Meta-Analysis. *International Journal of Aviation, Aeronautics, and Aerospace, 6*(1). https://doi.org/10.15394/ijaaa.2019.1305

This Article is brought to you for free and open access by the Journals at Scholarly Commons. It has been accepted for inclusion in International Journal of Aviation, Aeronautics, and Aerospace by an authorized administrator of Scholarly Commons. For more information, please contact commons@erau.edu.

Introduction

Advances in technology and automation have led to a gradual decline in the number of crewmembers employed in commercial airliner flight decks. Aircraft manufacturers incorporating these new technologies eliminated the positions of radio operators, navigators, and flight engineers by the 1980s, resulting in the current two-pilot model. The commercial aviation industry is poised to make yet another reduction as the aviation industry contemplates the concept of single-pilot operations as the next logical step.

These new technologies are developing at just the right time; as air transport routes expand globally, an industry-wide shortage of pilots persists. Boeing (2018) forecasts a global requirement for over 790,000 new pilots by the year 2037 in order to meet the demand. Reducing crew requirements to single-pilot operations presents a means to alleviate the demand but introduces new technological and human factors challenges.

Lim, Bassien-Capsa, Ramasamy, Liu, and Sabatini (2017) described managing and distributing workload, maintaining pilot situational awareness, and interface design as some of the key challenges to implementing single-pilot operations. Bilimoria, Johnson, and Schutte (2014) further detailed the need for automation to change between tasks and roles without being "rigidly prescribed" (p. 6) and function much as an active crewmember. Conceptually, these challenges illuminate the necessity to simplify the user interface, facilitate coordination between the pilot and automation, and simultaneously increase the extent and complexity of tasks to be automated.

Statement of the Problem

Speech interfaces present a novel opportunity to address the emerging requirements of automation in the single-pilot operation environment. Speech is a simple and intuitive method of interacting with a system, as the interaction is limited by the ability of the system to recognize and interpret the input, rather than by the finite space of controls on an instrument panel. The further step of interpreting natural, spoken words, exemplified throughout the U.S. population in digital assistants in smart phones and smart home devices, such as Apple's Siri and Amazon's Alexa, demonstrated the possibilities of using speech interfaces in existing technology to simplify the interface to complex tasks.

While speech interfaces present opportunities to reduce pilot workload overall, they are still considered an emerging technology, especially in aviation. Consequently, there is little research describing what effects such systems can have in the flight deck and in human performance. With increasing automation in the single pilot environment, some form of simplified interface will be required; how speech interfaces compare to traditional mechanical or touch screen interfaces in the flight deck remains unknown.

Research Questions

This research sought to determine what effects the use of speech interfaces with automation have on primary task performance, compared to traditional manual interfaces.

RQ1: Does use of a speech interface change user workload rating compared to use of traditional automation?

RQ2: Does use of a speech interface change user attention to primary task compared to use of traditional automation?

RQ3: Do the number of errors made differ when using a speech interface compared to a traditional interface?

RQ4: Is the time to complete an interaction using a speech interface different from that of traditional automation?

Literature Review

Speech Interfaces

Despite the simplification of the interface, speech interfaces introduce new human factors challenges that may affect performance by other means. One unique feature of such an interface is the transition of the system to a more overt social actor. Nass and Lee (2001) described a wide body of work supporting the "Computer as a Social Actor" (CASA) theory and demonstrated that humans ascribe personality to computers in text-to-speech applications. Knott and Kortum (2006) found that intentional personification of an automated system, through assigning a name and the virtual actor through spoken dialog, affected users' engagement with the system.

The system anthropomorphization did not stop at perceptions of the system; in automation studies with speech input, users altered the way they interacted with the system to include emotion and social niceties. One study of such a system in a driving simulator saw operators employ politeness in response to the system requesting input, and praising and thanking it in response to confirmation of simple tasks such as setting the radio (Large, Clark, Quandt, Burnett, & Skrypchuk, 2017). One can assume it is a comparably rare occasion in which an airline pilot says "thank you" to the traditional knob-and-indicator autopilot for reaching an assigned altitude.

While the implications of personifying automated systems are vast and represent a fascinating avenue for future study, the present research is concerned with how introducing such a system may affect pilot performance. Past studies have shown that the perceived attributes of automated systems and the user's mood (Nass et al., 2005) or personality (Knott & Kortum, 2006) can affect user performance in different roles.

Furthermore, despite social behavior entering the automation interface, systems accepting speech input are emphatically not human or truly sentient

systems. They still retain minimal ability to process commands outside their domain, rely upon clear input, and can frustrate users with responses if the input is unclear or framed incorrectly. Such problems can increase user workload or increase errors.

