Integrating Rigorous Qualitative Methods into the Design and Evaluation of Safety-Critical Systems

Romane Dubus[1,2][a], Anke M. Brock[2][b] and Wendy E. Mackay[1][c]
[1]Université Paris-Saclay, CNRS, Inria, Orsay, France
[2]Fédération ENAC ISAE-SUPAERO ONERA, Université de Toulouse, Toulouse, France

Keywords: Mixed-Methods, Comparative Structured Observation, Safety-Critical Systems, Qualitative Methods, Quantitative Methods.

Abstract: Traditional aviation research includes quantitative or qualitative studies of pilots' behavior and cognition, followed by quantitative evaluations of current and novel system designs. However, qualitative and quantitative methods are rarely combined, making it difficult to link pilot behavior to specific design implications. This paper discusses how aviation researchers can benefit from a mixed-method approach that explicitly includes rigorous qualitative methods into the process of designing and evaluating safety-critical systems. We describe our use of the *comparative structured observation* method, where pilots perform at least two directly comparable realistic tasks using selected design variants, and are then asked to reflect deeply on the advantages and disadvantages associated with each. The goal is to obtain a more nuanced understanding of specific design trade-offs from the pilot's perspective.

1 INTRODUCTION

This paper focuses on the design and evaluation of autopilot systems, which were introduced to reduce cognitive overload and relieve pilots of monotonous tasks. However, these systems still require pilots to maintain situation awareness as they monitor the state of the aircraft, since pilots are responsible for detecting deviations and making effective decisions under rapidly changing circumstances. The lack of situation awareness has contributed to multiple incidents and accidents (Kharoufah et al., 2018) making it an essential design consideration.

Unfortunately, the traditional approach for designing such systems separates studies of pilots' reflections from decisions about cockpit design. Thus researchers who study pilots rarely contribute directly to the design of novel interactive systems, and aviation system designers rarely include pilots' reflections into their design process. Our review of the literature highlights the lack of studies that take advantage of pilots' reflections either to directly inform design or to interpret the evaluation of those designs. We argue that a mixed-method research approach that combines rigorous qualitative and quantitative methods will provide greater insights into both the design and the evaluation of safety-critical systems. We provide an example of how a rigorous mixed-method—*comparative structured observation*—can be successfully applied to autopilot design.

[a] https://orcid.org/0009-0009-1113-9650
[b] https://orcid.org/0000-0002-0017-396X
[c] https://orcid.org/0000-0001-8261-2382

2 RELATED WORK

2.1 Qualitative Approaches for Assessing Pilot Behavior

Qualitative research is essential for understanding pilot behavior, especially with respect to the loss of situation awareness and its effect on successful flight operation. Typical research methods (Curry, 1985; Wiener, 1985; Wiener, 1989) involve gathering and interpreting observational data, e.g. pilots who solve problems in flight similators (Sarter and Woods, 1992; Sarter and Woods, 1997), incident reports (Bureau of Air Safety Investigation, 1998) and pilot interviews, e.g. using *critical incident technique* (Flanagan, 1954), semi-structured or open-ended interviews, and questionnaires, e.g. with

Dubus, R., Brock, A. and Mackay, W.
Integrating Rigorous Qualitative Methods into the Design and Evaluation of Safety-Critical Systems.
DOI: 10.5220/0013022000004562

In *Proceedings of the 2nd International Conference on Cognitive Aircraft Systems (ICCAS 2024)*, pages 104-109
ISBN: 978-989-758-724-5

NASA TLX (Hart and Staveland, 1988a) or Likert-style questions (Likert, 1932). Such studies provide field data that contributes to the development of principles of human interaction with automated systems, and suggest specific implications for design.

One of the most popular strategies for understanding problems related to situation awareness is to conduct experimental studies with pilots in flight simulators. Sarter and Woods (1994) examined pilots' decision-making processes and situation awareness of the Flight Management System (FMS). They asked 20 pilots about their general knowledge of FMS functionality and then asked them to describe their reactions to hypothetical incidents that could not be simulated due to time restrictions.

Analyzing incident and accident data offers real-world examples of breakdowns in situation awareness. For example, Johnson and Pritchett (1995) conducted a study inspired by Air Inter 148 crashes, which were caused when pilots became confused by the autopilot interface (Bureau d'enquêtes et d'Analyse, 1992). They conducted an experimental simulator study that introduced mistakes in autopilot mode selection to test how well pilots could detect errors. They recorded when (or if) the error was detected and asked pilots what they thought caused the problem. This study improved understanding of the cues pilots use to maintain mode awareness and led to specific design implications.

