2021

Aviation Human-in-the-Loop Simulation: Best Practices for Subjective Performance Measurement

John Kleber
Embry-Riddle Aeronautical University

Beth Blickensderfer
Embry-Riddle Aeronautical University

Aviation Human-in-the-Loop Simulation: Best Practices for Subjective Performance Measurement

John Kleber, M.S., Beth Blickensderfer, Ph.D.
Embry-Riddle Aeronautical University, Daytona Beach, FL

Subjective performance measurements are a useful tool for researchers and instructors to evaluate tasks that are difficult to quantify with objective data pulled from a simulator or the physiological data of pilots. Subjective performance measurements are non-intrusive measures typically conducted by human raters. Some recommendations for utilizing subjective measures include reducing the workload of the human raters, concealing the aim of the experiment from participants, utilizing multiple raters to evaluate each participant, providing raters with proper training, developing error-resistant rater forms and, including both subjective and objective measures when possible.

INTRODUCTION

Simulations have been a vital component of aviation training and research since the early days of flight. Today, using digital technology, aviation simulations have continual improvements in physical fidelity at lower cost. The purpose of this paper is to review the characteristics of subjective performance measures used in aviation simulation research and discuss best practices for implementing these measures.

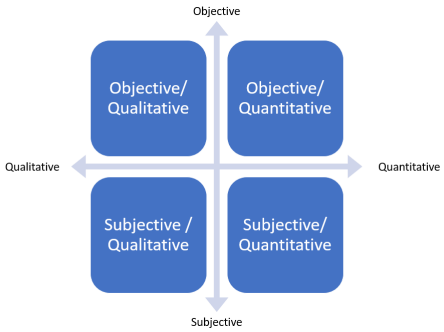
Human Performance in Aviation Simulation

In broad terms, human performance is the ability of an individual or team of individuals to complete a task (Gawron, 2019). Human Factors practitioners use human performance measures in system evaluations, research, and training. Examples of using human performance measures in conjunction with aviation simulation include: comparing crew communications in a two-pilot cockpit to a three-pilot cockpit, measuring a pilot's proficiency and providing feedback in a simulated flight, assessing a pilot's ability to effectively execute flight maneuvers as part of testing a particular control system, and measuring a flight crew's ability to coordinate during an emergency or a pilot's communication with Air Traffic Control (ATC).

Hall and Brannick (2009) classified performance measures along two dimensions (Objective versus Subjective; and Quantitative versus Qualitative) (see Figure 1). Regarding the objective – subjective dimension, objective measures often include data collected directly from the simulator (airspeed, altitude, etc.) or physiological measures (heart rate, pupil dilation, etc.). One example of an objective measure is a pilot's response time. Landman et al. (2017) measured response time by programming a simulator to calculate the time from tailwind onset to first action (i.e., disengaged auto-throttle, change in rudder input, etc.). Conversely, subjective measures derive from the opinions or judgments of the subjects (i.e., a self-rating) or human observer-raters. Effective human observer-raters have expertise in the domain of interest. In aviation, one common example of human raters are flight instructors tasked with grading some aspect of a pilot's or flight crews' performance. For example, a flight instructor may assign a "Pass or Fail" grade for a pilot's execution of flight maneuvers.

Figure 1

Classification of Performance Measures



Performance measures can also be categorized as either quantitative or qualitative (Hall and Brannick, 2009). Quantitative measures are numerical and can be either discrete or continuous. Conversely, qualitative measures are non-numerical and provide categorical information.

Objective Performance Measures. In aviation simulations, objective performance measures often incorporate quantitative data pulled directly from the simulator. This quantitative data can be used in various ways, such as providing frequency counts, making simple comparisons (e.g., comparing a pilot's flight route to the optimum flight route), or various metrics can be combined to create composite scores. For example, Taylor et al. (2007) compared the raw scores of more than 20 variables to the means and standard deviations of previously collected baseline measures to produce z-scores that aggregated the scores to produce an overall flight score.

While less common, objective measures can also be qualitative (see upper left quadrant of Figure 1). One way of generating objective qualitative measures is to compare data from a simulation to set standards. For example, a researcher may program a simulator to compare a plane's flight path with the location of weather systems to categorize the flight as either compliant with Visual Flight Rules (VFR) or non-compliant.

A benefit of using objective performance measures is that simulators can automatically collect data at a speed and consistent level of precision that is difficult to match with human raters (Atkinson et al., 2018). While collecting objective data is fast and reliable, there are some limitations to this data. One limitation is that objective measures lack sensitivity to underlying factors. For example, King (2020) recorded a frequency count of how often a pilot accessed Automatic Terminal Information Service (ATIS) inflight. However, the frequency data alone does not account for why the pilot was or was not frequently accessing ATIS. The pilot may routinely check ATIS for updates or because they are struggling to hear or interpret the weather information. On the other hand, a pilot may check ATIS once and receive all the data necessary or, they may review ATIS once and deem it uninterpretable.