The cognitive effects of speaking while performing other tasks can potentially affect pilot performance when using speech interfaces. Spence, Jia, Feng, Elserafi, and Zhao (2013) reported in a literature review that speaking uses finite cognitive resources and reduces visual attention. It is important to clarify and reiterate that Spence and colleagues (2013) stated that that the act of speaking itself, regardless of task relation, reduced attention, field of view, and reaction time. Thus, speech interface use in aviation may affect performance differently depending on when and how the system is used.

Assessing Single-Pilot Performance

The measures of pilot performance in the era of single-pilot operations remain nebulous, as the operating concept is still in its infancy. Instead, one can examine the current two-pilot flight deck model and identify other key performance tasks involving aircraft management and automation monitoring. The FAA (2017) defined the roles of pilots in a two-pilot operation as the Pilot Flying (PF) and Pilot Monitoring (PM). In design, the PF is responsible for physically flying the aircraft and managing the autopilot, while the PM is responsible for monitoring systems. In the single-pilot construct, the pilot likely fulfills parts of both roles (Billimoria et al., 2014). The FAA describes the characteristics of effective PM duties as including communicating deviations to crewmembers, managing distractions, and remaining vigilant. Liu, Gardi, Ramasamy, Lim, and Sabatini (2016) described several responsibilities for a single pilot, which, in broad terms, included monitoring the environment, manually flying the aircraft, managing and monitoring systems, and communicating with air traffic control. Aside from the addition of manual control, these concepts align with the FAA's description of modern PM duties.

The FAA succinctly assessed that "high workload, distraction, and inattention can all lead to monitoring errors" (FAA, 2017, p. 6-2). Notably, these are described in terms of performance effects on the pilot's primary task. For example, an altitude deviation while entering a new route into the flight management computer is an example of inattention to the primary task.

Measuring workload accurately across studies may be difficult. De Waard and Lewis-Evans (2014) argued that workload self-assessments are not contemporaneous with the work undertaken and workload cannot be experimentally manipulated during the measurement. Therefore such assessments may instead be measuring perceptions of performance. De Winter (2014) stated that such constructs should be augmented with other sources of information if possible, but such constructs are still useful for prediction. As workload is inherently

subjective and depends on the definitions used, workload is analyzed here alongside other measures of performance and assessed in terms of standard effect sizes that are comparable across measures.

A systematic review of studies of speech interfaces revealed a wide range of literature. While there is some research in the aviation domain, the much of the recent work in speech interfaces has been conducted in automotive studies. When comparing to broad concepts such as inattention and workload, some can be used as an analogs to pilots' duties. Notably, several studies tested GPS navigation entry, which serves as a stand-in here for pilots entering a flight route. Similarly, phone dialing or vehicle radio tuning represent the number-sequence entry of changing aircraft radio frequencies. Of particular note, many studies continued the phone dialing action to study phone conversations while performing a primary task; these were not used here as an analog for pilot duties, as a conversation ceases to be a function of interface interaction and becomes an enduring secondary task.

There are several relevant meta-analyses and literature reviews of voice input systems in automobiles that include a wide range of voice tasks, including phone conversations (Barón & Green, 2006; Simmons, Caird, & Steel, 2017). The present research shares some references to underlying studies but differs in inclusion criteria by including tasks related to interacting with automated systems and only tasks analogous to what an aircraft pilot may be expected to perform.

Methodology

This study employed a random effects meta-analysis of relevant research, as the populations and methods vary between sources. Analysis was completed with the *Meta-Essentials* analysis tool (Suurmond, van Rhee, & Hak, 2017). Standard effect sizes of participant performance for each performance category are used as reported (as available) or computed from available data and assessed at 95% confidence intervals.

Population

As the present research discusses the implications of a future trend, the commercial aviation industry currently does not employ speech interfaces. Accordingly, there are limited studies regarding such interfaces using the ideal population of airline pilots. The FAA's (2018) report of U.S. Civil Airmen Statistics was used to understand the demographics of airline pilots by examining the qualities of pilots holding an active Airline Transport Pilot (ATP) certificate.

Available information indicated U.S. ATP certificate holders are all 20 years or older (mean = 50.6) and range to over 80 years old (no upper limit specified). All studies included in the meta-analysis have participants aged 20 years or more. Gender was typically evenly divided in the included studies, and no studies reported differences in performance based on gender. Other demographic information, such as race and ethnicity, were neither included in the FAA

demographics nor the studies accessed. Expansion to the general population over 20 years old was further justified by the fact that the constructs measured in the research relate to human factors rather than piloting abilities or aviation-specific knowledge. Nonetheless, the researcher acknowledges the limitation and the potential for unforeseen and unique implications given the broadening of the research population.

Variables

The independent variable is the use of automation with a speech interface to assist with a primary task such as driving. The dependent variables are measures of performance of the primary task. It is important to reiterate that the present research does not include studies in which a speech interface is used to accomplish a task secondary to that which is automated, which ensures that the performance measured is related to the use of the interface, rather than a function of distraction.

