Human-Centered Aerospace Systems and Sustainability Applications, Vol. 98, 2023, 101–111
https://doi.org/10.54941/ahfe1003952

Speech-Enhanced and Context Dependent Alerts: Future Implications for Spacecraft Design

Ian Robertson[1], Kritina Holden[2], Ryan A. Lange[3], Durand R. Begault[4], Tyler L. Duke[2], and Ryan Z. Amick[1]

[1]KBR at the NASA Johnson Space Center, Houston, TX USA
[2]Leidos at the NASA Johnson Space Center, Houston, TX USA
[3]Geologics at the NASA Johnson Space Center, Houston, TX USA
[4]NASA Ames Research Center, Moffett Field, CA USA

ABSTRACT

In the future, NASA missions will involve many different space vehicles, habitats, and surface assets working together to provide safe, productive environments for crew. Because these systems will be provided by multiple commercial companies working with NASA, it will be very different from missions of the past, bringing new challenges. One of the challenges is related to whether NASA should move beyond simple tone annunciation alerting systems, to more advanced systems that include speech. The other is related to determining the level of consistency required of safety-critical alert systems across spacecraft. Two studies were completed to address these important issues. The first study investigated the advantages and disadvantages of a tone+speech alert relative to the traditional tone-only alert. Results indicate that speech-enhanced alerts initially take longer to silence (the default action to which NASA personnel are trained), due to the need to listen to the entire message, but ultimately provided for faster understanding of the alert situation. Speech-enhanced alerts were also preferred by a large majority of astronaut-like study participants. An unexpected finding was that participants took longer to respond to tone-only alerts that were heard in the same session as speech-enhanced alerts. Participants waited to hear a speech message even for alerts they were trained to know did not contain speech components. This performance error is believed to be due to negative transfer of training. A second study focused on task and alert performance using a common set of tones across two contexts (e.g., vehicles, habitats, suits) versus performance with a different set of tones for each context. Participants were able to manage two different alert sets successfully; results indicate that discriminability of the two alert sets played a major role in their success. Implications for the design of spacecraft alerts are discussed, and future areas of research are identified.

Keywords: Human systems integration, Alerts, Human-centered design, Alarm design, Alert design

INTRODUCTION

NASA's Artemis program aims to return astronauts to the lunar surface and establish a permanent presence on the moon. In this program, vehicles and

systems are being provided by multiple commercial companies, and currently there are few overarching requirements related to commonality. While there are benefits to this approach (e.g., innovative designs, diversity of ideas), it also may allow for human error due to inconsistency among designs. Alerting system design is of special concern, as it is an important part of the infrastructure of any safety-critical environment.

Whenever multiple independent alerting systems are used together, significant human factors concern arise over consistency, discriminability, and human memory load. Astronauts will be moving from one system/vehicle to another during their day and will have to have some familiarity with all systems to perform their tasks. Even if astronauts are trained as specialists in one system versus another, in an off-nominal situation, any astronaut could be faced with responding to alerts from any of the vehicles. From a human factors point of view, if crew must traverse daily among multiple vehicle systems that each have their own unique set of alerts, there is real danger of high cognitive workload and increased errors, as crewmembers must remember what each alert tone means. The greater the number of independent systems, the greater the problem. The concern is that if crewmembers must use cognitive resources for remembering multiple set of alerts, those resources may not be available for other, more important things, such as problem solving or emergency management.

Two potential countermeasures to this design issue have been identified. The first strategy is to add speech messages to critical alerts. Even if the tones used to communicate an alert vary by system, adding a speech component to higher level events (warning and emergencies) could communicate the essential information an astronaut would need to properly respond. However, it is critical to assess the relative advantages and disadvantages of including a speech component in the design of multimodal alerting systems.

Another solution would be to require a common alert set. Human Factors guidelines generally advocate limiting the number of things that the user must remember, favouring recognition over recall (Nielsen, 2010). With respect to auditory signals, the research-based design guidelines in Yeh, et al. (2016) state that the total number of auditory signals should be limited to four to six sounds, or even three or four when workload and time pressure are high (Cardosi & Murphy, 1995; McAnulty, 1995).

Given human factors concerns over the Artemis scenario, and the safety-critical aspects of alerting, two studies have been completed toward the goal of identifying multimodal alert designs that will mitigate potential issues related to lack of commonality. The first study evaluates augmenting NASA's current approach to multimodal alerting (text and tones) with speech alerts. This enhanced capability could be used to bridge the gap among different alert systems by providing an auditory message describing what the issue is and where it is located. The second study evaluates performance advantages of an alert set common to multiple operational contexts (i.e., one alert vocabulary for multiple vehicles, spacesuits, or habitats), versus a unique alarm vocabulary per context.

