Collegiate Aviation Review International

Volume 40 | Issue 2

Peer Reviewed Article #8

11-14-2022

Creation of Tonal and Speech Alarm Efficacy Scales for Aviation

Ryan A. Lange
Embry-Riddle Aeronautical University

Stephen Rice
Embry-Riddle Aeronautical University

Cameron M. E. Severin
Embry-Riddle Aeronautical University

Jonah S. L. Chiu
Embry-Riddle Aeronautical University

Keith J. Ruskin
Embry-Riddle Aeronautical University

Connor Rice
Embry-Riddle Aeronautical University

Alarms have been in use for decades in aviation; however, it is still the case that many alarms are sub-optimally designed and do not perform well. Some alarms are so poorly designed that they increase workload, confuse the user, and/or cause a severe loss of trust. When users are asked about alarm efficacy, they often say that the alarm is either good or bad. While this provides some useful subjective information, we would argue that a quantitative scale offers more value. Using a consensus research method to ensure construct validity, we solicited 2362 participants across a four-phased, one-year study in the development of a Tonal Alarm Efficacy Scale and a Speech Alarm Efficacy Scale. A factor analysis using principal components and varimax rotation provided strong evidence of validity, while Cronbach's Alpha and Guttman's Split Half tests were used to ensure high consistency and reliability, respectively. Follow-up analyses highlight the sensitivity of the scales. These types of quantitative scales can provide a means for users, designers, engineers, and human factors experts to communicate in a common language to design more effective alarms for our society. The present study attempts to fill a gap in the current literature by providing Tonal and Speech Alarm Efficacy Scales for use applications in aviation.

Recommended Citation:
Lange, R. A., Rice, S., Severin, C. M. E., Chiu, J. S. L., Ruskin, K. J. & Rice, C. (2022). Creation of tonal and speech alarm efficacy scales. *Collegiate Aviation Review International,* 40(2), 132-145. Retrieved from http://ojs.library.okstate.edu/osu/index.php/CARI/article/view/8709/8435

Introduction

Alarms are a critical component of modern automation. They allow users of complex systems to focus on primary tasks rather than monitoring a multitude of dynamic information sources (Parasuraman et al., 2000). Many devices, including those operated by general users with no formal training, use alarms to indicate system status. For example, an automobile may sound a chime to indicate that the door is ajar or that the driver's seatbelt is not fastened, or the TCAS alarm in an aircraft may emit a vocalization indicating an imminent traffic conflict. Because of their ubiquity and the increasing complexity of modern technology, many people rely on these signals to maintain situational awareness. Poor design may impair an alarm's ability to attract the user's attention (International Organization for Standardization, 2003). Poor design may also contribute to false or nuisance alarms, which reduce the user's response to a signal (Breznitz, 1982; Dixon, Wickens & McCarley, 2007; Rice, 2009; Wickens & Dixon, 2007). Allowing users to participate in the design process may improve overall system performance (Nielsen & Levy, 1994). Therefore, the question is: Can we develop scales to evaluate the efficacy of tonal and speech-based alarms? This study aims to develop two coexistent, independent scales of efficacy that allow users to evaluate both tonal and speech alarms, creating more effective signals for aviation applications and a variety of other domains.

Literature Review

Under Parasuraman et al.'s (2000) model for levels of automation, tonal alarms can be classified as Stage 2 automation because alarm systems monitor multiple information sources within a complex system and alert the user to potential issues. Speech alarms might be classified as Stage 3 automation, as they provide verbal guidance or instructions.

85% of people in modern industrialized societies use an alarm clock to wake them (Roenneberg, 2012). The smoke detector is another commonly-used alarm that reduces the risk of fatalities by 55% (National Fire Protection Agency, 2021). These alarms are so common that we often fail to appreciate their design. Many industries—including aviation, healthcare, and nuclear power generation—employ safety-critical processes that require operators to maintain a precise mental model of system function (Carroll & Olson, 1988). These industries rely on many alarms to report system status information. The number of alarms to which users are exposed, combined with the high rate of false and nuisance alarms in some settings, suggests that a deeper understanding of their purpose and design is required (Ruskin & Hueske-Kraus, 2015).