While the narrow definition of the independent variable has the unfortunate consequence of ruling out much of the recent body of literature, it does ensure the studies that remain in the present research are more aligned with the concept of a single pilot using automation to support the flight task. While many distracted driving studies are largely excluded here, most remaining studies still do take place in the driving environment, as it presents a well-defined primary task that can be supported by automation. Such instances include speech input for navigation, radio tuning, and phone dialing as they support the primary task and are representative of tasks a single pilot must accomplish.

The dependent variables used are measured differently throughout the literature, but many concern workload, errors, distraction, and time to complete a task. Terms used are coded here so that in all cases a higher value indicates worse performance: high workload, more errors, more missed cues, and longer task times. All reported results are directionally presented as speech interfaces as compared to manual interfaces; a positive effect indicates worse performance in the voice input condition.

Sampling Strategy

The researcher conducted a search for relevant literature in Embry-Riddle Aeronautical University's Hunt Library databases, which included citation indices from ProQuest, Taylor and Francis, and Sage Journals among others. The initial search used the phrase "(voice or speech or language) and (workload or attention or distraction or error)" and was limited to scholarly or peer reviewed sources. The initial search yielded 1,816 results and was narrowed by scanning the titles and abstracts for those that may be applicable to the present study. Studies selected for review were read in full, sorted by the inclusion criteria, and the reference sections were scanned for additional sources to review. Those additional reference sections yielded new search terms and studies, and a snowball method was used to continue

expanding searches through the university, Google Scholar, and the broader internet until no new sources arose in searches.

Studies selected for meta-analysis were required to be experimental design, peer reviewed or scholarly, and original research with quantitative data. The method of the research was required to measure an analog to tasks performed by pilots, and the voice interface method must include natural language (i.e. more than merely single word prompts). Many books, systematic reviews, reports, and meta-analyses were reviewed for background information and additional references but were not used in the quantitative analysis here.

Additionally, studies focusing on the technical aspects of speech interfaces, input languages other than English, or non-native English speakers were excluded. While this limits the present research to domestic aviation applications, it allows the research to focus on the effects of speech input by controlling for technical limitations of speech interface systems.

Imputation of Missing Data

No studies included in the meta-analysis reported correlation coefficients, which were required by the analysis software for the quantitative comparison of within-subjects data. Correlation coefficient was estimated by calculating the standard effect size of each study, using Cohen's *d*. The estimated correlation coefficient was then calculated using the following formula.

$$r = \frac{d}{\sqrt{d^2+4}} \quad (1)$$

A frequent problem with meta-analyses is that the underlying studies do not reliably report standard deviations (Furukawa, Barbui, Cipriani, Brambilla, & Watanabe, 2006). In cases where the studies provided graphs, but no exact data, standard error was estimated by closely inspecting the images and counting pixels between the scale bars, whiskers, and graph axes to reach as close an estimate as possible. Standard deviation was then calculated by multiplying the standard error of the mean by the square root of the sample size.

Some studies provided only mean values, and did not include standard deviation, standard error, or confidence intervals. Ma, Liu, Hunter, and Zhang (2008) recommended researchers use their "prognostic method" to predict missing standard error of mean (SEM) values; their method uses Error Theory, but weighted by each study's sample size, to estimate SEM. Conversions between SEM and standard deviation make Ma and colleagues' method functionally equivalent to averaging the standard deviations of similar studies. Thus, the mean standard deviation for both control and treatment from studies within the same analysis groupings (e.g. workload or time to complete task) and sub-groups (e.g. navigation entry or radio tuning) were used to impute missing standard deviation values in cases where sufficient data were otherwise unavailable.

Furukawa et al. (2006) found that averaging standard deviations from similar data is an acceptable approach to estimate missing standard deviations. Given that the objective of a meta-analysis is to include the body of relevant literature, discarding studies from the analysis for lack of complete data violates inclusivity, and one should err on the side of inclusion rather than exclusion. However, while the method is sound, one must acknowledge that estimation does reduce the credibility of the final analysis (Furukawa et al., 2006).

Results

Included Studies

Studies meeting the inclusion criteria and using applicable variables are included in the analysis below. The reference sections from each study were searched and reviewed iteratively until the reviews yielded no new sources. Of the 1,816 studies found in the original search results and those found in other studies' reference sections, 133 studies were subject to a detailed review. Of those, 37 were irrelevant to the present study, 1 was not available in English, 7 were not original research, 19 were not from peer reviewed or scholarly sources, 24 did not use applicable variables, 5 published no quantifiable data, and 24 studies were not focused on automation interactions supporting a primary task. Finally, 16 studies met inclusion criteria for the meta-analysis (see Table 1). An asterisk precedes the listing for each source included in the analysis in the reference section.