STUDY 1: SPEECH + TONE VS. TONE-ONLY ALERTS METHOD

Participants

Participants ($N = 25$) were recruited from the Human Test Subject Facility (HTSF) at the NASA Johnson Space Center (JSC) or were qualified personnel at JSC who agreed to participate. To maximize the generalizability of the results to astronauts, participants were screened and qualified based on several criteria: healthy, non-smoking, age 30 to 55 years, and a minimum of a Master of Science in a STEM discipline, or equivalent years of experience in a science/engineering field.

Experiment Design

This study used a 2x2 within-subjects design. One factor was Alert set: Tone-Only or Tone+Speech, and the other factor was Task Type: Electronic procedures or Mission Control Center (MCC)-read procedures. All participants completed all conditions. Presentation order of the two alert sets, and two task types were counterbalanced to control for potential carryover effects (e.g., learning).

Procedure and Materials

Participants came to the Human Factors Engineering Laboratory (HFEL) for two study sessions. One session consisted of all activities pertaining to the Tone+Speech alerts, and the other session pertained to the Tone-Only alert sessions. Half of the participants completed the Tone+Speech alerts session first, and then the Tone-Only alert session. The other half completed the study in the reverse order. Tone-only alerts consisted of non-verbal sounds (siren, klaxon, alternating tones, continuous tone etc). Tone+Speech alerts consisted of the tone alerts along with a synthetic female voice identifying the alert type, event, and location of the alert[1] (e.g., Siren tone + Emergency–Fire–HALO, Fire–HALO Emergency– Fire–HALO, Fire–HALO). This voice was selected based on the results of a prior internal preference test conducted at NASA.

After signing the study consent form, participants completed a hearing screening questionnaire (no participants were screened out). Participants then familiarized themselves with the alerts they would hear in that session. They would then complete 55 practice trials with feedback (Correct or Incorrect response). After the practice trials, the participants completed a 36-trial mastery test without feedback. They then completed the experimental task.

Participants were told to imagine they were an astronaut whose vehicle had experienced a power system issue, resulting in problems with the alerting system. Some alerts coming in were false alarms. Their job was to use procedures to reconfigure a backup electrical power system, and to report any alerts that came in to MCC for confirmation. Procedures consisted of several different types of tasks (i.e., navigating the interface, checking telemetry, and using a variety of types of controls). Because procedures were detailed (e.g., press this button, check this value, close this switch), the participants did not get specific training on the displays. Procedures were provided in either an

[1]Locations were either the Habitation and Logistics Outpost (HALO) or the Human Landing System (HLS).

