Analysis of Pilot Feedback Regarding the Use of State Awareness Technologies During Complex Situations

Emory Evans, Steven D. Young, Taumi Daniels, and Yamira Santiago-Espada
NASA Langley Research Center, Hampton, VA

Tim Etherington
Rockwell-Collins, Cedar Rapids, IA

***Abstract*— A flight simulation study was conducted at NASA Langley Research Center to evaluate flight deck systems that (1) predict aircraft energy state and/or autoflight configuration, (2) present the current state and expected future state of automated systems, and/or (3) show the state of flight-critical data systems in use by automated systems and primary flight instruments.**

Four new technology concepts were evaluated vis-à-vis current state-of-the-art flight deck systems and indicators. This human-in-the-loop study was conducted using commercial airline crews. Scenarios spanned a range of complex conditions and several emulated causal factors and complexity in recent accidents involving loss of state awareness by pilots (e.g. energy state, automation state, and/or system state).

Data were collected via questionnaires administered after each flight, audio/video recordings, physiological data, head and eye tracking data, pilot control inputs, and researcher observations. This paper strictly focuses on findings derived from the questionnaire responses. It includes analysis of pilot subjective measures of complexity, decision making, workload, situation awareness, usability, and acceptability.

Keywords—Energy, automation, system, complexity, decision making, usability, workload, situation awareness, acceptability

I. Introduction

The Automation and Information Management Experiment (AIME) was conducted at NASA Langley Research Center to evaluate flight deck systems that (1) predict aircraft energy state and/or autoflight configuration, (2) present the current state and expected future state of automated systems, and/or (3) show the state of flight-critical data systems in use by automated systems and primary flight instruments. Four new flight deck display concepts/technologies were evaluated vis-à-vis a representative baseline of current state-of-the-art flight deck systems and indicators.

AIME was conducted over several weeks from November 2015 to January 2016 using 11 commercial airline crews. Each crew completed on average 20 flight scenarios over two days in the flight simulator. In total, over 220 flights were completed.

Several types of data were recorded during each flight including aircraft state parameters, audio, video, and physiological measures (electrocardiogram, respiration, skin conductance). After each flight, each pilot completed a subjective measures questionnaire. A summary debrief was conducted at the end of the second day for each crew.

This paper focuses on the analysis of the questionnaire responses pertaining to complexity, decision making, workload, situation awareness, usability, and acceptability. Representative observations derived from pilot comments on the new technology concepts are also included. For a broader view of AIME, including motivation, purpose, and results not presented here, see [1], [2], and [3].

II. Technologies Under Evaluation

In this experiment, five display conditions were used. Boeing 787-like flight deck displays and indicators served as a reference or baseline (BL) condition. The other four conditions were new technology concepts that augmented this baseline. All five included the following displays:

- Primary Flight Display (PFD) and Head-Up Display (HUD)
- Navigation Display (ND) and Vertical Situation Display (VSD)
- Engine Indicating and Crew Alerting System (EICAS)
- Synoptics displays
- Lower MFD (LMFD), serving primarily as the pilot interface to the Flight Management System (FMS)
- ATC communications display
- Electronic Flight Bags (EFB)

These displays were organized in a B-787-like layout as shown in Figure 1.

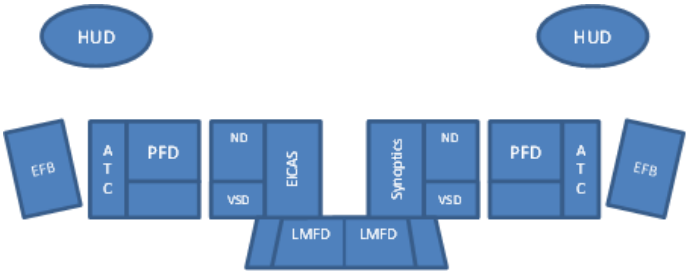


Fig. 1. AIME flight deck display layout.

The four new technology concepts evaluated were:

- ME - Maneuver Envelope display and effects estimation function; information provided on the PFD
- TP - Trajectory Prediction function; information provided on the VSD and ND
- SIS - System Interaction Synoptic; information provided as an additional tab on the Synoptics display, with associated simplified checklists on the EFBs
- PAE - Predictive Alerting of Energy-related problems; information provided on EICAS, ND, and VSD

These new indicators and functions, as implemented for AIME, are described in [1], [2], and [3]. Figures 2 through 6 provide examples of the display elements.

ME estimates and monitors the maneuvering envelop of the aircraft and provides visual awareness of the changing relationship between airspeed, bank angle, angle of attack, and lift via maneuverability bands (amber and red/black segments) that dynamically update in flight. The bands are shown on the airspeed indicator, vertical speed indicator, and bank angle indicator. For more on ME see [4].

Fig. 2. Example ME Indicators on the PFD.

TP predicts where the auto-flight system will take the aircraft and where mode changes will occur if the pilot were to make no additional inputs. It indicates this as a green line on the VSD and ND, with a constant length of five minutes. Green circles, labeled with the mode change, indicate where mode changes are predicted to occur. The white line (see Fig. 3) is not part of TP, but predicts where the aircraft will go based on inertia and is standard functionality on B-787-like displays. Pilot inputs (e.g. extending speedbrakes or changing commanded airspeed) cause these indicators to change based on new predictions. For more on TP see [5].