Researchers also review incident reports to identify common factors that affect pilot's awareness (Eldredge et al., 1992; Jones and Endsley, 1996). For example, Mumaw (2020, 2021) classified incidents and accidents according to the source of confusion with respect to the state of the autoflight system. Silva and Hansman (2015) performed a similar analysis with respect to automation mode confusion to identify when and why confusion occurs. Although essential for understanding the causes of lapses in situation awareness, few studies contribute directly to the design of new cockpits.

2.2 Quantitative Approaches for Assessing System Performance

Designers of safety-critical systems must determine whether a new design offers better support for system awareness than existing designs. The most common approach is to employ quantitative methods that measure aspects of the system and/or user performance. Wei et al. (2013) suggested four kinds of methodology to assess situation awareness, such as physiological measurement, memory probe measurement, performance measurement and subjective measurement. Nguyen et al. (2019) summarize and discuss the advantages and disadvantages of six key measurement approaches for assessing situation awareness, including freeze-probe and real-time Probe techniques (Endsley, 1988; Wei et al., 2013), post-trial self-rating (Taylor, 2017; Waag and Houck, 1994), observer-rating (Matthews and Beal, 2002), performance-based rating (Gugerty, 2017; Tang et al., 2024) and process indices-based rating. Munir et al. (2022) also discuss the challenges associated with quantification of situation awareness.

Some studies provide potentially interesting implications for design. For example, Li et al. (2016) ran an eye tracking experiment to assess two FMA positions: on the far left of the MCP and at the top of the PFD (baseline). They found that placing FMA next to the FCU did not negatively affect pilot performance and could potentially increase pilots' situation awareness. Indeed, participants who looked at the FMA from the FCU position were slightly faster on the FMA, perhaps because the FCU changes less frequently than the PFD. These results suggest that repositioning the FMA may have benefits, a promising direction for future research.

2.3 Designing New Systems

Aviation designers have proposed multiple autopilot designs that seek to enhance situation awareness. For example, Hutchins (1996) observed operational autoflight mode management issues when he was sitting in the jumpseat during an incident. He introduced the *Integrated Mode Management Interface*, which combines control and autopilot state monitoring into a single interface, with the goals of simplifying the interface while improving mode awareness. He ran a comparative cognitive walkthrough study that suggested that this approach will eliminate or reduce the occurrence of certain errors.

Feary et al. (1999) proposed new FMA labels that indicate the purpose of the system rather than what the aircraft controls. They first conducted a survey of how pilots interpret and use current FMA displays and then, based on the survey results, evaluated a new FMA whose design was inspired by the situation awareness global assessment technique (SAGAT) (Endsley, 1988). They use quantitative measures to assess situation awareness and qualitative methods to observe behavior, but generating specific implications for design remains a topic for future research.

Boorman et al. (2004) developed a new autoflight interface design that emphasizes the target and who chooses it—the system or the pilot—rather than abstract modes. In order to assess their level of situation awareness, they asked 17 pilots to perform tasks and answer questions about the autoflight system's behavior (Mumaw et al., 2006; Prada et al., 2006). However, they did not measure pilots' subjective reactions to the system. Mumaw (2021) introduced a *feedback-oriented* display consisting of a lateral view and a vertical view. They evaluated the display in terms of the pilots' performance (time to first action), workload (NASA TLX) (Hart and Staveland, 1988b), subjective situation awareness (SART) (Taylor, 2017) and system usability (Brooke et al., 1996), as well as pilots' general comments that suggest possible improvements for the next iteration.

Rouwhorst et al. (2017) describe the process of designing a novel touch screen for selecting targets and engaging advanced modes. They evaluated an early design iteration by asking study participants to perform various descent scenarios in a flight simulator, using both a baseline and their new design. Participants were also asked to rate their own level of situation awareness. The results strongly influenced a major redesign, which was assessed in the same way. Although the quantitative measures of situation awareness showed no significant improvement over the baseline, the post-experiment question analysis revealed that the new touch screen led to lower situation awareness than the conventional autopilot.

These results indicate the potential benefits of combining quantitative and qualitative results. Although each of these studies explore interesting new design directions for autopilot systems, few benefit from a comprehensive approach that combines performance data and in-depth analysis from pilots about the system's strengths and weaknesses.