Another limitation with objective measures for human-in-the-loop aviation simulations is the difficulties in evaluating interpersonal communications. A key component of flight is calm and efficient communications between the pilot and co-pilot as well as the pilot and air traffic controllers. Effective communication is particularly essential during emergency operations. While, objective measures can provide information on communication frequency, communication quality (e.g., inflection, tone, or meaning) requires a more nuanced approach. Furthermore, measures of communication quality have a significantly stronger relationship with performance than do measures of frequency (Marlow et al., 2018).

Subjective Performance Measures Subjective measures are typically provided by human raters (Hebbar & Pashilkar, 2016). Subjective measures have been used throughout the flight process to evaluate many aspects of pilot performance, including crew communications and execution of flight maneuvers. Raters can evaluate performance in real-time or after the fact using recordings of the simulation (e.g., King, 2020 and Müller & Giesa, 2002).

Looking again at Figure 1, Hall and Brannick (2009) split subjective measures into two groups (quantitative or qualitative). An example of a subjective, quantitative measure would be a human judge assigning a pilot's execution of a landing a score of 5 for safety on a rating scale of 1 to 10. However, if the rating sheet allowed the judge to follow-up the score with additional comments (e.g., the pilot entered at an unsafe angle and speed), the comments would be considered subjective and qualitative.

A primary concern with subjective measures is reliability. Maintaining consistency among evaluations can be difficult when using human judgments. Raters must attend to all aspects of interest during the simulation (Aamodt, 2007). When a rater observes a target behavior, they must hold the observation in memory, then compare the observation to the rating sheet, and decide where the observation falls in the rating scale. This can result in the opportunity for human error,

particularly in situations where a rater is evaluating performance in real-time (e.g., increased time-pressure). When assessing performance in real-time, the rater may take their eyes off the simulation to mark notes on the rating sheet. This can result in the rater missing an action performed by the pilot (Aamodt, 2007).

High workload amongst raters has also been shown to negatively affect subjective measures' accuracy (Bretz et al., 1992; Kahneman, 1973, as cited in Atkinson et al., 2016). Raters often must attend to multiple aspects of the simulation. If a pilot performs several actions in quick secession, the rater must continue to monitor while attending to each action in their working memory until they can evaluate and rate them individually. This can result in errors caused by either the amount of presented information surpassing the processing ability of the rater's working memory (Ryu and Myung, 2005) or from decay due to the increase in time between the rater witnessing the action and evaluating it (Barrouillet et al., 2011).

Another issue with subjective performance measures in aviation simulation is that pilots operating the simulation know that researchers monitor their actions. This awareness, also known as the Hawthorne effect, may bias the pilot's performance, resulting in either a positive or negative impact that potentially reduces the measure's external validity (McCambridge et al., 2014). For example, suppose a flight crew is aware that researchers are evaluating their communication. In that case, they might take extra caution to speak clearly and calmly (positive effect) or make more mistakes due to the anxiety of being watched (negative impact). While the measurement will be accurate for that simulation run, because of the potential bias, the assessment may not generalize to their actual, day-to-day flight behaviors.

BEST PRACTICES

This section describes some recommended best practices for the effective use of subjective performance measurements in aviation human-in-the-loop simulations.

Recommendation 1: Reduce the Workload of Raters.

As mentioned previously, a high rater workload negatively affects subjective measures' accuracy and effectiveness (Bretz et al., 1992; Kahneman, 1973, as cited in Atkinson et al., 2016). Researchers can reduce the workload by recording the simulation so that raters can pause or rewind the section of the simulation. Another option includes breaking up the areas of interest amongst multiple raters. This can be of particular use when raters have to evaluate the pilot in real-time without the ability to record and view the simulator at a later time.

Recommendation 2: Conceal the Aim of the Purpose for the Assessment

One precautionary measure to mitigate the potential for a Hawthorne Effect is to conceal the purpose of the assessment. Participants are less likely to be biased if they believe the researcher's attention is focused elsewhere. For example, Landman et al. disguised their study as a two-staged experiment where would validate the simulator's "aerodynamic model" then evaluate the fidelity of spatial disorientation illusions in the simulators (2017). The study's real aim was to assess the influence of surprise on airline pilot's ability to recover from a stall. The surprise condition occurred during the initial stage of the study, which researchers told them was designed not to evaluate their performance but instead the simulator's "aerodynamic model."

Recommendation 3: Utilize Multiple Raters to Evaluate Each Participant

Include at least two raters in the evaluation process. Increasing the number of raters who evaluate each subjective measure can improve the measure's overall reliability (Hall & Brannick, 2009). Researchers should calculate a measure of inter-judge reliability with their findings. To obtain accurate inter-judge reliability, researchers should require each rater to evaluate every participant's performance.

Recommendation 4: Provide Raters with Proper Training.

All raters should receive training on the tasks that they are assigned to evaluate. One of the most common rater-training methods is Frame-of-Reference (FOR) training (Bernardin & Buckley, 1981). FOR training focuses on provided raters with examples, practice, and feedback. Research has shown FOR training improves the accuracy of human raters (Roch et al., 2012).