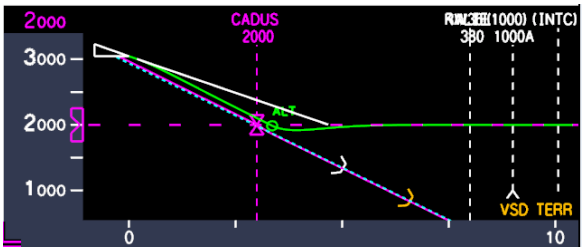


Fig. 3. Example TP indications on VSD; predicting ALT hold transition near CADUS.

SIS graphically depicts whether flight-critical data is valid, the state of the sources of the data, and the effect on systems that receive the data should a relevant failure occur. It supplements checklists by providing information in a graphical form that would otherwise typically be given as textual "notes" in the checklists. For AIME, the relevant standard non-normal checklists were shortened to remove this information, which was now conveyed by SIS. For more information on SIS see [1] and [2].

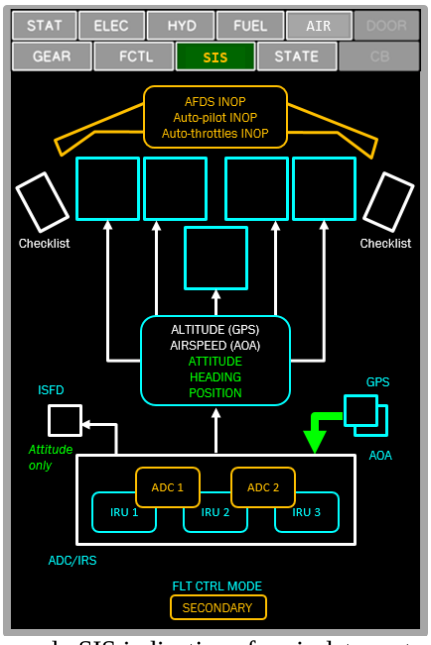


Fig. 4. Example SIS indications for air data system failure.

PAE provides predictive information related to where energy-related problems will occur if the current course of action is continued. The prediction is indicated as a cyan circle and label on the TP-generated green line. In conjunction with any predicted condition, an EICAS message states the type of energy-related problem being predicted. A triangle in front of the EICAS message indicates a predicted condition, not one that has already occurred. For more information on PAE see [1], [2], and [3].

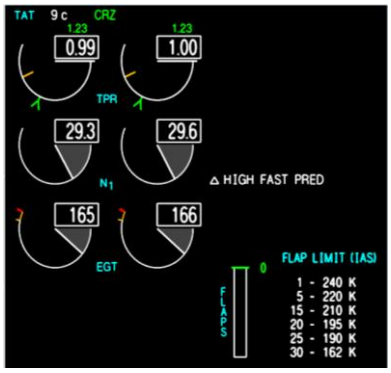


Fig. 5. Example PAE indication on EICAS; predicting high and fast on approach.

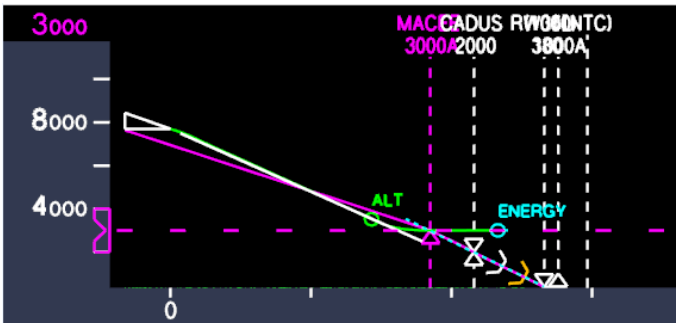


Fig. 6. Example PAE indications on the VSD; predicting high and fast on approach (ENERGY label and symbol).

ME and TP are the result of previous research and have a higher Technology Readiness Level (TRL). The SIS and PAE technology concepts are less mature, with AIME being the first opportunity for testing against design requirements and assessed for usability by pilots.

III. Experiment Overview

A. *Test Subjects*

Eleven two-pilot airline crews participated. Each crew spent two days on-site and consisted of a Captain and First Officer from the same airline. Type ratings covered Boeing 787, 777, 767, 757, 747, 737, and Airbus 330, 321, 320, 319. Flight experience ranged from 5,000 to 39,000 hours, with an average of 15,500 hours.

B. *Scenarios*

Six types of scenarios were defined for AIME:

1. Nominal
2. Loss of flight-critical data
3. Increased risk of low energy
4. Increased risk of high energy
5. Unanticipated automation interaction required
6. Distractions affecting workload management

Specifics for these are summarized in Table I. Types 2 – 6 include off-nominal events. Off-nominal events were included to help expose state awareness issues and/or to mimic relevant accidents/incidents. This resulted in 20 unique scenarios to draw from for the run sequence for each crew. The flight time for each scenario was on the order of 15-20 minutes, followed by the administration of questionnaires with a web-based tool developed for this experiment. Weather cases (Wx) in Table I are coded as:

0. VMC, Ceiling 1000 ft, Visibility 3 miles; Calm; Dusk
1. IMC, Ceiling 500 ft, Visibility 2400 ft; Storm, light turbulence and winds; Night
2. IMC, Ceiling 200 ft, Visibility 1800 ft; Storm, moderate turbulence and winds; Night
3. IMC, Ceiling 100 ft, Visibility 1000 ft; Storm, moderate turbulence and winds; Night
4. Clear-and-visibility ok (CAVOK); Dusk

TABLE I. AIME SCENARIO SUMMARIES BY TYPE.