Healthcare professionals, for example, must observe and comprehend a constant flow of data that reflects their patients' conditions. Medical equipment is designed with a comprehensive system of alarms, which notify the relevant personnel of changes in patient status hundreds of times per day (Lewandowska et al., 2020). The design of these alarms may have detrimental effects. Nurses are often subject to alarm fatigue and nuisance alarms (Ruskin & Hueske-Kraus, 2015), which can negatively impact patient safety. In a 2016 survey of clinical care providers, 30% of hospital medical staff indicated that their healthcare institution experienced adverse patient events or outcomes related to clinical alarms (Clark, 2016). In these clinical settings, an average of 150-400 alarms are generated per patient per shift, which comprises 35% of the working time of an ICU nurse (Lewandowska et al., 2020, Li et al., 2018). Recognizing an alarm, identifying its source, and interpreting its meaning require additional cognitive effort from users, especially when alarms have not been adequately designed (Ruskin & Hueske-Kraus, 2015).

In aviation, alarm-related incidents, particularly those involving the Minimum Safe Altitude Warning (MSAW), have been reported in air traffic control (Ruskin et al., 2021). These incidents are also associated with the Ground Proximity Warning System (GPWS) and the Traffic Collision Avoidance System (TCAS) (Bliss et al., 1999). Alarm fatigue may be attributable to the acoustic properties of the alarm (Edworthy, 2013). The perceived urgency of some alarms may also be difficult to judge based on their design, preventing the user from correctly responding to the alarm (Arrabito et al., 2004; Burt et al., 1995). Accurately measuring an alarm's efficacy early in the design process is a crucial factor in developing safety-critical systems.

Several measures have been used to evaluate the application of an alarm within a system. Jian et al. (2000) developed a scale that measures the level of trust a user places in an automated system. Singh et al. (1993) developed a scale that indicates the potential for complacency with an automated system by measuring attitudes toward commonly encountered automated devices. Arrabito et al. (2004) and Burt et al. (1995) have described the measurement of perceived alarm urgency using a Likert scale, which is a quick and effective way to determine perceived alarm urgency. This method also provides a way to evaluate and compare candidate alarms. Some studies have evaluated the perception of an alarm's acceptability with a single-item rating (Taylor & Wogalter, 2012). Although this may be an indicator of efficacy, an alarm that users perceive to be acceptable may not be effective for a given application. Further, many measures of existing alarm systems are focused on the metrics of alarm system performance and not the measurement of alarm system efficacy (Dorgo et al., 2021).

The Current Study

While these scales are useful for their intended purposes, they do not rate the overall efficacy of an alarm or breakdown where the problems may occur during design. A validated Likert-type scale remains to be developed that allows users to determine an alarm's perceived efficacy based on its inherent qualities. The value of single-factor rating systems is that they are quick and easy to administer. Such single-factor instruments can be delivered to operators who rely on a given alarm in the form of a survey and make it possible to easily evaluate the perceived efficacy of the alarm. The survey can be administered by various personnel, including engineers, system designers, alarm researchers, and even people without prior experience in alarm design or management. The results of these Likert-type scales are easily interpreted by the person responsible for administering and evaluating the survey, allowing the administrator to make quantitative decisions about alarms.

Prior research has shown that poor alarm design can negatively impact operator performance across many industries, including aviation. While scales have been created to measure other qualities affecting alarm performance, such as system trust and perceived urgency, none have been developed that measure perceived alarm efficacy. We have filled this research gap by developing two scales that may capture the perceived efficacy of both new and existing tonal and/or speech alarms from the users' perspective. In order to develop these scales with construct validity, we used a consensus research method (Hinkin, 1998) across a four-phased study. Participants were recruited to identify potential items for inclusion. They then narrowed those items down to a final scale. Validity, consistency, and reliability were tested using Factor Analysis, Cronbach's Alpha, and Guttman's Split Half, respectively (see Table 1 and Table 2).