3 MIXED-METHOD APPROACH

Traditional aviation research uses both quantitative and qualitative methods to design and evaluate autopilot systems, but rarely at the same time. Qualitative methods are more common in the early stages of a user-centered design process (Mackay and Beaudouin-Lafon, 2023) and focus on better understanding the challenges that arise from a lack of situation awareness. They can provide rich insights into pilots' experiences, perceptions and behavior and help researchers consider nuances in dynamic and complex safety-critical systems.

By contrast, quantitative methods are more often used at the end of the design process, primarily to evaluate the effectiveness of a particular design, expressed in terms of statistical significance. Unfortunately, despite their potential for offering rich insights into both causes and mitigating factors related to situation awareness, qualitative methods remain marginal due to their supposed lack of rigor and objective data.

Even so, some researchers (Denzin, 2009) have shifted away from the idea that qualitative research fails to "adhere to canons of reliability and validation" (LeCompte and Goetz, 1982, p.31). Mackay and Fayard (1997) argue in favor of triangulating across research methods so as to mitigate the limitations of using a single approach. Mixed-method approaches that combine both quantitative and qualitative methods offers researchers complementary perspectives (Johnson and Onwuegbuzie, 2004) and help address the complexities of designing and evaluating safety-critical systems.

4 CASE STUDY: COMPARATIVE STRUCTURED OBSERVATION

Mackay and McGrenere (2024) introduce *comparative structured observation*, a mixed-method approach that borrows from best practices in the design of controlled experiments, including creating and ordering the presentation of comparable tasks, but emphasizes the collection of rich qualitative data over quantitative measures. This method diverges from traditional approaches that prioritize quantitative over qualitative data and takes advantage of expert users' ability to reflect upon and compare their experiences with alternative design variants. *Comparative structured observation* is well adapted for use within a participatory design approach (Mackay and Beaudouin-Lafon, 2023).

Comparative structured observation involves first constructing tasks that are grounded in real-world user practices and usually provide a challenge to the user. The researcher then observes as participants perform equivalent tasks with different design variants that are organized according to established experimental design practices, such as counter-balancing for order. Participants are asked to reflect on each design variant and compare them to each other. This results in richly detailed, grounded assessments of the advantages and disadvantages of each design variant.

Researchers can compare the design variants, but also compare their observations of participant behavior with the participants' analysis of their own behavior. Of course, *comparative structured observation* studies can also gather performance data, if the design prototype is sufficiently advanced. The goal is to gather nuanced insights about each design's strengths and weaknesses, based on diverse measures of situation awareness, to further the design of future safety-critical systems. The following elements should be considered when conducting a *comparative structured observation* in aviation to determine the impact of each variant on situation awareness.

Participants. Ideally, participants should be experts in the field of study. However, finding groups of experienced pilots to perform these tasks is challenging, due to their limited availability. An alternative is to involve advanced student pilots who have a deep grounding in the material but may be less likely to be biased in favor of one existing system or another. Despite their more limited experience, they are also more likely to uncover design flaws or usability issues that more experienced pilots would overlook given their over-training with the design.

Set-up. When assessing situation awareness in safety-critical systems, researchers face the choice of conducting experiments “in-the-field” (Salmon et al., 2006) or using flight simulators. Each option involves a set of considerations and constraints. In the context of autopilot design, real-flight experiments offer the most ecologically valid environments, but are severely limited by logistical challenges and safety concerns. On the other hand, flight simulators provide a controlled setting, but the authenticity varies greatly, ranging from high-fidelity simulators that replicate a fully interactive cockpit, including sounds, physical movement, and a realistic outside view, to low-fidelity simulators that are not interactive and do not accurately represent the cockpit.

Since the cockpit is a complex environment where information is distributed over various displays, the use of low-fidelity simulators may affect pilots' information-gathering strategy, with a corresponding impact on their level of situation awareness. Even so, this lack of information can also inform the researcher about a pilot's strategy for constructing their situation awareness for a given task. Similarly, providing ultra-realistic information displays may draw novice pilots' attention from their primary task of evaluating and comparing the design variants. The choice of simulator should align with the stage of the system's development, the chosen tasks and the participants' profiles. The key is to strike a balance between the fidelity of the environment and the pilot's ease of use and access.

Protocol. *Comparative structured observation* studies always employ a within-participant protocol where participants are exposed to equivalent tasks with different design variants, which allows them to make grounded, detailed comparisons. The study must also include at least two systems, either a baseline system that is compared to one or more design variants, or multiple design variants. Finally, tasks and systems should be counter-balanced for order, both within and across participants, for example, by using a Latin square.