Type	Wx	Off-nominal(s)
1	0	None, follow published STARs (KMEM) or SID (KDEN), use airline procedures
2	1,3	(a, b) Failure within the pitot-static system due to icing/blockage of the pitot and/or static ports, and/or failure of the pitot heat system; airspeed and/or altitude become unreliable; (c, d) IRS, IRU, and/or AHRU fail
3	0,3	(a) Radar altimeter reports -8 ft at ~2000 ft after GS captured; causes A/T to go to ROLLOUT FLARE mode, returns throttles to idle position even if the pilot pushes them forward (i.e. A/T remains engaged unless pilot disconnects); (b) Aircraft in front slows; ATC asks ownship to slow to maintain spacing; (c) At mid-point between turn-to-final and FAF, A/T disconnects; (d) ATC issues vectors for ILS approach, then "descend to 3000, slow to 180", after "Contact tower" issues "Runway change to 18R" (farther parallel runway); after A/T retracts to idle, A/T disconnects
4	0,4	(a) NOTAM: "Glideslope Out of Service"; ATC issues clearances that set crew up high, then clears for visual approach; (b) ATC issues two "direct-tos" causing crew to be high and fast at 4000 ft, ATC says expect runway 36L; but after turn onto final, ATC issues clearance to parallel (closer) runway; if crew doesn't go around by 300 ft, ATC calls for go-around; 4-knot tailwind on final approach
5	0,1,2	(a) Tailwind encountered is significantly higher than forecast winds in the FMS, causes unexpected VNAV transitions during downwind leg; (b) After passing 6000 ft, ATC issues "Level at 4000, hold at present position", after clearance, A/T disconnects without audible annunciation; (c) During an RNAV approach, ATC calls for unexpected go-around at low altitude (~300 ft AGL)
6	1,2	(a) EICAS message "HYD SYS PRESS (CENTER ONLY)" shortly after start; later at 3000 ft, ATC issues "Low level wind shear, +/- 10 knots, 1000 ft on approach to Runway 18C"; (b) ANTISKID warning on EICAS after start; Use shorter runway, runway conditions reported as "poor"

C. Facilities

AIME was conducted in the Research Flight Deck (RFD) within the Cockpit Motion Facility (CMF) at NASA's Langley Research Center (Figure 7). For AIME, the RFD was configured to emulate B-787 displays and functions.

Fig. 7. RFD internal (top) and external (bottom).

IV. Results and Discusion

Questionnaires were collected over a testing period of several weeks, with more than 220 flights completed. As shown in Table II, the total number of questionnaires received was 400. For some flights, questionnaires were not completed either due to inadequate time, or a connectivity issue with the questionnaire tool. Data were collected for Pilot Flying (PF) and Pilot Monitoring (PM). Each pilot individually completed a questionnaire after each flight. The pilots alternated roles periodically so that each pilot assumed each role for about half the scenarios. Roles were not changed during a flight, only between flights.

Selected results from the subjective measures are presented in the following sections for complexity, decision making factors, workload, situation awareness, usability, and acceptability. For each of these sets of data, analysis of variance was conducted to assess the impact of PF vs PM. There was no statistically significant difference between means as determined by one-way ANOVA ($\alpha = 0.05$). Therefore no post-hoc statistical tests were performed.

Table II. Completed Pilot Questionnaires By Scenario Type and Display Condition

	Scenario Type						Total
	1	2	3	4	5	6	
BL	10	36	32	28	40	16	162
ME	16	0	8	12	18	0	54
TP	37	0	14	8	18	10	87
SIS	0	41	0	0	0	0	41
PAE	0	0	28	22	4	2	56
Total	63	77	82	70	80	28	400

Comparisons involving PAE were the closest to showing statistically significant differences between PF and PM.

Some scenarios were flown twice by the same crew, once with the baseline (BL) technology condition being evaluated and once with one of the new technologies being evaluated. This occurred 14 times for ME and PAE, 16 for SIS, and 18 for TP. Order was randomized. For each "same scenario" pair, the difference in score between the evaluated technology and baseline was calculated. These score differences were then averaged for each technology to show the change from baseline.

A. Complexity

After each flight, pilots rated the complexity of (a) the task, (b) the operational environment, (c) the system/automation, and (d) the information provided. Ratings were based on a scale from 1 (not complex) to 10 (extremely complex). Figure 8 shows the mean perceived complexity of the task and the operational environment, along with +/-2 standard errors (SE) to represent the variability in responses.

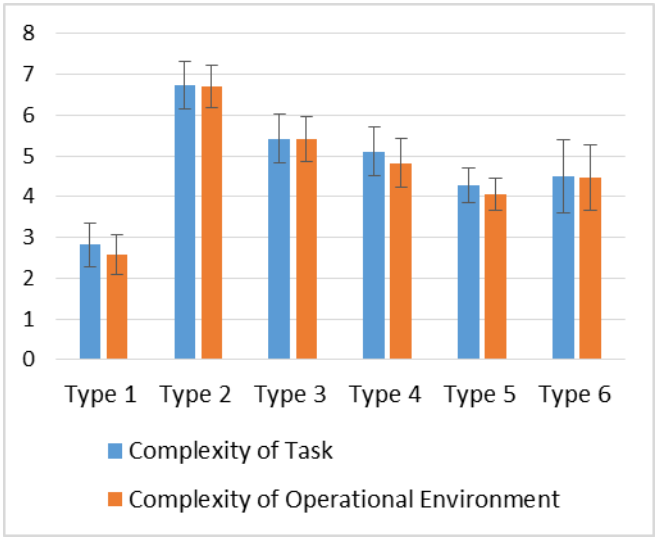


Fig. 8. Complexity by scenario type.