Methodology – Tonal Alarm Efficacy Scale

The study was conducted with ethics approval from the university review board prior to participant recruitment.

Scale Building

The Tonal Alarm Efficacy Scale was developed in four phases: 1) item generation, 2) nominal pairing of the items, 3) Likert scale pairing, and 4) factor analysis and sensitivity test. The methods employed in this study were guided by Hinkin's (1998) framework for scale development. Similar approaches have been used to create scales in prior studies (Rice et al., 2014; Rice et al., 2015).

Phase 1: Item Generation

The overall goal of Phase 1 was to identify potential items (the written options that respondents can select from when they provide answers to questions in a survey) for integration into the final scale. This goal was accomplished by recruiting participants to complete an internet-based survey. In addition, other items were added to the list through a literature review and eight subject matter experts (SMEs) who represented human factors professionals, spaceflight engineers, airline pilots, and an anesthesiologist. Collecting inputs from SMEs has been shown to be an effective way of contributing content validity to research (Burns & Grove, 1993, p. 343).

Participants. Two hundred and two participants (94 Female, 105 Male, and 3 Other) were recruited via a convenience sample using Amazon's ® Mechanical Turk ® (MTurk) program—a crowdsourcing marketplace that allows people to participate in studies for monetary compensation. This data collection method has proven to be reliable and better represents the general population compared to laboratory data (Buhrmester et al., 2011; Thomas & Clifford, 2017). All participants were at least eighteen years of age, with a mean age of 39.92 (*SD* = 13.14) years. The participants were not screened for any specific background or technical proficiency for this phase or any other phase. Participants were paid USD $0.25 upon completion of this phase. A sample size of 200 was deemed necessary to complete the task and generate as many potential items as possible.

Procedure, Materials, and Stimuli. Participants gave their consent electronically via Google Forms ®. Following this, participants were presented with a simplified definition of what constitutes an alarm as follows:

> *An alarm is any type of auditory or visual cue that lets people know there is an ongoing danger that requires immediate action. Some alarms convey information using only tones (Tonal Alarms). Some alarms may include speech components to convey additional information (Speech Alarms).*

Participants were then presented with the following instructions:

> *In the context of the design of an alarm system, please enter 5 words or phrases that you feel are strongly relevant to the concept of a successful TONAL ALARM system design. In other words, what words or phrases would describe the qualities of a TONAL ALARM system that uses tones to convey information? For example, the alarm should be "loud," "urgent," etc. Each answer should include a word or a one-sentence phrase.*

After answering these questions, participants responded to demographic questions. Lastly, participants were compensated and thanked for their assistance.

Results. After completing the survey and consulting with the SMEs and research team, 392 unique items were generated for tonal alarms and 531 items for speech alarms. The research team reviewed the data to ensure that duplicates were either combined or removed from the list. Longer phrases were shortened to their most basic form when possible (e.g., "easy to understand" became "understandable"). In addition, researchers verified that all phrases were grammatically correct. The post-truncation item list contained 126 items for tonal alarms and 147 items for speech alarms.

Phase 2: Nominal Pairing

The goal of Phase 2 was to further refine the item list. In this phase, participants were asked to judge each term based on its relevance to the topic of "tonal alarms."

Participants. Two hundred and six participants (84 Female, 119 Male, and 3 Other) were recruited from MTurk. The average age was 41.74 (*SD* = 13.56) years. Participants were paid USD $0.50 upon completion of this phase. A sample size of 200 was deemed necessary to complete the task and maintain consensus.

Procedure, Materials, and Stimuli. The items collected in Phase 1 were presented to each participant, who rated each item on its relevance to "tonal alarms." Participants were able to rank each term as 'relevant,' 'not relevant,' or 'I don't know.'

Results. A minimum relevancy score of 70% was used to determine if an item would be included in Phase 3. Thirty-seven items met or exceeded the criteria.