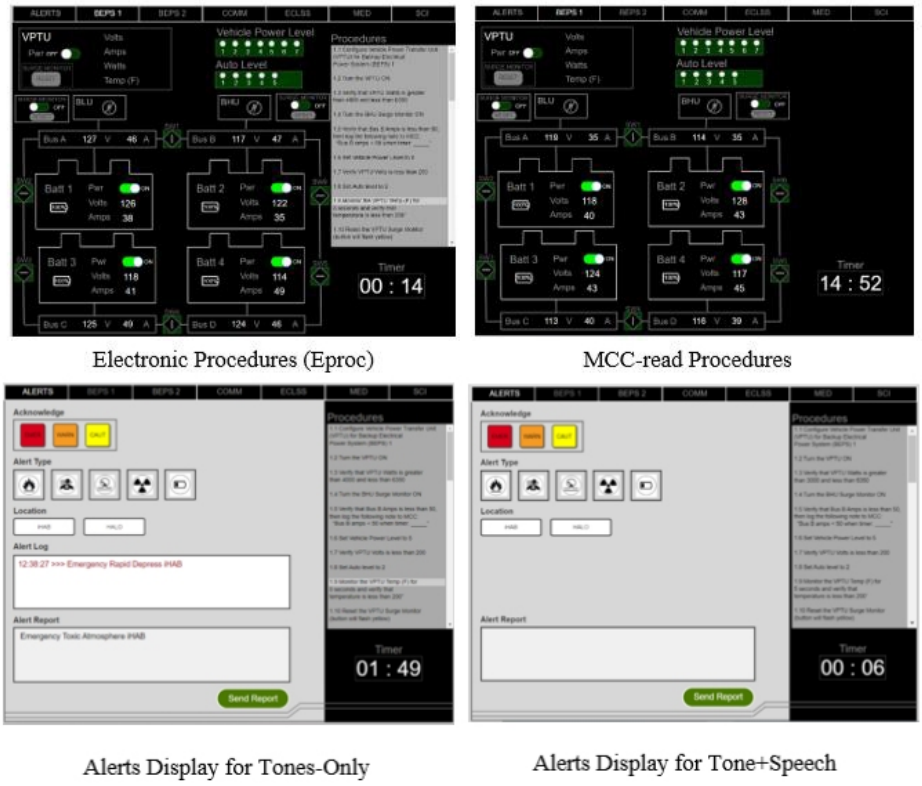

Figure 1: Screenshots of alert task displays.

electronic format embedded in the display (Eproc) or were read to the participant by MCC (the test conductor). When the participant had procedures read to them by MCC (MCC- read), no written procedures were visible to them. They had to listen carefully and could converse to clarify steps. They were instructed to tell MCC to stand by/pause reading when they heard an alert. It is assumed that this would be a realistic response in the actual operational scenario. Participants completed both versions of the task as blocks within the same session. Half of the participants completed the Eproc version of the task first and the MCC-read version second; the other half completed it in the reverse order. The experimental displays are depicted in Figure 1.

The second session was conducted the same as the first, except that the alternate condition alert set was heard (i.e., Tone-only alert or Tone+Speech alert). After completing the second session, participants were asked to indicate which alert set they believed would work best for space vehicle/habitats, and why. Participants were given the chance to make any final suggestions or comments, and then were thanked and dismissed.

RESULTS

Only correct trials were used in the analysis of response times. Alert Type, Location, and Send Report response times were combined and averaged to

represent "time to identify the alert." Because there were no voice messages for Caution alerts, analyses were split into non-Caution and Caution to keep the comparison between Tone-Only and Tone+Speech for alert messages for which they were truly equivalent in nature (Emergencies and Warnings).

Overall Accuracy and Workload

Overall, participants were more accurate (Accuracy $M = 95\%$) when responding to Tone+Speech alerts than when responding to Tone-Only alerts (Accuracy $M = 92\%$). However, a formal test conducted with a paired Brunner-Munzel test determined that the difference was not statistically significant, BMT = 0.90, $p = .38$. Subjective workload ratings, as measured by the Bedford Workload Scale (Roscoe, 1984), were extremely similar across all conditions (3.16 for Tone+ Speech and 3.14 for Tone-Only); thus, no further analysis was performed.

Time to Acknowledge Alert

Time to Acknowledge Alert is defined as the time from alert annunciation until the pressing of the correct alert acknowledge/severity button (Emergency, Warning, or Caution). Participants were faster to acknowledge Tone-Only alerts ($M = 5.40$ seconds, $SD = 1.06$) compared to the Tone+Speech alerts ($M = 7.96$ seconds, $SD = 0.89$). Participants time to acknowledge alerts were about the same for the Eproc version ($M = 6.44$ seconds, $SD = 1.70$) and the MCC-read version of the task ($M = 6.92$ seconds, $SD = 1.5$).

A formal test on time to acknowledge was performed using repeated measures ANOVA. There was a statistically significant main effect for alert set, $F(1, 24) =$ ($p < .001$) and task type, $F(1, 24) = 12.51$, $p = .002$. The observed difference in time to acknowledge by alerts set was large in effect (Cohen's $d = 2.11$). The observed difference between time to acknowledge by task was moderate in effect (Cohen's $d = 0.53$). Results indicated that the interaction between alert type and task type was also statistically significant, $F(1, 24) = 4.91$, $p = .036$. Participants were faster to press acknowledge in the Eproc ($M = 5.04$ seconds) version of the task than the MCC-read version ($M = 5.76$ seconds) of the task for Tone-Only alerts (a moderate effect, Cohen's $d = 0.66$), but were about the same for Tone+Speech alerts (a small effect, Cohen's $d = 0.40$). The fact that in both alert sets, the Eproc condition had faster acknowledge times may indicate that conversing with MCC slightly delayed responding to the alert, but these differences are so small from a practical standpoint, that there is little value in further discussion of the interaction.