Recommendation 5: Develop Error-Resistant Rater Forms.

Along with training, another way of reducing error amongst raters is designing user-friendly rater forms. Rater forms should clearly state instructions for raters and provide well-defined grading parameters. User-friendly forms remove uncertainty for raters and reduce training time.

Recommendation 6: Include Both Subjective and Objective Measures.

As established previously, both subjective and objective measures have limitations. For example, objective measures would have a limited ability to determine if a pilot giving orders to the co-pilot is speaking assertively or aggressively. On the other hand, subjective measures may suffer from low reliability if the rater is not adequately trained or their instructions are too ambiguous. However, the use of both subjective and objective measures can give a more holistic and meaningful assessment (Hebbar & Pashilkar, 2016).

CONCLUSION

In summary, this paper describes the characteristics and issues associated with using subjective measures used for evaluated performance in aviation simulations. Based on previous literature, we discussed best practices for conducting simulation performance evaluations with subjective measures:

- Reduce the Workload of Raters.
- Conceal the Aim of the Purpose for the Assessment.
- Utilize Multiple Raters to Evaluate Each Participant
- Provide Raters with Proper Training.
- Develop Error-Resistant Rater Forms.
- Include both Subjective and Objective measures.

DISCLAIMER

The views expressed in this paper are those of the authors and do not necessarily represent the organization with which they are affiliated.

REFERENCES

Aamodt, M. G. (2007). Evaluating employee performance. In J. Kim & D. Moneypenny (Eds.), *Industrial/organizational psychology: An applied approach* (5th ed., pp. 244-246). Thomson-Belmont Wadsworth.

Atkinson, B., Tindall, M., Sheehy, M., & Bailey, H. (2016). Answering the call for analytics within the maritime patrol community. *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) Proceedings*.

Atkinson B., Tindall M.J., Killilea J.P., & Anania E.C. (2018, July) *Advancing performance assessment for aviation training*. *Proceedings of the AHFE 2018 International Conference on Human Factors in Training, Education, and Learning Sciences,* Orlando, FL, United States.

Barrouillet, P., Portrat, S., Vergauwe, E., Diependaele, K., & Camos, V. (2011). Further evidence for temporal decay in working memory: Reply to Lewandowsky and Oberauer (2009). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(5), 1302–1317.

Bernardin, H. J., & Buckley, R. M. (1981). Strategies in rater training. *Academy of Management Review, 6,* 205–212.

Bretz, R. D., Milkovich, G. T., & Read, W. (1992). The current state of performance appraisal

research and practice: Concerns, directions, and implications. *Journal of Management, 18*(2), 321-352.

Gawron, V. J. (2019). Human Performance. *Human performance and situation awareness measures* (3rd ed., pp. 13). CRC Press. https://doi-org.ezproxy.libproxy.db.erau.edu/10.1201/9780429019562

Hall, S., & Brannick, M. T. (2009). Performance assessment in simulation. *Human factors in simulation and training*, 149-168.

Hebbar, P. A., & Pashilkar, A. A. (2017). Pilot performance evaluation of simulated flight approach and landing manoeuvres using quantitative assessment tools. *Sādhanā, 42*(3), 405-415.

Kahneman, D. (1973). *Attention and effort* (p. 246). Englewood Cliffs, NJ: Prentice-Hall.

King, J. (2020). An aviation weather preflight decision support tool to improve GA pilots preflight and inflight performance. *[Doctoral dissertation, Embry-Riddle Aeronautical University]*. ERAU Campus Repository. https://commons.erau.edu/edt/538/

Landman, A., Groen, E. L., Van Paassen, M. M., Bronkhorst, A. W., & Mulder, M. (2017). The influence of surprise on upset recovery performance in airline pilots. *The International Journal of Aerospace Psychology, 27*(1-2), 2-14.

Marlow, S. L., Lacerenza, C. N., Paoletti, J., Burke, C. S., & Salas, E. (2018). Does team communication represent a one-size-fits-all approach?: A meta-analysis of team communication and performance. *Organizational Behavior and Human Decision Processes, 144*, 145-170. https://doi.org/10.1016/j.obhdp.2017.08.001

McCambridge, J., Witton, J., & Elbourne, D. R. (2014). Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of clinical epidemiology, 67*(3), 267-277.

Müller, T., & Giesa, H. G. (2002). Effects of airborne data link communication on demands, workload and situation awareness. *Cognition, Technology & Work, 4*(4), 211-228.

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta‐analytic review of frame‐of‐reference training. *Journal of Occupational and Organizational Psychology, 85*(2), 370-395. https://doi.org/10.1111/j.2044-8325.2011.02045.x

Ryu, K., & Myung, R. (2005). Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics, 35*(11), 991-1009. https://doi.org/10.1016/j.ergon.2005.04.005

Taylor, J. L., Kennedy, Q., Noda, A., & Yesavage, J. A. (2007). Pilot age and expertise predict flight simulator performance: A 3-year longitudinal study. *Neurology, 68*(9), 648-654.