Phase 3: Likert Scale Pairing

The goal of Phase 3 was to more accurately determine which items were relevant to a tonal alarm efficacy scale. In this phase, participants read through the 37 items from Phase 2 and rated them on the following scale: 0 (Not at all related to tonal alarm qualities), +1 (Slightly related to tonal alarm qualities), +2(Somewhat related to tonal alarm qualities), +3(Quite related to tonal alarm qualities), +4 (Extremely related to tonal alarm qualities).

Participants. Two hundred and thirty-nine participants (79 Female, 159 Male, and 1 Other) were recruited through a convenience sample using MTurk. The average age was 35.59 (*SD* = 10.00) years. Participants were paid USD $0.30 upon completion of this phase. A sample size of 200 was deemed necessary to complete the task and maintain consensus.

Results. Six terms met the inclusion criteria and were included in the final Tonal Alarm Efficacy scale: *effective*, *attention-grabbing*, *audible*, *loud*, *useful*, and *identifiable*.

Phase 4: Factor Analysis and Sensitivity Test

In this phase, the final scale of six items was assessed for validity, reliability, and sensitivity. Participants were given a hypothetical scenario in which an alarm would be necessary and were instructed to listen to one of three alarm stimuli. The stimuli were created with Audacity® audio editing software version 3.1.3 and uploaded onto YouTube®. Following this, participants responded to the scenario using the Tonal Alarm Efficacy scale (see Appendix A).

Participants. Six hundred and nine participants (347 Males and 262 Females) were recruited using a convenience sample from MTurk. The average age was 38.90 (*SD* = 11.29) years. Participants were paid USD $0.20 upon completion of this phase. A sample size of at least 600 was deemed necessary for conducting the factor analysis. As a general rule, the sample size required for factor analysis is roughly ten times the number of variables you are testing (Comrey & Lee, 1992).

Procedure, Materials, and Stimuli. Participants were randomly divided into three groups based on alarm quality and presented with the following scenario:

> *Imagine a scenario in which you are in a building. A serious event has occurred and the building's evacuation alarm sounds. The following alarm is sounded and you must respond accordingly. Play the video to hear the alarm, and then answer the following question*.

Each group listened to one of three alarm stimuli that were designed to be either “low-quality” (n = 243), “mid-quality” (n = 229), or “high-quality” (n = 157). If needed, participants could listen to the alarm as many times as they wished. Participants then responded to the Tonal Alarm Efficacy scale (see Appendix A).

Results. A factor analysis using the principal components and varimax rotation showed that all items strongly loaded on a single factor for each group, with 59-62% of the variance explained for each model (see Table 1). The scale also had a very high level of internal consistency, as determined by Cronbach's alpha values of 86-88%. Guttman's Split Half tests indicated very high reliability with results between 88-90%.

Table 1
Statistical Analysis Results

Analyses performed	Low-Quality Alarm	Mid-Quality Alarm	High-Quality Alarm
Variance explained	0.62	0.59	0.62
Cronbach's alpha	0.88	0.86	0.88
Guttman Split Half Test	0.90	0.88	0.89

Before analysis, scores of the scale were averaged to create a single "alarm efficacy" score for each participant. The scores for the three groups were then compared using a one-way ANOVA with an LSD post hoc test. Alarm efficacy was statistically significantly different between the alarm quality groups $F(2, 606) = 30.63$, $p < .001$, $\eta^2 p = 0.092$. Alarm efficacy increased from the Low-Quality alarm group ($M = 0.49$, $SD = 0.89$) to the Mid-Quality alarm group ($M = 1.03$, $SD = 0.71$) and to the High-Quality alarm group ($M = 0.98$, $SD = 0.74$), indicating that the scale is sensitive to different levels of efficacy.

Methodology – Speech Alarm Efficacy Scale

Participants. One thousand one hundred and six participants (501 Female, 596 Male, and 9 Other) were recruited from MTurk. The average age was 38.32 (*SD = 11.43*) years. Participants were compensated in the same manner as Phases 1, 2, 3, and 4 of the Tonal Alarm Efficacy Scale development process (USD $0.25, $0.50, $0.30, and $0.20, respectively). Sample size determinations were also identical to the development of the Tonal Alarm Efficacy Scale.