Time to Categorize Alerts and Send Report (*Identification Task*)

Time to Categorize Alerts and Send Report is defined as the time from release of the acknowledge button until the press of the Send Report button (also referred to as the *Identification Task*). Participants were faster to identify the type of event and location and send that information when responding to the Tone+Speech alerts ($M = 2.45$ seconds, $SD = 0.88$), as compared to the Tone-Only alerts ($M = 3.77$ seconds, $SD = 0.88$). Participants reading

procedures ($M = 3.02$ seconds, $SD = 0.94$) performed about the same as when they were communicating with MCC ($M = 3.20$ seconds, $SD = 1.04$).

A formal test on time to complete the Identification Task was conducted using a repeated measures ANOVA. The main effect for alert type was statistically significant, $F(1, 24) = 81.64$, $p < .001$, and practically significant (Cohen's $d = 1.47$, a large effect size). The main effect of task type was not significant, $F(1, 24) = 1.71$, $p = .20$, nor was the interaction between alert type and task type, $F(1, 24) = 1.04$, $p = .32$. The difference in participants' performance on the alert identification task (i.e., event/location/send) is believed to be primarily due to slower reading of the alerts log (Tone-Only condition), versus getting the information more quickly and directly from the speech message (Tone+Speech condition).

Overall Preference

After completing the study, participants were asked about their preference for an alert set. After training and using both alert sets in a semi-realistic task, most participants preferred Tone+Speech alerts ($N = 20$) over the Tone-Only alerts ($N = 3$). A few participants ($N = 2$) stated no preference.

Caution Responses

In the experimental task, participants were trained that Caution alerts were low priority and were not accompanied by a speech message. During the study, the test conductors noticed that participants appeared to hesitate when responding to Caution alerts during the Tone+Speech session. It was hypothesized that when conditioned to listen for speech alerts, participants generalized waiting for speech to the Caution alerts, even though they had been trained that Caution alerts did not have a speech component. This effect could be explained by negative transfer of learning. This tendency to apply the same "rule" (waiting for the speech component), could result in unnecessary delays in responding to alerts.

To empirically determine if participants took longer to respond to Caution alerts in the Tone+Speech condition than the Tone-only condition, time to acknowledge Caution alerts was compared for each alert set. A paired t-test reveals that participants were significantly faster to respond to Caution alerts in the Tone-Only condition ($M = 5.40$ seconds, $SD = 1.18$) than in the Tone+Speech condition ($M =$ seconds, $SD = 1.41$), $t(24) = 2.60$, $p = .016$. The observed difference (0.78 seconds) was moderate in size as judged by Cohen's $d = 0.52$.

Discussion

This study demonstrated that speech alerts can offer more quickly digestible bits of information than a tone but require listening to the entirety of the alert message (versus reading it or identifying the tone). There was about a 2.5 second advantage of pressing the acknowledge button after hearing a Tone versus a Tone+Speech message. There was about a one second advantage of identifying alert details in the Tone+Speech condition. This results

in a net advantage of 1.5 seconds for Tone alerts. However, this study presented a best-case scenario for the Tone-only condition. If participants had to physically move to a computer to see alert details, the response time for the Tone-Only condition would have the penalty of additional seconds to minutes in travel time, and then display reading time (unless the details could be represented as part of a more complex tone alert). Participants receiving the Tone+Speech alert would be ready to respond as soon as they heard one iteration of the alert. Astronauts onboard a spacecraft may spend significant time away from a computer display performing maintenance, housekeeping, food prep, stowage, sleeping, and other activities. In these cases, the ability to understand an alert and plan a response while moving to a computer would have significant advantages – especially in the case of emergencies. In sum, the real benefit of a speech component comes when the number of unique tones overly taxes memory limits (not observed with the tone sets in this study), or when the listener is not situated in front of a display that can provide alert details. Finally, the Tone+Speech alerts were overwhelmingly preferred by participants (20 of 25).

The task type manipulation (i.e., electronic procedures versus MCC-read procedures) was included to determine whether dialoguing with MCC during the task would introduce any interference with hearing and processing the alert speech message. Results did not show much difference between performance in these two tasks. It should be clarified that the participant did not hear MCC conversation while responding to the alerts but heard conversation up until the alert sounded and/or they asked MCC to stand by. In the operational spacecraft environment, it is likely that much of the conversation is ceased when an alert occurs, but a more robust test of possible speech interference should be performed in a future study.

Another interesting finding in this study is that when speech and non-speech alerts are used together, response delays can result. Participants were trained that Caution tones would not be accompanied by a speech message due to their low priority but took longer to respond to Cautions in the Tone+Speech condition. Several participants commented that during the Tone+Speech session Cautions always made them pause because they were used to waiting for a speech message. These results indicate there may be a delay in responding to Tone-Only alerts if a speech message is expected. This finding was interpreted to suggest that if alert sets use a combination of speech-alerts and tone-only alerts, speech should be reserved for higher priority alerts.