The primary measure is the participants' qualitative assessment of each design variant, based on their experience using it to perform one or more tasks. When possible, participants should be asked to talk aloud during each task scenario and encouraged to reflect on their current experience with the system. After experiencing at least two variants, participants should be asked to compare them and explicitly consider both the positive and negative aspects of each.

In all cases, a researcher should observe participants as they perform the assigned tasks. At the end of the session, the researcher should run a debriefing interview to gather each pilot's general reflections about the tasks, scenarios, and, of course, the design variants being examined. Researchers may also include questions during the session, such as freeze-probe or real-time probe techniques for assessing situation awareness. The researcher should record qualitative data, including video, transcripts and hand-written notes, and, when relevant, capture subjective data, e.g. from Likert-scale questionnaires or ratings, and performance measures such as speed or error rate.

5 CONCLUSION

Traditional aviation research uses both quantitative and qualitative approaches to study pilots' situation awareness but rarely combines them to assess new design concepts. This paper argues that researchers who study safety-critical systems can benefit from using a mixed-method approach that explicitly combines quantitative and qualitative methods. We present *comparative structured observation*, a mixed-method approach that gathers rich insights from users about design variants combined with relevant subjective and performance measures, and briefly describe how to conduct a successful *comparative structured observation* study. The goal is not only to understand how users will interact with novel aviation designs, but also to identify potential design problems and implications for future designs. We hope that this paper will benefit researchers and practitioners working in aviation specifically and on safety-critical systems more generally, to gain a deeper understanding of how pilots will interact with innovative proposed designs.

REFERENCES

Boorman, D. J., Mumaw, R. J., Pritchett, A., and Jackson, A. (2004). A new autoflight/fms interface: Guiding design principles. In *Proceedings of the international conference on human-computer interaction in aeronautics*, pages 303–321.

Brooke, J. et al. (1996). Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.

Bureau d'enquêtes et d'Analyse, . . (1992). Accident survenu le 20 janvier 1992 près du mont sainte-odile (bas rhin) à l'airbus a 320 immatriculé f-gged exploité par la compagnie air inter. Technical report, BEA, FR.

Bureau of Air Safety Investigation, . . (1998). Advanced technology aircraft safety survey report. Technical report, BASI, AU.

Curry, R. E. (1985). The introduction of new cockpit technology: A human factors study. *NASA TM 86659*.

Denzin, N. K. (2009). The elephant in the living room: Or extending the conversation about the politics of evidence. *Qualitative research*, 9(2):139–160.

Eldredge, D., Mangold, S., and Dodd, R. S. (1992). A review and discussion of flight management system incidents reported to the aviation safety reporting system. Technical Report 5855, U.S. Department of Transportation, Bureau of Transportation Statistics.

Endsley, M. R. (1988). Situation awareness global assessment technique (sagat). In *Proceedings of the IEEE 1988 national aerospace and electronics conference*, pages 789–795. IEEE.

Feary, M., Alkin, M., Polson, P., McCrobie, D., Sherry, L., and Palmer, E. (1999). Aiding vertical guidance understanding. *Air & Space Europe*, 1(1):38–41.

Flanagan, J. C. (1954). The critical incident technique. *Psychological bulletin*, 51(4):327.

Gugerty, L. J. (2017). Situation awareness during driving: Explicit and implicit knowledge in dynamic spatial memory. In *Situational awareness*, pages 379–404. Routledge.

Hart, S. G. and Staveland, L. E. (1988a). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Hancock, P. A. and Meshkati, N., editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland.

Hart, S. G. and Staveland, L. E. (1988b). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier.

Hutchins, E. L. (1996). *The Integrated Mode Management Interface*. NASA, USA.

Johnson, E. N. and Pritchett, A. R. (1995). Experimental study of vertical flight path mode awareness. *IFAC Proceedings Volumes*, 28(15):153–158.

Johnson, R. B. and Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational researcher*, 33(7):14–26.

Jones, D. G. and Endsley, M. R. (1996). Sources of situation awareness errors in aviation. *Aviation, space, and environmental medicine*, 67(6):507–512.

Kharoufah, H., Murray, J., Baxter, G., and Wild, G. (2018). A review of human factors causations in commercial air transport accidents and incidents: From to 2000–2016. *Progress in Aerospace Sciences*, 99:1–13.

LeCompte, M. D. and Goetz, J. P. (1982). Problems of reliability and validity in ethnographic research. *Review of educational research*, 52(1):31–60.