Table 1
Summary of Included Studies and Measures of Performance

Study	N	Attention	Workload	Errors	Task Time
Beckers et al., 2017	24	V	V		V
Carter & Graham, 2000	32	V	V		M
Gärtner et al., 2002	16			I	M
Gellatly & Dingus, 1998	12				M
Harbluk et al., 2007	16	I			M
Jenness et al., 2002	24	V		M	
Maciej & Vollrath, 2009	30	V			
Mazzae et al., 2004	54		V		M
McWilliams et al., 2015	40	V	V		V
Mountford & North, 1980	10			V	V
Munger et al., 2014	30		V		
Noyes & Starr, 2007	16			V	V
Owens et al., 2010	21	V	V		V
Schreiner, 2006	12	V			M
Schreiner et al., 2004	37	V			
Tsimhoni 2004	24				V

Note. "V" indicates better performance in the voice interface condition, "M" in manual, and "I" is inconclusive or mixed results. As multiple measures are summarized, no claims to statistical significance are made here.

Attention

Each study that measured attention involved driving as the primary task. There were a variety of measures of participant attention used throughout the relevant literature which fell into two broad categories with sufficient data for analysis: deviations in speed and position and gaze behavior. While the studies included many more measures of attention, no other measures were found frequent enough to warrant meta-analysis.

Six studies analyzed deviations in speed and position, reported in ten categories (k = 10, n = 236). See Figure 1. The speech input condition resulted in significantly lower mean deviation position and speed (d = -1.07, 95% CI [-1.75, -0.38]). Subgroup analysis did not indicate meaningful differences when grouped by type of deviation (speed or position), type of task (radio tuning, navigation entry), or type of manual input (touch screen, buttons). An Egger Regression did not indicate significant publication bias (t = -.51, p = .624).

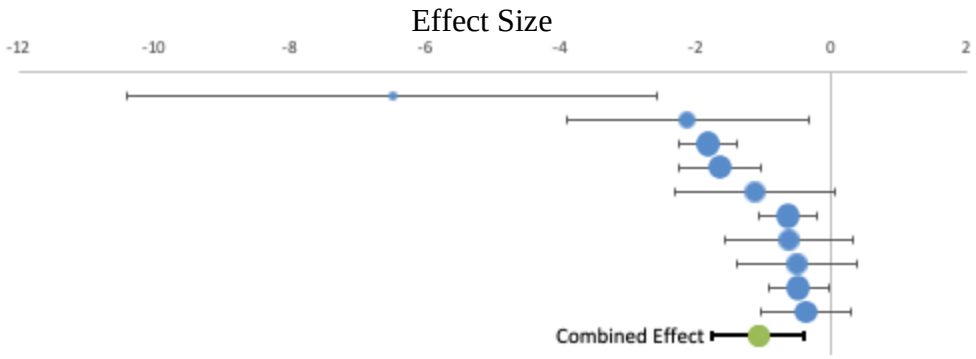

Figure 1. Standard Effect Size (95% CI) of Speed and Position Deviations

Seven studies included a number of different metrics to assess distraction as a function of gaze performance (k = 12, n = 335). The commonly used methods were the number times a participant glanced away from the road and the total time spent looking away. Overall, performance was better when using speech interfaces, as participants focused more intently on the primary task (d = -5.12, 95% CI [-5.74, -4.49]). Figure 2 illustrates how participants using voice interfaces both glanced away from the road less frequently (k = 3, n = 69, d = -4.72, 95% CI [-9.15, -0.30]) and for less total time (k = 9, n = 266, d = -5.32, 95% CI [-8.49, -2.15]). An Egger Regression indicated significant publication bias (t = -8.56, p < .001). *Meta-Essentials* (Suurmond et al., 2017) recommended an adjusted effect size based upon imputed unpublished studies of d = -4.30 with a 95% confidence interval of -7.51 to -1.09, still indicating improved gaze behavior in the voice input condition.