STUDY 2: ALERT COMMONALITY AND CONTEXT

Participants

Participants ($N = 19$) were recruited from the set of participants who completed the previous experiment. This reduced training time, since the same alert set was used in this study (along with a new set), and the experimental task was also highly similar.

Experiment Design

This study used a mixed methods approach, with one between-subjects variable (Alert Set: *Common Alert Set* or *Multiple Alert sets*), and one within-subjects variable (Origin/Location of alert: *HLS* or *HALO*). Half of the participants ($N = 10$) trained on and used the Common Alert Set and half ($N = 9$) trained on and used the Multiple Alert sets.

Procedure and Materials

After signing the study consent form, participants completed a hearing screening questionnaire (no participants were screened out). They then completed alert familiarization, training, and the experimental task very similar to Experiment 1. Participants in the Common Set condition completed familiarization and training with the tones learned in Experiment 1. They were trained to use these alert tones in both the HALO and HLS locations. Participants in the Multiple Alert Sets condition completed familiarization and training with that same tone set (referred to as the "HALO" tone set). These participants also received training on an alternate tone set (referred to as the "HLS" tone set). All participants received 24 trials of training and 24 test trials on their assigned alert sets. Each group then received 48 practice trials with the experimental task, to practice applying the response rules they had learned (described below).

The electrical power system configuration task and alert identification task used in Study 1 were used in this study as well, with two exceptions: 1) procedures were always electronic and displayed within the electrical power display, participants were given different response instructions depending upon where they were performing the task (i.e., HALO or HLS). Instructions were as follows:

- If an alert occurs in a vehicle where you are located:
  - Acknowledge alert, log alert type, and send to MCC (as in Study 1)
- If an alert occurs in the other vehicle, or is a Caution:
  - Acknowledge the alert, and then send to MCC (no logging of details)

To simulate astronauts moving from one docked vehicle to another and responding to alerts, participants changed locations during the study. Two adjoining rooms within the lab were used in the study to represent two vehicles docked together. One room was referred to as the *HLS* vehicle, and the other was referred to as the *HALO* vehicle. Signs with these names were displayed on the walls in each room. The interfaces for the experimental task were virtually identical to those used in Experiment 1. The laptop computer interfaces were the same in each room, except for color scheme. The HALO interface had a black and green theme, and the HLS interface had a blue theme. The goal was to loosely represent the actual operational environment and to determine if context (knowing where you are performing the task) would help with identification of the location of the alert.

They first completed the experimental task in the location (i.e., HALO or HLS) in which they familiarized, trained, and tested on their assigned alert set. This starting location was counterbalanced, with half of each condition

starting in HALO and half in each condition starting in HLS. At the end of the procedure, they completed rating scales, and then physically moved to the other location (i.e., HALO or HLS adjoining room) and completed the procedural task again. They switched back to the original location for a third procedure and switched once more for the fourth and final iteration of the task.

Analyses

Due to the small sample size, data were analyzed with the Brunner-Munzel test (BMT). The Brunner-Munzel test is appropriate for small samples and does not have any assumptions for the distribution of the data (e.g., does not need to have a normal distribution). The statistical hypothesis being tested is whether the outcome of interest tends to be of equal size (stochastic equality) under two conditions or between two groups (Brunner & Munzel, 2000). Effect sizes are reported using Vargha and Delaney's A (VDA; 2000), along with an accompanying 95% confidence interval, and interpreted using their suggested guidelines. Values for VDA ranging from .56 to .64 are considered small effects, values ranging from.64 to.71 are considered medium effects, and values of .71 or greater are considered large effects.

RESULTS

Overall Accuracy and Workload

Participants were slightly more accurate when responding to the Common Alert Set (Accuracy $M = .93$, $SD = 0.04$) than when responding to the Mixed Alert Set (Accuracy $M = .90$ $SD = 0.09$). A Brunner-Munzel test indicated that the difference is not statistically significant, BMT(10.52) = 0.48, $p = .64$. There were no meaningful differences observed between the overall workload participants experienced in the Common Alert Set condition (M = 3.58, SD = 1.60), and the workload experienced by participants in the Mixed Alert Set condition (M = 3.58, SD = 1.08).