Procedure, Materials, and Stimuli. The Speech Alarm Scale was developed exactly like the previous Tonal Alarm Scale. In Phase 1, 531 unique items were generated. These were pared down to 32 items in Phase 2 and seven items in Phase 3: *effective, attention-grabbing, simple, audible, clear, reliable, and understandable* (see Appendix B).

Results. A factor analysis using the principal components and varimax rotation revealed that all items strongly loaded on a single factor for each group, with 59-62% of the variance explained for each model (see Table 2). The scale also had a very high level of internal consistency as determined by Cronbach's alpha values of 86-88%. Guttman's Split Half tests indicated very high reliability, with results between 90-91%.

Table 2
Statistical Analysis Results

Analyses performed	Low-Quality Alarm	Mid-Quality Alarm	High-Quality Alarm
Variance explained	0.62	0.59	0.58
Cronbach's alpha	0.88	0.86	0.86
Guttman Split Half Test	0.91	0.90	0.90

Before analysis, scores of the scale were averaged to create a single "alarm efficacy" score for each participant. The scores for the three groups were then compared using a one-way ANOVA with an LSD post hoc test. Alarm Efficacy was statistically significantly different between the alarm quality groups $F(2, 665) = 14.70$, $p < .001$, $\eta^2 p = 0.042$. Alarm efficacy increased from the Low-Quality alarm group ($M = 0.36$, $SD = 0.94$), to the Mid-Quality alarm group ($M = 0.47$, $SD = 0.85$), and to the to the High-Quality alarm group ($M = 0.79$, $SD = 0.78$), indicating that the scale is sensitive to different levels of efficacy.

Discussion and Conclusions

The purpose of the current study was to create and validate two scales that can be used to study users' perceived efficacy of alarms. To achieve this, we conducted a four-phased study to ensure construct validity. Participants, including eight SMEs, were recruited to identify potential items for inclusion and then to narrow those items down into a final scale. Validity, consistency, and reliability were tested using Factor Analysis, Cronbach's Alpha, and Guttman's Split Half, respectively.

The use of a consensus methodology for scale development contributed to the construct validity of the two scales (Hinkin, 1998). Factor analysis further supported the validity of the scales, showing that all the items for both the Tonal Alarm Efficacy Scale and the Speech Alarm Efficacy Scale contribute to a single factor: "tonal alarm efficacy" and "speech alarm efficacy", respectively. The results of the Cronbach's Alpha calculation indicated a high level of internal consistency among the items. The reliability of the scales was tested using Guttman's Split Half tests. Unlike the test-retest method, this method only requires one administration of the scales, reducing the need for participants and allowing for easier administration over the internet.

Other scales developed prior to this study focused primarily on evaluating one or more factors that may contribute to an alarm's efficacy, such as user trust or complacency (Jian et al., 2000; Singh et al., 1993). Many studies have focused on alarm system performance metrics (Dorgo et al., 2021), but none have sought to fully capture a user's perceived efficacy of an alarm. The alarm efficacy scales will facilitate the integration of user input in the development of alarm systems. Ultimately, this will allow designs to be driven by human factors principles (Nielsen & Levy, 1994) while providing quantitative data to support design decisions.

The alarm efficacy scales are highly suitable for use in the aviation industry. They are easy to administer and implement into the design process and capable of producing actionable results. Due to their generalizable nature, the scales are also suitable for use in a variety of other industries. For example, an administrator working on the design of the flight deck for a new aircraft might begin by collecting or creating the alarms to be evaluated. Next, the administrator recruits a sample of participants who represent the alarm system's intended user: pilots. The pilots are invited to participate in the evaluation individually. Each pilot is first given information and context about the system that contains the candidate alarms. The pilot must have as much context as possible about the situation in which the alarm is utilized so that they may form an adequate opinion about the alarm's properties. The pilot is also briefed about the purpose of the scale and how to fill it out.

Next, the pilot listens to one of the candidate alarms and evaluates it by filling out the Tonal or Speech scale. The pilot is then presented with the next candidate alarm, and the process is repeated until the pilot has listened to and rated each of the candidate alarms. The administrator can then compare the scores and choose the best alarm.