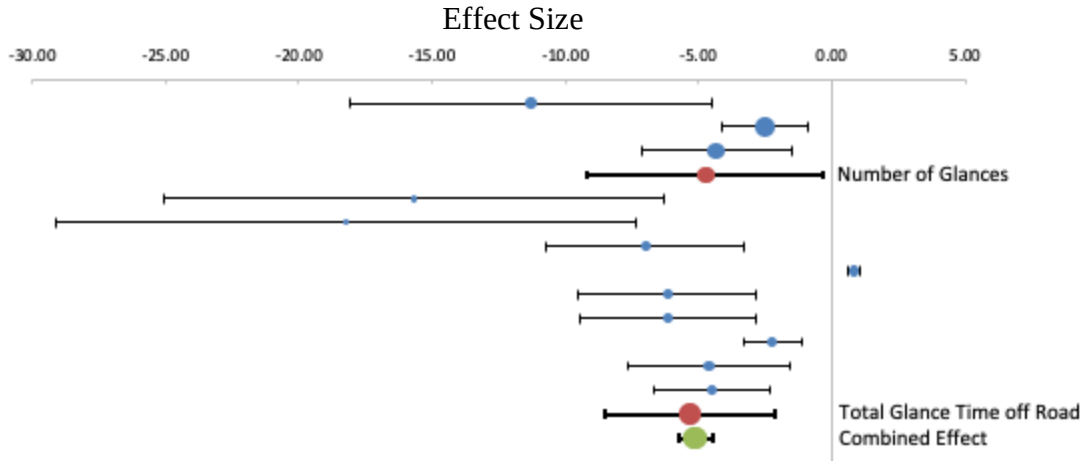

Figure 2. Standard Effect Size (95% CI) of Gaze Performance

Workload

Six studies included subjective, self-reported assessments of workload on Likert-type scales. Of those that reported results of that data ($k = 5$, $n = 169$), participants in the speech input condition reported significantly less workload than in the manual input condition (d = -2.82, 95% CI [-4.48, -1.16]). See Figure 3. The remaining study that did not report the results of the quantitative workload assessment qualitatively agreed, "while driving, the speech control conditions were rated lowest workload" (Carter & Graham, 2000, p. 3-289). An Egger Regression indicated publication bias was significant (t = -12.18, p = 0.001). *Meta-Essentials* (Suurmond et al., 2017) imputed missing unpublished studies and estimated the adjusted effect size, with still significantly less workload in the voice input condition (d = -1.93, 95% CI [-3.83, -0.03]).

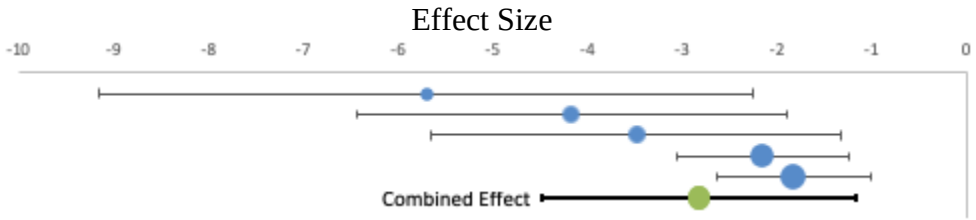

Figure 3. Standard Effect Size (95% CI) of Workload

Errors

Only four articles included in the meta-analysis reported errors associated with an independent variable of speech or manual interface. Each reported a different type of error, precluding quantitative comparison. Qualitatively, in the speech input condition, there were fewer errors in primary driving tasks of maintaining vehicle speed and lane position (Gärtner, König, & Wittig, 2002), tracking an object with a joystick (Mountford & North, 1980), and deviation from a tracking task (Noyes & Starr, 2007). Voice input was associated with more data input errors (Jenness, Lattanzio, O'Toole, Taylor, & Pax, 2002).

Task Time

The time to complete the interaction with the interface was measured across 12 studies, two of which reported data in two categories (k = 14, n = 281). The time taken time to complete a task was not significantly associated with input modality (d = .55, 95% CI [-1.34, 2.66]). There was a high level of heterogeneity ($p < .001$), warranting subgroup analysis. The studies were first categorized by task: number entry and radio tuning, navigation entry, and completing an aviation checklist. Only one study involved the aviation checklist task, precluding further subgroup analysis. Subgroup analyses of task type and task complexity did not indicate significant differences.

First, the subgroup of navigation entry was analyzed (k = 6, n = 108), and the effect of input modality was not significant (d = 1.13, 95% CI [-7.96, 10.21]). See Figure 4. In voice command systems that provided feedback prompts for navigation entry (k = 2, n = 32), time to complete the task was significantly slower when using speech input (d = 4.94, 95% CI [0.67, 9.20]). When navigation entry systems did not prompt entry or provide feedback until the end (k = 4, n = 76), the voice input was significantly faster than manual input (d = -2.15, 95% CI [-3.70, -0.60]). An Egger Regression did not indicate significant publication bias (t = -1.96, p = .121).

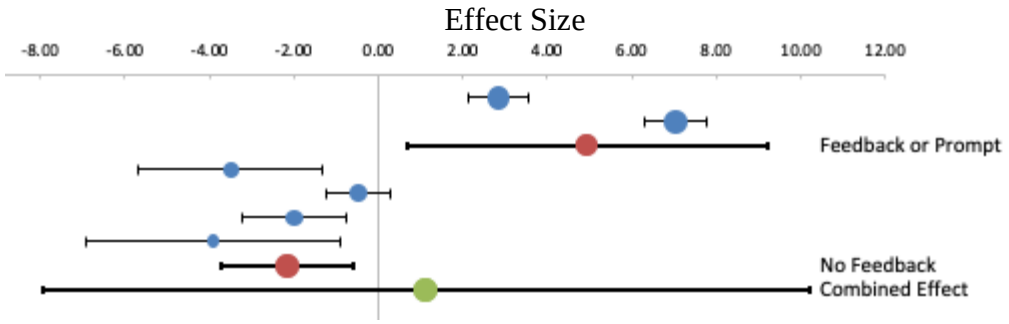

Figure 4. Standard Effect Size (95% CI) of Time to Complete Navigation Data Entry