Time to Acknowledge Alert

Time to acknowledge alerts for participants in the Common Alert Set condition ($M = 5.17$ seconds, $SD = 1.20$) was about 0.20 seconds faster than those in the Mixed Alert Set condition ($M = 5.37$ seconds, $SD = 0.99$). There was no evidence to suggest that participants were stochastically different in their average time to press the acknowledge button, BMT(16.51) = 0.40, $p = .70$, VDA = .56 [.26,.85].

Time to Categorize Alerts (Alert Originated From Participant's Location)

Time to categorize alerts was calculated as mean time to press the Event Type button when responding to alerts originating from a participant's work location. Participants in the Common Alert Set condition were about 0.62 seconds slower on average (M = 2.48 seconds, SD = 0.88) than those in the Mixed Alert Set condition (M = 1.86 seconds, SD = 0.27). The participants

were not stochastically equal in their average time to press the event button, BMT (9.39) = 2.90, p = .017, VDA = .83 [.57, 1.00]. The evidence suggests that participants in the Mixed Alert Set condition generally were faster to select the event type and had an advantage for determining whether the alert originated from their location, compared to those in the Common Alert Set condition.

Time to Send Report (Alert Originated From Other Location)

Time to press the Send Report button was calculated as mean time to press the Send Report button when responding to alerts not originating from participant's work location. Participants in the Common Alert Set group were about 0.88 seconds slower on average ($M = 2.45$ seconds, $SD = 1.01$) than those responding to the Mixed Alert Set ($M = 1.57$ seconds, $SD = 0.54$). The participants were not stochastically equal in their average time to press the Send Report button when the alert originated from the other location, BMT (14.11) = 2.79, $p = .014$, VDA = .80 [.57, 1.00]. The evidence was interpreted to suggest that participants in the Mixed Alert Set condition were generally faster to press the Send Report button and had an advantage for determining whether the alert originated from the other location compared to those in the Common Alert Set condition.

DISCUSSION

Both the Common and Multiple tone set conditions provided for accurate performance and similar workload. The a priori hypothesis was that increasing the number of auditory tone sets to be learned and used would negatively impact performance by overloading cognitive resources. With the tone sets used in this study, that hypothesis was not supported. Participants in the Mixed Alert set group were faster than participants in the Common alert set group in identifying alert type and recognizing origin of the alert. While there were some participant comments about workload required to manage multiple tone sets, indications are that the study did not significantly overload the participants, and they were able to manage working with two separate tone sets.

The main takeaway from this study is that two alert sets are manageable without reduced performance, if the two alert sets are very distinct. In practice, the use of multiple tone sets will require more training time upfront and additional maintenance or refresher training. As the number of unique spacecraft and space- related assets (e.g., spacesuits, rovers, habitats) continues to grow, it will be critically important to better understand the factors that can impact the success of alerting systems, so that appropriate risk assessment and design decisions can be made. The threshold for number of unique alert sets that can be successfully used is still unknown, as well as how much mitigation distinctiveness can offer. These questions must be answered with a more robust study that varies number of alert sets and level of distinctiveness.

REFERENCES

Brunner, E., and Munzel, U. (2000). The nonparametric Behrens-Fisher Problem: Asymptotic theory and a small-sample approximation. *Biometric Journal*, *42*(1), 17–25.

Cardosi, K. M. and Murphy, E. D. (1995). *Human Factors in the Design and Evaluation of Air Traffic Control Systems* (DOT-VNTSC-FAA-95-3; DOT/FAA/RD-95-3). Washington, DC: US DOT Office of Aviation Research. Available from: http://ntl.bts.gov/lib/33000/33600/33633/33633.pdf.

McAnulty, D. M. (1995). Guidelines for the Design of GPS and LORAN Receiver Controls and Displays (DOT-VNTSC-FAA-95-7; DOT/FAA/RD-95-1). Cambridge, MA: US DOT Volpe National Transportation Systems Center.

Nielsen, J. (2010, November 15). 10 Usability Heuristics for User Interface Design. NN/g Nielsen Normal Group. https://www.nngroup.com/articles/ten- usability-heuristics/

Roscoe, A. H. Assessing pilot workload in flight. Flight test techniques. (1984). In Proceedings of NATO Advisory Group for Aerospace Research and Development (AGARD) (AGARD-CP373). Neuilly-sur-Seine, France: AGARD.

Vargha, A., and Delaney, H. D. (2000). A critique and improvement of the "CL" common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25*(2), 101–132.

Yeh, M., Swider, C., Jo, Y., and Donovan (2016). *Human Factors Considerations in the Design and Evaluation of Flight Deck Displays and Controls*, Version 2.0. Federal Aviation Administration: Washington DC.