As can be seen, Type 2 scenarios presented the greatest complexity. Type 1 scenarios, which had no off-nominal events, presented the least. For this graphic, the results for PF and PM are combined. These results help to validate the scenario designs (e.g. Type 2 was designed to be the most complex of all the scenarios).

Figure 9 shows the mean difference in perceived complexity of the system/automation and the information provided for the evaluated technologies versus the BL condition, based on same-scenario pairwise comparisons.

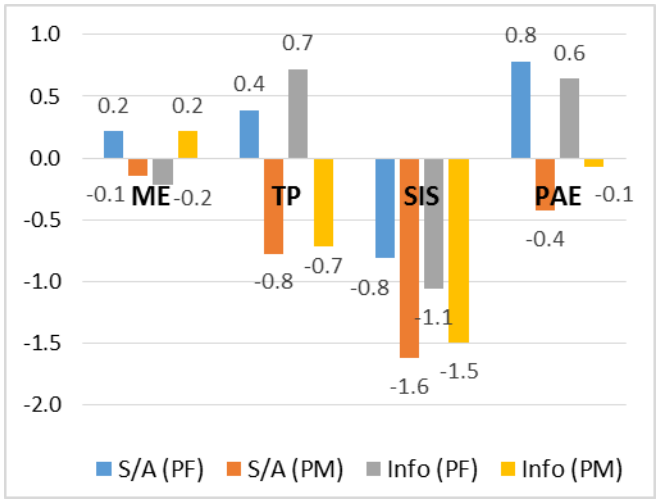


Fig. 9. Difference in system/automation and information complexity versus baseline.

Positive values indicate higher perceived complexity for the evaluated technology than the baseline (BL). Negative values indicate lower complexity than BL. For ME the differences from BL are not remarkable, indicating that there is minimal perceived effect on clutter. With TP and PAE, the complexity of the system/automation and the information provided increased from BL for PF and decreased for PM. This is particularly encouraging for the PM role as reference to these indicators is more in line with the monitoring function. With SIS, this form of complexity was reduced for PF, and even more so for PM. This reinforces the design concept that such graphical representations can be more intuitive.

B. *Decision Making*

After each flight, pilots were asked to rank the top three factors, in order, that most influenced their decisions with respect to maintaining safety of flight. The aggregate responses for the number one factor are shown in Figure 10.

The highest ranked decision-making factor across all pilots and all flights was flight path management (for both PF and PM). Below that was autoflight system state for PF and aircraft system state for PM. Communication with ATC was ranked first the least.

Overall the top three factors for maintaining safety of flight were: (1) Flight path management, (2) Autoflight system state, (3) Energy state / Aircraft system state (tie).

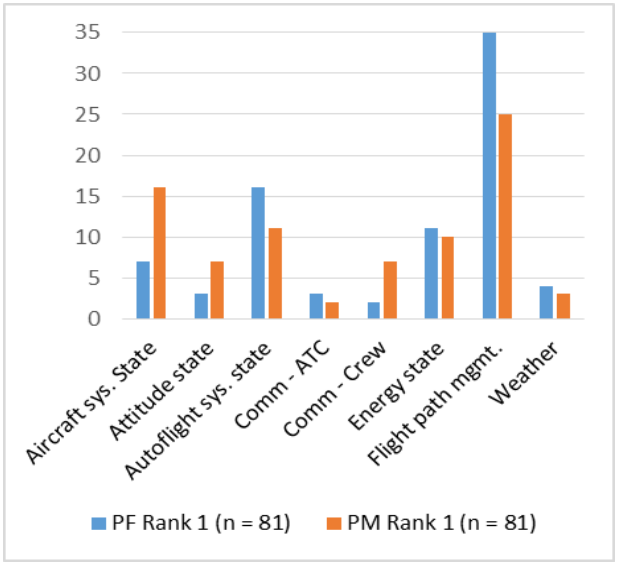


Fig. 10. Number one ranked decision-making factor for safety of flight across all BL flights.

C. *State Awareness*

Pilots were asked to indicate which baseline indicators they relied on most for awareness of (a) energy state, (b) autoflight system state, and (c) aircraft system state.

Energy state awareness results are shown in Figure 11 for PF and PM combined. PFD airspeed and altitude were the predominant responses, as expected. Lowest ranked were EICAS indications, HUD AOA, and audible alerts. Perhaps the most striking finding here is the number of indicators used by pilots for energy state awareness. The average number of indicators checked off by pilots for this question was 4.5. Interestingly, aircraft sound and feel was cited more often than some of the other indicators, including audible alerts.