Radio tuning and number entry were grouped together for analysis (k = 7, n = 157) due the similar nature of the tasks. In some studies, participants tuned radios by using numeric phrases to change a frequency, thereby bridging both categories. There was no significant difference in input modality as the confidence interval included zero (d = 0.97, 95% CI [-2.28, 4.22]). The subgroup analysis hinted at different effects between systems that provided feedback and those that did not, although neither subgroup demonstrated a significant difference at the 95% confidence level (see Figure 5). Finally, an Egger Regression did not indicate publication bias for the radio tuning and number entry tasks (t = -1.99, p = .103).

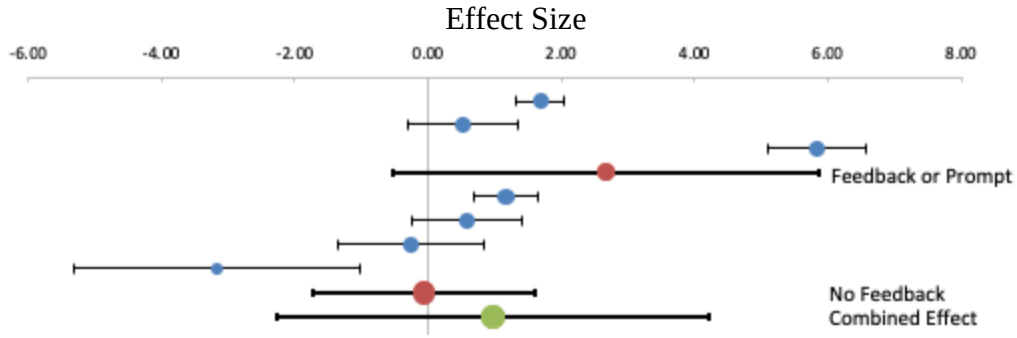

Figure 5. Effects of Time to Complete Number Entry and Radio Tuning Tasks

Moderators

Voice recognition accuracy does affect performance (Gellatly & Dingus, 1998), which may affect underlying results in the studies analyzed. This was partially controlled by selectively using results in some of the underlying studies; in cases where the study manipulated voice recognition accuracy, only the 100% accurate condition was used. Other studies employed a "Wizard of Oz" approach with a researcher simulating the computer voice recognition, without the participant's knowledge, again with 100% accuracy. However, in most studies, voice recognition accuracy was not reported. As voice recognition technologies improve over time, the more recent studies may have more accurate voice recognition systems as a result, also potentially moderating the results.

Voice recognition accuracy and the year of publication may introduce moderating effects (Simmons et al., 2017). In the studies included in the meta-analysis, the year of the study's publication did not significantly moderate input modality and attention when assessing attention by either number of glances away from the road ($F[1, 10] = .14$, $p = .72$) or total time looking away ($F[1, 8] = .15$, $p = .71$). Similarly, there was no significant moderation for workload ($F[1, 3] = .86$, $p = .42$) or time to complete task ($F[1, 11] = 0.03$, $p = .87$).

Participant age may also be a moderator, as indicated by performance differences between age groups in some included studies. There was insufficient data to explore that relationship here. The majority of studies included in the meta-analysis did not report results for separate age groups, preventing meta-analysis of age as either a subgroup or moderator.

Discussion

Speech interfaces may be a valuable tool to assist pilots in single pilot operations. Despite the recent proliferation in speech input technologies and digital assistants, very few studies consider their application in the aviation environment, restricting the meta-analysis to using automotive studies as an analog. Driving similarly requires attention and accuracy to accomplish safely, and tasks such as

entering destinations and tuning radios are similar to tasks performed by pilots interacting with avionics systems.

The available literature indicated that there are fewer vehicle control deviations and fewer glances away from the primary task when using voice input. Additionally, subjective workload was significantly decreased when using speech interfaces. Together, these indicate that speech interfaces may be able to assist pilots with complex system interactions while allowing them to focus on the task of safely flying an aircraft.

The few studies that did measure errors each did so in a different manner, preventing useful meta-analysis of the results. While the research question regarding the effect of interface modality on errors was unable to be addressed with the available data, it is worthwhile to note that in general, studies reported decreases in errors in most cases when using speech interfaces. The notable exception is that there were more input errors when using speech input (Jenness et al., 2002). This itself is worthy of further study, especially considering how input errors may affect highly complex and automated aircraft systems differently than automobile systems..