Li, W.-C., White, J., Braithwaite, G., Greaves, M., and Lin, J.-H. (2016). The evaluation of pilot's situational awareness during mode changes on flight mode annunciators. In Harris, D., editor, *Engineering Psychology and Cognitive Ergonomics*, pages 409–418, Cham. Springer International Publishing.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55.

Mackay, W. and McGrenere, J. (2024). Comparative structured observation. *ACM Transactions on Computer-Human Interaction (TOCHI)*, pages 1–25. to appear.

Mackay, W. E. and Beaudouin-Lafon, M. (2023). Participatory design and prototyping. In *Handbook of Human Computer Interaction*, pages 1–33. Springer.

Mackay, W. E. and Fayard, A.-L. (1997). Hci, natural science and design: A framework for triangulation across disciplines. In *Proceedings of the 2nd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, DIS '97, page 223–234, New York, NY, USA. Association for Computing Machinery.

Matthews, M. D. and Beal, S. A. (2002). Assessing situation awareness in field training exercises. *US Army Research Institute for the Behavioral and Social Sciences*, 31.

Mumaw, R., Boorman, D. J., and Prada, R. (2006). Experimental evaluation of a new autoflight interface. In *Proceedings HCI-Aero 2006, International Conference on Human Computer Interaction, Seattle, WA*.

Mumaw, R. J. (2020). Addressing mode confusion using an interpreter display. *NASA Contractor Report*.

Mumaw, R. J. (2021). Plan b for eliminating mode confusion: An interpreter display. *International Journal of Human–Computer Interaction*, 37(7):693–702.

Munir, A., Aved, A., and Blasch, E. (2022). Situational awareness: techniques, challenges, and prospects. *AI*, 3(1):55–77.

Nguyen, T., Lim, C. P., Nguyen, N. D., Gordon-Brown, L., and Nahavandi, S. (2019). A review of situation awareness assessment approaches in aviation environments. *IEEE Systems Journal*, 13(3):3590–3603.

Prada, L. R., Mumaw, R. J., Boehm-Davis, D. A., and Boorman, D. J. (2006). Testing boeing's flight deck of the future: A comparison between current and prototype autoflight panels. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(1):55–58.

Rouwhorst, W., Verhoeven, R., Suijkerbuijk, M., Bos, T., Maij, A., Vermaat, M., and Arents, R. (2017). Use of touch screen display applications for aircraft flight control. In *2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC)*, pages 1–10. IEEE.

Salmon, P., Stanton, N., Walker, G., and Green, D. (2006). Situation awareness measurement: A review of applicability for c4i environments. *Applied ergonomics*, 37(2):225–238.

Sarter, N. B. and Woods, D. D. (1992). Pilot interaction with cockpit automation: Operational experiences with the flight management system. *The International Journal of Aviation Psychology*, 2(4):303–321.

Sarter, N. B. and Woods, D. D. (1994). Pilot interaction with cockpit automation: Ii. an experimental study of pilots' model and awareness of the flight management system. *The International Journal of Aviation Psychology*.

Sarter, N. B. and Woods, D. D. (1997). Team play with a powerful and independent agent: Operational experiences and automation surprises on the airbus a-320. *Human factors*, 39(4):553–569.

Silva, S. S. and Hansman, R. J. (2015). Divergence Between Flight Crew Mental Model and Aircraft System State in Auto-Throttle Mode Confusion Accident and Incident Cases. *Journal of Cognitive Engineering and Decision Making*, 9(4):312–328.

Tang, H., Lee, B. G., Towey, D., and Pike, M. (2024). The impact of various cockpit display interfaces on novice pilots' mental workload and situational awareness: A comparative study. *Sensors*, 24(9):2835.

Taylor, R. M. (2017). Situational awareness rating technique (sart): The development of a tool for aircrew systems design. In *Situational awareness*, pages 111–128. Routledge.

Waag, W. L. and Houck, M. R. (1994). Tools for assessing situational awareness in an operational fighter environment. *Aviation, space, and environmental medicine*, 65(5 Suppl):A13–9.

Wei, H., Zhuang, D., Wanyan, X., and Wang, Q. (2013). An experimental analysis of situation awareness for cockpit display interface evaluation based on flight simulation. *Chinese Journal of Aeronautics*, 26(4):884–889.

Wiener, E. L. (1985). Human factors in cockpit automation: A field study of flight crew transition. *NASA CR 177333*.

Wiener, E. L. (1989). Human factors of advanced technology (" glass cockpit") transport aircraft. *NASA CR 177528*.