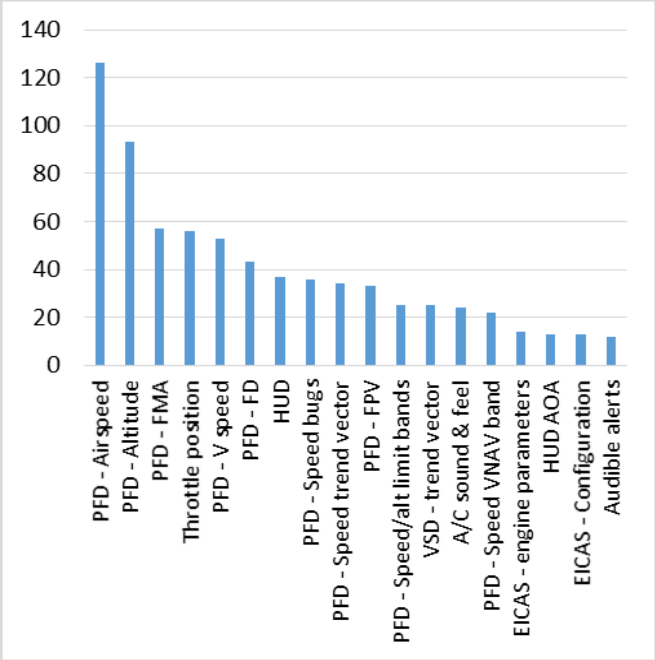


Fig. 11. Ranking of BL energy state awareness indicators.

For autoflight system state awareness, the top three baseline indicators were, as expected: (1) FMA, (2) MCP, and (3) PFD commanded speed. For systems state awareness the predominant baseline indicator was EICAS.

All of this data hints at the dichonomy that exists relative to the amount of opaqueness in the design of flight deck systems and functions. Energy state is inferred by the pilot from a host of information elements; while autoflight system state and the state of onboard systems is provided more explicitly but with little or no transparency regarding the information elements used to generate the state indication.

D. Workload

Perceived workload was measured two ways: the NASA Raw Task Load Index (RTLX) [6] [7] [8] and a four-factor assessment.

RTLX derives an overall workload score based on the unweighted average of ratings of six subscales (mental demand, physical demand, temporal demand, performance, effort, and frustration level). Each subscale rating, except performance, is scored on a scale of 0 (very low) to 100 (very high). Performance is scored on a scale of 0 (perfect) to 100 (failure).

Figure 12 shows the mean differences in perceived workload, as measured by RTLX, for the evaluated technologies versus baseline based on same-scenario pairwise comparisons. Positive values indicate higher perceived workload for the new technology vs. baseline. Negative values indicate lower perceived workload using the new technology.

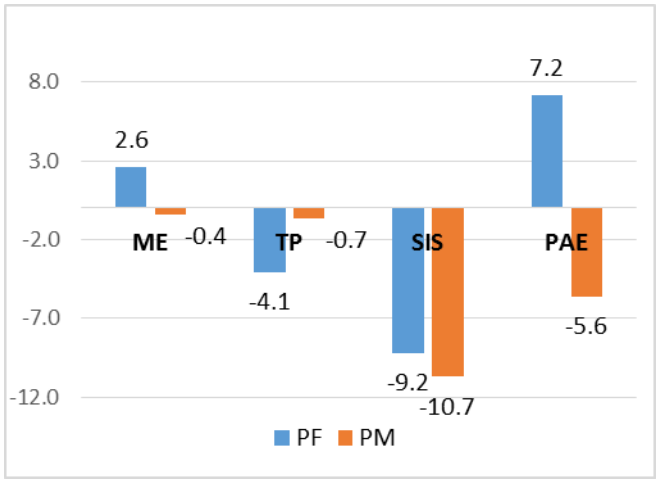


Fig. 12. Difference in workload vs baseline (RTLX).

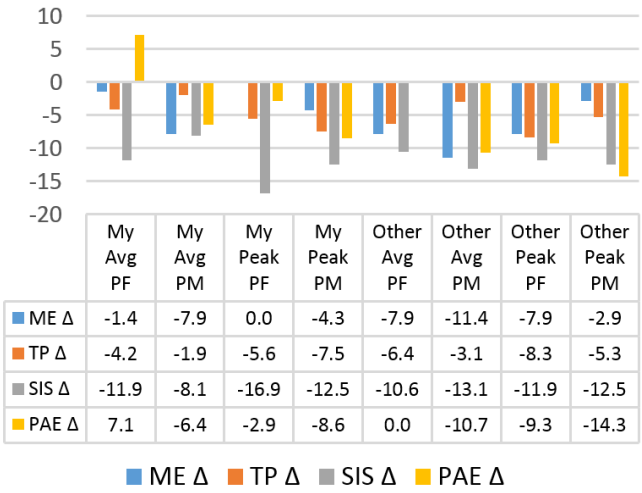
Use of ME resulted in a modest increase in workload for PF and a slight decrease for PM. For TP, workload decreased modestly for PF and very slightly for PM. For SIS, workload decreased markedly for PF and PM. For PAE, workload increased for PF and decreased for PM. These findings are quite consistent with the intended functions of these technologies. For example, ME was designed to primarily aid the PF.

A substantial workload decrease was seen for SIS PM. This makes sense because the PM sees the greatest immediate benefit of SIS, with shortened checklists to work through. A notable workload increase was seen for PAE by the PF. PF scored PAE workload higher on each of the six RTLX subscales compared with ME, TP, and SIS. This may reflect the low TRL of this nascent technology, and the fact that crew procedures for the use of PAE were not well-defined. Feedback from the crews will help to define appropriate procedures for future tests as this technology matures.

Workload was also assessed based on pilot ratings of the following four factors from 0 (Easier) to 10 (Harder):

1. My average workload.
2. My peak workload.
3. Other crewmember's average workload.
4. Other crewmember's peak workload.

The mean difference in workload for each technology versus baseline is shown in Figure 13, again based on same-scenario pairwise comparisons. Positive scores indicate pilots judged workload for the new technology to be higher than baseline. Negative scores indicates lower workload for the new technology.