The time it takes to complete an interaction with a speech system may be affected by the type of system and its capabilities. While relatively short interactions such as radio tuning and number entry did not significantly differ in the time to complete the interaction depending on the input modality, longer interactions presented an interesting finding. On first inspection, the time to direct navigation to a destination did not significantly differ depending on modality. Yet when analyzed in groups, the nature of the interface divided the results. When the system allowed unprompted natural speech input, the interaction was faster when using speech input systems. However, in systems that prompted users to speak or provided feedback, the voice input took longer than manual entry. Given recent advancements in natural language system interfaces, such as those in mobile assistants, it is recommended to design new systems that do not rely on user prompts.

Speech interfaces present opportunities to decrease inattention and workload when interacting with complex automation and performing a safety-critical task. The airline flight deck is characterized by such automation and could benefit from more natural system interfaces that improve pilot performance. While speech interfaces have many benefits that apply to pilots, there is insufficient direct research on the topic in aviation. Future experiments of the performance effects of such interfaces on pilot performance or comparisons of voice systems may provide useful evidence to aid the industry in adopting such a potentially beneficial tool.

References

Barón, A., & Green, P. (2006). *Safety and usability of speech interfaces for in-vehicle tasks while driving : A brief literature review* (Technical Report UMTRI-2006-5). Ann Arbor, MI: Transportation Research Institute. doi:10.1518/001872008X288394

Beckers, N., Schreiner, S., Bertrand, P., Mehler, B., & Reimer, B. (2017). Comparing the demands of destination entry using Google Glass and the Samsung Galaxy S4 during simulated driving. *Applied Ergonomics, 58*, 25-34. doi:10.1016/j.apergo.2016.05.005

Bilimoria, K., Johnson, W., & Schutte, P. (2014). Conceptual framework for single pilot operations. Paper presented at the *International Conference on Human-Computer Interaction in Aerospace*, 1-8. doi:10.1145/2669592.2669647

Boeing. (2018). *Pilot outlook: 2018-2037.* Retrieved from https://www.boeing.com/commercial/market/pilot-technician-outlook/2018-pilot-outlook/

Carter, C., & Graham, R. (2000). Experimental comparison of manual and voice controls for the operation of in-vehicle systems. *Human Factors and Ergonomics Society Annual Meeting Proceedings, 44*(20), 286-289.

de Waard, D., & Lewis-Evans, B. (2014). Self-report scales alone cannot capture mental workload: A reply to De Winter, Controversy in human factors constructs and the explosive use of the NASA TLX: A measurement perspective. *Cognition, Technology and Work*, *16*(3), 303–305. doi:10.1007/s10111-014-0277-z

de Winter, J. C. F. (2014). Controversy in human factors constructs and the explosive use of the NASA-TLX: A measurement perspective. *Cognition, Technology and Work, 16*(3), 289–297. doi:0.1007/s10111-014-0275-1

Federal Aviation Administration. (2017). *Standard operating procedures and pilot monitoring duties for flight deck crewmembers.* (AC 120-71B). Retrieved from https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_120-71B.pdf

Federal Aviation Administration. (2018). *U.S. civil airmen statistics*. Retrieved from https://www.faa.gov/data_research/aviation_data_ statistics/civil_airmen_statistics/

Furukawa, T. A., Barbui, C., Cipriani, A., Brambilla, P., & Watanabe, N. (2006). Imputing missing standard deviations in meta-analyses can provide accurate results. *Journal of Clinical Epidemiology, 59*(1), 7-10. doi:10.1016/j.jclinepi.2005.06.006

Gärtner, U., König, W., & Wittig, T. (2002). Evaluation of manual vs. speech input when using a driver information system in real traffic. Paper presented at the *Driving Assessment Conference,* Iowa City.

Gellatly, A. W., & Dingus, T. A. (1998). Speech recognition and automotive applications: Using speech to perform in-vehicle tasks. *Human Factors and Ergonomics Society Annual Meeting Proceedings, 42*(17), 1247-1251. doi:10.1177/154193129804201715

Harbluk, J. L., Burns, P. C., Lochner, M., & Trbovich, P. L. (2007). Using the lane change test (LCT) to assess distraction: Tests of visual-manual and speech-based operation of navigation system interfaces. Paper presented at the *Proceedings of the Fourth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*.