Additionally, a designer could compare a current alarm with an improved version. For example, the designer might use the Speech Alarm Efficacy scale to have users rate a newly-designed speech alarm being considered for use in an updated version of the TCAS alarm in an aircraft. Based on the ratings of this alarm, the designer can implement specific changes to its design, creating an iterative process where user feedback is solicited whenever new changes are made to the candidate alarm. Ultimately, a highly refined version of the candidate alarm would be selected for use in the finished product.

This four-phased study was conducted to develop and test a scale for measuring the perceived efficacy of tonal and speech alarms. These scales are intended for use by equipment designers, manufacturers, and users. Both scales are short and easy to administer, making them ideal for use in the iterative design processes used to create and evaluate alarms over the phases of a design project. The Tonal and Speech Alarm Efficacy Scales will give users a voice in the design of alarms, ultimately improving the safety of those who rely on them.

Limitations

The current study has several limitations. First, participants were recruited using convenience sampling techniques. Since participants reported that they live in the United States, these results may be limited in perspective to western ideologies, leaving room for future studies to enhance the generalizability of the scales to international audiences. Additionally, control of the environment in which participants responded to auditory stimuli was not regulated. The volume of the stimuli, repetitions, the presence or absence of background noise and headphones, and the possibility that participants were distracted could not be controlled using the current study design. Finally, participants were not screened for hearing deficiencies prior to listening to the alarm samples, which may have impaired their ability to judge the efficacy of the alarms.

References

Arrabito, G. R., Mondor, T. A., & Kent, K. J. (2004). Judging the urgency of non-verbal auditory alarms: a case study. *Journal of Ergonomics*, *47(*8*)*, 821-840. https://doi-org.ezproxy.libproxy.db.erau.edu/10.1080/0014013042000193282

Bliss, J. P., Freeland, M. J., & Millard, J. C. (1999). Alarm related incidents in aviation: A survey of the aviation safety reporting system database. *Proceedings of the Human Factors and Ergonomics Society. Annual Meeting*, *1*, 6.

Breznitz, S. (2013). *Cry wolf: The psychology of false alarms*. Psychology Press.

Buhrmester, Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Burns, N., & Grove, S. K. (1993). The practice of nursing research: Conduct, critique & utilization (2nd ed). Sanders.

Burt, J. L., Bartolome, D. S., Burdette, D. W., & Comstock, J. R. (1995). A psychophysiological evaluation of the perceived urgency of auditory warning signals. *Ergonomics*, *38*(11), 2327-2340. doi:10.1080/00140139508925271

Carroll, J. M., & Olson, J. R. (1988). Mental models in human-computer interaction. *Handbook of human-computer interaction*, 45-65.

Clark, T. (2016, June 5). *HTF Update: 2016 National Clinical Alarm Survey Results*. AAMI Conference and Expo, Tampa, FL. http://thehtf.org/documents/2016%20National%20Clinical%20Alarms%20Survey%20Results.pdf

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed). L. Erlbaum Associates.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors*, *49*(4), 564-572.

Dorgo, G., Tandari, F., Szabó, T., Palazoglu, A., & Abonyi, J. (2021). Quality vs. quantity of alarm messages - How to measure the performance of an alarm system. *Chemical Engineering Research and Design*, *173*, 63-80.

Edworthy, J. (2013). Medical audible alarms: A review. *Journal of the American Medical Informatics Association: JAMIA*, *20*(3), 584-589. doi:10.1136/amiajnl-2012-001061

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*(1), 104–121. http://dx.doi.org/10.1177/109442819800100106

International Organization for Standardization (2003). Ergonomics — Danger signals for public and work areas — Auditory danger signals (ISO Standard No. 7731:2003). https://www.iso.org/standard/33590.html

Jian, J. Y., Bisantz, A., & Drury, C. (2000). Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, *4*, 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Lewandowska, K., Weisbrot, M., Cieloszyk, A., Mędrzycka-Dąbrowska, W., Krupa, S., & Ozga, D. (2020). Impact of Alarm Fatigue on the Work of Nurses in an Intensive Care Environment-A Systematic Review. *International journal of environmental research and public health*, *17*(22), 8409. https://doi.org/10.3390/ijerph17228409

Li, T., Matsushima, M., Timpson, W., Young, S., Miedema, D., Gupta, M., & Heldt, T. (2018). Epidemiology of patient monitoring alarms in the neonatal intensive care unit. Journal of Perinatology, 38(8), 1030-1038.