	My Avg PF	My Avg PM	My Peak PF	My Peak PM	Other Avg PF	Other Avg PM	Other Peak PF	Other Peak PM
ME Δ	-1.4	-7.9	0.0	-4.3	-7.9	-11.4	-7.9	-2.9
TP Δ	-4.2	-1.9	-5.6	-7.5	-6.4	-3.1	-8.3	-5.3
SIS Δ	-11.9	-8.1	-16.9	-12.5	-10.6	-13.1	-11.9	-12.5
PAE Δ	7.1	-6.4	-2.9	-8.6	0.0	-10.7	-9.3	-14.3

Fig. 13. Difference in workload vs baseline (Four-Factor).

Figure 13 shows pilots generally rated their workload and their crewmember's workload lower with the new technologies. The exception is PF assessment of increased workload with PAE. This is consistent with the RTLX results. Notable decreases in workload are seen across the scale for SIS. In general the assessment of the other pilot's workload slightly overestimated the actual decrease in workload. Also, these results show slightly lower assessments of "my workload" than the RTLX results, but the assessments of the other crewmember's workload substantially overestimated decreases in workload shown by the RTLX results.

E. Situation Awareness

Situation awareness (SA) was measured by using the SA Rating Technique (SART) [9]. The three dimensions of SART are (1) Demand on pilot attentional resources, (2) Supply of pilot attentional resources, and (3) Understanding by the pilot of the situation. Pilots rated Demand, Supply, and Understanding after each flight on a scale of 0 (very low) to 100 (very high). SA is calculated using SART as:

SA = Understanding – (Demand – Supply)

Figure 14 shows the mean difference in SART scores for the evaluated technologies versus baseline based on same-scenario pairwise comparisons. Positive values indicate higher SA for the evaluated technology than for baseline. Negative values indicate lower SA than baseline. For PF, SA increased using ME, TP, and SIS versus BL but decreased using PAE. This reinforces the PF perceived workload increase for PAE seen earlier. The PM saw increases in SA with TP, SIS, and PAE, while the change in SA for ME was essentially flat. Overall, SIS offered substantial increase in SA. This corresponds with the substantial decrease in workload seen for SIS earlier.

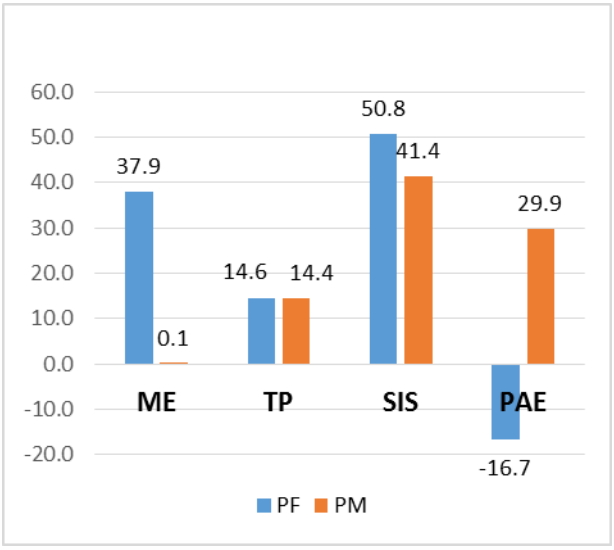


Fig. 14. Change in SA vs Baseline.

F. Usability

ISO defines the usability of a system as the extent to which it can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction, in a specified context of use [10]. The System Usability Scale (SUS) questionnaire [11] [12] [13] [14] was used to gauge how pilots assessed the perceived usability of ME, TP, SIS, and PAE in the context of the scenario just flown. SUS was chosen because it offered several benefits. (1) It could be administered quickly between flights. (2) It has been shown to be valid and reliable. (3) It can provide a good assessment of usability even with small sample sizes. (4) Since it was introduced in 1986 it has become the most widely used questionnaire for measuring perceived usability, with a track record of thousands of assessments in many fields.

Per the SUS method, after each flight the PF and PM scored 10 statements for the technology under evaluation:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

The 10 statements are arranged such that positive and negative statements alternate to encourage the respondent to think before responding. Each statement had five Likert scale response options ranging from Strongly Agree to Strongly Disagree. During post-hoc analysis, SUS scores were calculated, producing overall scores for each technology. SUS scores can range from 0 to 100, but they are not percentile ranks. However, they can be converted to percentile ranks. As a way to intuitively interpret SUS scores, [15] proposed a correlation between SUS scores and letter grades. Based on extensive research [16] showed specific letter grades can be associated with corresponding SUS score ranges. This mapping is used here to put the SUS scores in the context of letter grades.

Pilot usability scores for ME, TP, SIS, and PAE are shown in Figure 15. ME scored an A for usability for both PF and PM. TP also rated well, at A- for both PF and PM. For the PM, SIS scored an outstanding rating at A+. SIS also rated high for the PF, at A-. PAE rated C for both PF and PM. Results are shown as mean values, along with +/-2 SE to represent the variability in responses.