Jenness, J. W., Lattanzio, R. J., O'Toole, M., Taylor, N., & Pax, C. (2002). Effects of manual versus voice-activated dialing during simulated driving. *Perceptual and Motor Skills, 94*(2), 363-379. doi:10.2466/PMS.94.2.363-379

Knott, B. A., & Kortum, P. (2006). Personification of voice user interfaces: Impacts on user performance. *Human Factors and Ergonomics Society Annual Meeting Proceedings, 50*(4), 599-603. doi:10.1177/154193120605000411

Large, D. R., Clark, L., Quandt, A., Burnett, G., & Skrypchuk, L. (2017). Steering the conversation: A linguistic exploration of natural language interactions with a digital assistant during simulated driving. *Applied Ergonomics, 63*, 53-61. doi:10.1016/j.apergo.2017.04.003

Lim, Y., Bassien-Capsa, V., Ramasamy, S., Liu, J., & Sabatini, R. (2017). Commercial airline single-pilot operations: System design and pathways to certification. *IEEE Aerospace and Electronic Systems Magazine, 32*(7), 4-21. doi:10.1109/MAES.2017.160175

Liu, J., Gardi, A., Ramasamy, S., Lim, Y., & Sabatini, R. (2016). Cognitive pilot-aircraft interface for single-pilot operations. *Knowledge-Based Systems, 112*, 37-53. doi:10.1016/j.knosys.2016.08.031

Ma, J., Liu, W., Hunter, A., & Zhang, W. (2008). Performing meta-analysis with incomplete statistical information in clinical trials. *BMC Medical Research Methodology, 8*(1), 56-56. doi:10.1186/1471-2288-8-56

Maciej, J., & Vollrath, M. (2009). Comparison of manual vs. speech-based interaction with in-vehicle information systems. *Accident Analysis and Prevention, 41*(5), 924-930. doi:10.1016/j.aap.2009.05.007

Mazzae, E. N., Ranney, T. A., Watson, G. S., & Wightman, J. A. (2004). Hand-held or hands-free the effects of wireless phone interface type on phone task performance and driver preference. *Human Factors and Ergonomics Society Annual Meeting Proceedings, 48*(19), 2218-2222. doi:10.1177/154193120404801903

McWilliams, T., Reimer, B., Mehler, B., Dobres, J., & Coughlin, J. F. (2015). Effects of age and smartphone experience on driver behavior during

address entry: A comparison between a Samsung Galaxy and Apple iPhone. Paper presented at the *7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, New York. 150-153. doi:10.1145/2799250.2799275

Mountford, S. J., & North, R. A. (1980). Voice entry for reducing pilot workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *24*(1), 185-189. doi:10.1177/107118138002400153

Munger, D., Mehler, B., Reimer, B., Dobres, J., Pettinato, A., Pugh, B., & Coughlin, J. (2014). (2014). A simulation study examining smartphone destination entry while driving. Paper presented at the *6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '14,* New York. 1-5. doi:10.1145/2667317.2667349

Nass, C., Jonsson, I., Harris, H., Reaves, B., Endo, J., Brave, S., & Takayama, L. (2005). Improving automotive safety by pairing driver emotion and car voice emotion. Paper presented at the *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, Portland, OR. 1973-1976. doi:10.1145/1056808.1057070

Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied, 7*(3), 171-181. doi:10.1037/1076-898X.7.3.171

Noyes, J. M., & Starr, A. F. (2007). A comparison of speech input and touch screen for executing checklists in an avionics application. *The International Journal of Aviation Psychology*, *17*(3), 299-315. doi:10.1080/10508410701462761

Owens, J., McLaughlin, S., & Sudweeks, J. (2010). On-road comparison of driving performance measures when using handheld and voice-control interfaces for mobile phones and portable music players. *International Journal of Passenger Cars - Mechanical Systems, 3*(1), 734-743. doi:10.4271/2010-01-1036

Schreiner, C. S. (2006). The effect of phone interface and dialing method on simulated driving performance and user preference. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, *50*(22), 2359-2363. doi:10.1177/154193120605002202

Schreiner, C., Blanco, M., & Hankey, J. M. (2004). Investigating the effect of performing voice recognition tasks on the detection of forward and peripheral events. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, *48*(19), 2354-2358. doi:10.1177/154193120404801932

Simmons, S. M., Caird, J. K., & Steel, P. (2017). A meta-analysis of in-vehicle and nomadic voice-recognition system interaction and driving performance. *Accident Analysis and Prevention*, *106*(November 2016), 31–43. doi:0.1016/j.aap.2017.05.013

Spence, I., Jia, A., Feng, J., Elserafi, J., & Zhao, Y. (2013). How speech modifies visual attention. *Applied Cognitive Psychology, 27*(5), 633-643. doi:10.1002/acp.2943

Suurmond, R., van Rhee, H., & Hak, T. (2017). Introduction, comparison and validation of Meta-Essentials: A free and simple tool for meta-analysis. *Research Synthesis Methods, 8*(4), 537-553. doi:10.1002/jrsm.1260.

Tsimhoni, O., Smith, D., & Green, P. (2004). Address entry while driving: Speech recognition versus a touch-screen keyboard. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 46*(4), 600-610. doi:10.1518/hfes.46.4.600.56813