National Fire Protection Agency (2021). Smoke Alarms in US Home Fires. Retrieved from https://www.nfpa.org/News-and-Research/Data-research-and-tools/Detection-and-Signaling/Smoke-Alarms-in-US-Home-Fires

Nielsen, J., and Levy, J. (1994). Measuring usability — preference vs. performance. Communications of the ACM 37, 4 (April), 66–75.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, *30*(3), 286-297. https://www.ida.liu.se/~769A09/Literature/Automation/Parasuraman,%20Sheridan,%20Wickens_2000.pdf

Rice, S. (2009). Examining single and multiple-process theories of trust in automation. *Journal of General Psychology, 136*(3), 303-319.

Rice, S. C., Mehta, R., Winter, S., & Oyman, K. (2015). A trustworthiness of commercial airline pilots (T-CAP) scale for American consumers. *Journal of Aviation Technology and Engineering*, *4*(2), 55.

Rice, S., Mehta, R., Steelman, L. A., & Winter, S. R. (2014). A trustworthiness of commercial airline pilots (T-CAP) scale for Indian consumers. *International Journal of Aviation, Aeronautics, and Aerospace*, *1*(3), 3.

Roenneberg, Till. *Internal Time : Chronotypes, Social Jet Lag, and Why You're So Tired*, Harvard University Press, 2012. *ProQuest Ebook Central*, http://ebookcentral.proquest.com/lib/erau/detail.action?docID=3301120.

Ruskin, K. J., & Hueske-Kraus, D. (2015). Alarm fatigue: impacts on patient safety. *Current Opinion in Anesthesiology*, *28*(6), 685-690.

Singh, I. L. Molloy, R., & Parasuraman, R. (1993). Automation-induced "complacency": Development of the Complacency-Potential Rating Scale. *The International Journal of Aviation Psychology*, *3*(2), 111-122

Taylor, J. R. I., & Wogalter, M. S. (2012). Acceptability of Evacuation Instruction Fire Warnings. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 1753–1757. https://doi.org/10.1177/1071181312561352

Thomas, K. A., & Clifford, S. (2017). Validity and mechanical turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*, 184-197. https://doi.org/10.1016/j.chb.2017.08.038

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201-212.

Appendix A –Tonal Alarm Efficacy Scale

Please respond how strongly you agree or disagree with the following statements.

	Strongly Disagree	**Disagree**	**Neither Agree nor Disagree**	**Agree**	**Strongly Agree**
1. The alarm is Effective.	-2	-1	0	1	2
2. The alarm is Attention-grabbing.	-2	-1	0	1	2
3. The alarm is Audible.	-2	-1	0	1	2
4. The alarm is Loud.	-2	-1	0	1	2
5. The alarm is Useful.	-2	-1	0	1	2
6. The alarm is Identifiable.	-2	-1	0	1	2

The final alarm efficacy score will be the average of the six responses.

Appendix B - Speech Alarm Efficacy Scale

Please respond how strongly you agree or disagree with the following statements.

	Strongly Disagree	**Disagree**	**Neither Agree nor Disagree**	**Agree**	**Strongly Agree**
1. The alarm is Effective.	-2	-1	0	1	2
2. The alarm is Attention-grabbing.	-2	-1	0	1	2
3. The alarm is Simple.	-2	-1	0	1	2
4. The alarm is Audible.	-2	-1	0	1	2
5. The alarm is Clear.	-2	-1	0	1	2
6. The alarm is Reliable.	-2	-1	0	1	2
7. The alarm is Understandable	-2	-1	0	1	2

The final alarm efficacy score will be the average of the seven responses.