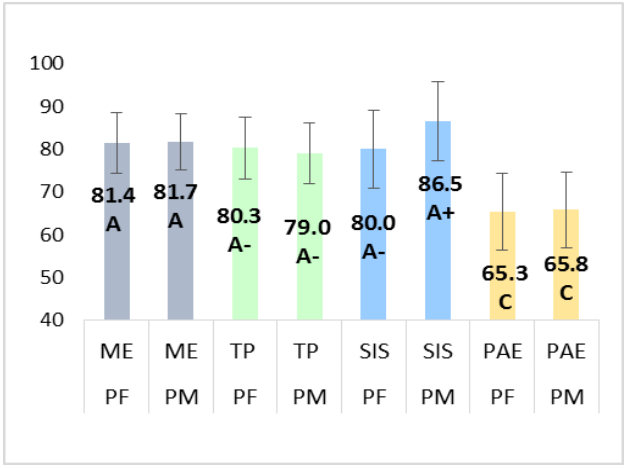


Fig. 15. Usability of ME, TP, SIS, and PAE.

Within these overall scores, mean scores for the individual 10 statements are shown in Figure 15.

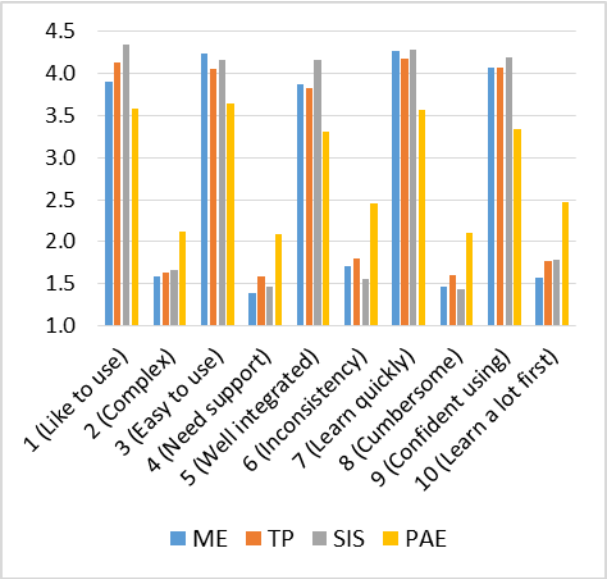


Fig. 15. SUS statement scores for the new technologies.

The scores for ME, TP, and SIS are high for the odd numbered items and low for the even numbered items, indicating good usability – commensurate with the overall scores. The scores for PAE are not as high for the odd items and not as low for the even items, indicating where more work is needed to mature the technology.

G. *Acceptability*

Pilots rated the acceptability of the technologies on a 1 to 7 scale. Where

1 = Very unacceptable. I did not like the technology and would not use it in normal operations.

4 = Average. I liked the technology and would use it in normal operations, but would like to see some improvements.

7 = Very acceptable. I like the technology very much and would use it without any improvements.

Mean acceptability results are shown in Figure 16, along with +/-2 SE to represent the variability in responses. All acceptability ratings were found to be above average. The highest overall rating is for SIS PM, which also had very good results for workload and usability. The acceptability rating for PAE is consistent with its usability results and its early stage of development; its high variability in responses is largely due to performance problems for some of the predictors for some of the flights.

The number of flights used to derive the usability and acceptability ratings are shown in Table III.

Table III. NUMBER OF FLIGHTS FOR USABILITY AND ACCEPTABILITY RATINGS.

ME		TP		SIS		PAE	
PF	PM	PF	PM	PF	PM	PF	PM
27	27	44	44	21	20	29	29

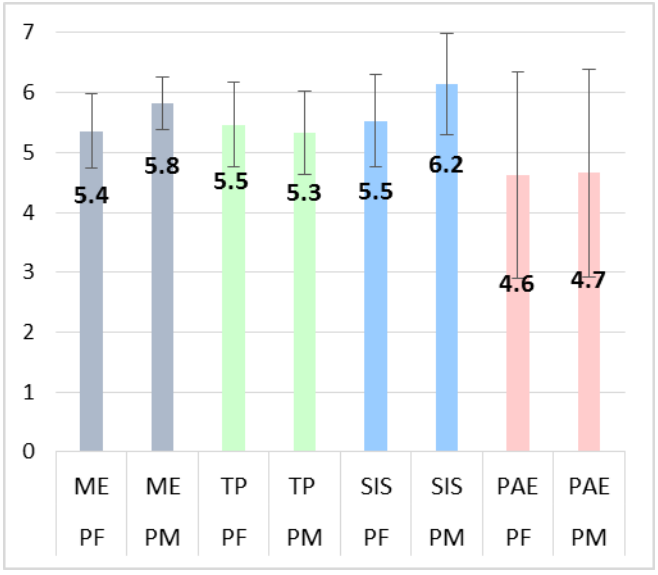


Fig. 16. Acceptability of ME, TP, SIS, and PAE.

H. *Pilot comments*

Discussions between pilots, and between pilots and researchers, were held after each run and during the post-experiment debriefs. Representative observations are paraphrased below for each technology.

ME

1. … tends to clutter the PFD and block other information in some situations.
2. … if the flight director fails, it could help decide what pitch to go to.
3. … speed bands are more useful than the roll bands.
4. …the bank angle indications would be much more effective as a pop-up feature, when the aircraft comes close to exceeding a safe bank angle; would be especially helpful during an engine-out scenario.
5. … would be nice to have when banking at higher altitudes.
6. … a nice situational awareness display that could really save the day on a heavyweight takeoff with sharp turn during cleanup.

TP

1. Several general positive comments (e.g. Awesome; I like this thing; Great aid to situational awareness; Would be used on a regular basis.)
2. … be careful not to create clutter.
3. … need to better distinguish, and train, differences between the green line predictor and the white line predictor.
4. … might be better for less experienced pilots who may not have seen some of the possible mode changes that can occur.

5. … was good in showing where aircraft will intercept the localizer and glide slope.
6. … provided another place to look besides Flight Mode Annunciator for automation states and state changes.

SIS

1. … the synoptics make it so much better, gives the big picture of the situation at a glance.
2. … like that it pops up and provides the needed information. Tells what systems we have and don't have.
3. … increases SA significantly, and avoids PM becoming buried in long checklists.
4. … need to be careful not to rely solely on this function; still need to crosscheck against other information.
5. … expedites working through otherwise lengthy checklists. Allows both pilots to stay in the flight.
6. … backed up the EICAS message for 100% understanding of the malfunction. EICAS only gives a title to the problem. This was like opening a book and immediately being given the rest of the story to understand what exactly the effects were.

PAE

1. …borders on too much information.
2. …too many failures or erroneous indications in this version to judge utility at the present time.
3. … phraseology is awkward; "ENERGY" term isn't intuitive.
4. … indications too small and should come as an aural as well. Hard to notice.
5. …Good system concept but needs more development to improve performance and gauge true usability.

SUMMARY

This paper presents results from questionnaires administered to pilots during the Automation and Information Management Experiment (AIME). Four new flight deck display concepts were evaluated by 11 crews (22 pilots) with respect to complexity, decision making, workload, situation awareness, acceptability, and usability. Each was evaluated vis-à-vis current state-of-the-art systems and indicators. Flight scenarios spanned a range of conditions designed to help expose state awareness issues where these technologies were intended to be used/useful. In general, SIS showed notable reduction in workload and increase in situation awareness. ME, TP, and SIS all ranked high for usability and acceptability. PAE ranked average for these despite its development state being very preliminary.

ACKNOWLEDGMENTS

The authors would like to thank Robert Myer, David Jenkins, and Michelle Johnson for their contributions to the development of the tablet-based questionnaire tool.

REFERENCES

[1] S. Young, T. Daniels, E. Evans, K. Shish, S. Schuet, T. Etherington, M. Uijt de Haag, and D. Kiggins, "Evaluating Technologies for Improved Airplane State Awareness and Prediction", AIAA Science and Technology Forum and Exposition (SciTech), San Diego, CA, January 4-8, 2016.

[2] **S.** Young, T. Daniels, E. Evans, E. Dill, M. Uijt de Haag, T. Etherington, "Flight Simulation Study of Airplane State Awareness and Prediction Technologies," Proceedings of the 35th AIAA/IEEE Digital Avionics Systems Conference, Sacramento, CA, Sept 2016.

[3] M. Uijt de Haag et. al.; "Energy State Prediction Methods for Airplane State Awareness," Proceedings of the AIAA/IEEE Digital Avionics Systems Conference, Sacramento, CA, September 25-29, 2016.

[4] T. Lombaerts, S. Schuet, D. Acosta, J. Kaneshige, and K. Shish, "Piloted Simulator Evaluation of Maneuvering Envelope Information for Flight Crew Awareness," paper no. AIAA 2015-1546, AIAA SciTech 2015, Kissimmee, FL, Jan 2015.

[5] K. Shish, J. Kaneshige, D. Acosta, S. Schuet, T. Lombaerts, L. Martin, and A. Madavan, "Trajectory Prediction and Alerting for Aircraft Mode and Energy State Awareness," paper no. AIAA 2015-1113, AIAA SciTech 2015, Kissimmee, FL, Jan 2015.

[6] S. G. Hart, L. E. Staveland, "Development of a NASA TLX (Task Load Index): Results of empirical and theoretical research," P. S. Hancock and N. Meshkati (Eds.), Human Mental Workload, Amsterdam: North Holland Press, pp. 139-183, 1988.

[7] S. G. Hart, "NASA-Task Load Index 20 Years Later," Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting, pp. 904 – 908, Santa Monica: HFES, 2006.

[8] W. F. Moroney, D. W. Biers, F. T. Eggemeier, J. A. Mitchell, "A comparison of two scoring procedures with the NASA Task Load Index in a simulated flight task," Proceedings of the IEEE National Aerospace and Electronics Conference 2, Dayton, OH, pp. 734-740, 1992.

[9] S. J. Selcon and R. M. Taylor; "Evaluation of the situational awareness rating technique (SART) as a tool for aircrew systems design," AGARD-CP-478, Situational Awareness in Aerospace Operations (pp. 5-1 to 5-8). Neuilly Sur Seine, France: Advisory Group Aerospace Research & Development (AD-A223939), April 1990.

[10] ISO 9241-210, Ergonomics of human-system interaction — Human-centered design for interactive systems, 2010.

[11] J. Brooke, "SUS – A quick and dirty usability scale," in P. W. Jordan, B. Thomas, B. A. Weerdmeester, A. L. McClelland, Usability Evaluation in Industry, London: Taylor and Francis, 1996.

[12] J. Brooke, "SUS: A Retrospective," Journal of Usability Studies, Vol. 8, Issue 2, pp. 29 – 40, Feb 2013.

[13] "System Usability Scale (SUS)," https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html.

[14] J. Sauro, "Measuring Usability with the System Usability Scale," http://www.measuringu.com/sus.php, Feb 2, 2011.

[15] A. Bangor, P. Kortum, J. Miller, "Determining What SUS Scores Mean: Adding an Adjective Rating Scale," Journal of Usability Studies, Vol. 4, Issue 3, pp. 114 – 123, May 2009.

[16] J. Sauro, A Practical Guide to the System Usability Scale, Denver: Measuring Usability, 2011.