SURVEY DEVELOPMENT AND INITIAL DATA: FLIGHT CONTEXT AND PILOT TECHNIQUES IN EVERYDAY FLIGHTS

Dorrit Billman
NASA Ames Research Center
Moffett Field, CA
Alan Hobbs, Lucas Cusano, & Nóra Szládovics
San Jose State University Research Center
Moffett Field, CA

The aviation industry is recognizing that flight crews routinely contribute to system safety in ways that go beyond adherence to standard operating procedures (SOPs). Our research goals were to explore a) whether a survey could shed light on pilots' contributions to adaptation and resilience in everyday flights and b) relevant assessment methods. The survey focused on challenges faced by pilots in normal operations, and on the ways that pilots anticipate and monitor those challenges. We collected responses concerning revenue flights from two pilot groups; one group also provided responses concerning a simulated scenario. The results indicated that relatively few flights proceeded exactly as in the original flight plan. Pilots routinely anticipated and adapted to changing circumstances. We discuss some design and assessment challenges encountered for a survey on this topic, we provide 5 approaches to assessment, and we present example findings as illustrations. We hope assessment methods such as these will lead to useful surveys of resilience in flight.

Much of our knowledge about human performance in flight safety has come from the analysis of undesired events, whether accidents, incidents, or crew behaviors identified via flight exceedance monitoring or observational techniques. Recent years have seen an acknowledgement that operational personnel are not merely sources of "human error", but also make a unique human contribution to safe outcomes. In a few celebrated cases, this takes the form of "heroic saves", but on many more occasions, operational personnel contribute to safety through everyday, barely-noticed, actions that turn potentially hazardous situations into non-events.

An emerging approach to safety, frequently referred to as "Safety II," proposes that the positive human contribution is an important, largely untapped source of safety information. Some airlines have successfully trained observers to identify and record the positive behaviors exhibited by the crew over the course of a flight. In other cases, flight crew are interviewed about good practices or positive behaviors. However, each of these methods are relatively limited in scale and resource intensive. A survey could provide a relatively low-cost approach to systematically gather this information on a larger scale.

Our research focus is methodological, investigating prospects and challenges for surveying "Safety II" activities. Throughout this paper we include empirical findings from our trial surveys to illustrate both our approach to survey development and assessment, and the potential benefits of a survey focused on the positive human contributions. We hope that such a survey could be both a research tool as well as a safety management aid to the aviation industry.

Survey Development and Response Collection

This paper describes the iterative development and assessment of a survey to examine the human contribution to resilience in routine airline operations. Each survey version was critiqued by airline pilot advisors and completed by a sample of airline pilots. Several design considerations shaped the scope and prioritized the coverage of the survey:

- Our survey was directed at adaptive behaviors that are not specified in (SOP) or standard practices.

- Resilient behavior has been described as monitoring, anticipation, responding, and learning (Hollnagel, 2015). Our survey focused on the more proactive over reactive aspects, in part because this is less studied than reactions to triggering events.
- We limited the initial scope of the survey to the descent phase of flight as we anticipated that this would provide us with numerous opportunities for resilient pilot behavior. For example, Standard Terminal Arrivals (STARs) can require complex interactions with the autoflight system, well-timed actions, and an understanding of automation, ATC, and the airspace.
- To understand the intent of pilot behavior, it is necessary to understand the operational context in which the behavior occurred. Therefore, we included some situational questions, primarily about ATC actions and weather.

Several principles guided the organization and design of questions, to make them as clear and easy to answer as feasible:

- We aimed to avoid abstract terminology or jargon that might be used in the research community but not necessarily familiar to pilots. For example, a major airline (AA LIT White Paper 2020) uses specially trained personnel who observe flights from the jump seat and record instances of resilient performance using a standard set of terms. Pilots lacking specialized training might vary widely in how they interpreted such terms.
- Our focus was on adaptive activities in ordinary circumstances that were unlikely to be particularly striking or memorable. Therefore, to minimize interference, we focused on the most recent flight.
- Unless phrased carefully, questions about resilient behavior can imply a "correct" or desirable answer. For example, a survey question asking if a potential threat was included in a briefing could imply that the threat should have been included. We framed the majority of questions to be "matter of fact" descriptions about the flight and what the crew did.

We used a variety of question formats, including checkbox items, rating scales, and free text responses. A checkbox *item* consists of a question and *response* choices, allowing multiple choices. Throughout the survey development process, a variety of airline pilots with research backgrounds helped us to ensure that questions were relevant and phrased appropriately.

Survey development was guided by these considerations of content, question design, and question format. We iterated through four major cycles of development and response collection. The versions in the last two cycles were similar. Respondents were airline pilots completing the survey for their most recent line flight. Twenty-five of these pilots were participants in a flight simulation study conducted at NASA Langley Research Center as part of NASA's SOTERIA[1] study (Stephens et al., 2021). An additional 65 respondents were Line Check Pilots (LCP) in the airline industry who were not participants in the SOTERIA study. SOTERIA pilots (n=22) also completed the survey regarding a simulator scenario with video recording. The pilots who completed the survey versions should by no means be considered a random or a representative sample of airline pilots. Table 1 shows an overview of topics and formats of survey questions in the most recent iteration.

Table 1. Survey question content and format for the 4th iteration (LCP).

Format	What happened	What did you do						Evaluate	
	Op. Context	Proactive/Anticipatory				(Re?/active)	Explc. "Learn"		
		briefing	info gathering	assessment	"monitoring"				
Checkbox	6	3	2	3	2	5			21
Rating		1:eval	2:eval		1 + 6:eval			2 +9	12
Text	1	2			1		2	4	10
	7	5 (+1)	2 (+2)	3	4 (+6)	5	2	14	43

[1] System Wide Safety Operations and Technologies for Enabling Resilient In-Time Assurance (SOTERIA)

Results and Assessment

Keeping our methodological focus, we describe five *approaches* to assessing the value of this survey, considering validity and reliability. We use selected results as illustrations.

Approach 1: questions reviewed individually for interesting but reasonable findings. For an individual question there were few or no cases where responses seemed inconsistent with how the world is, though many provided novel information. We consider examples from text and checkbox items. Text responses were coded into categories based both on our expectations and what was observed.
Example 1A: SOTERIA-revenue pilots described what was most challenging and in the next question how they managed it. Responses from the 25 participants were coded into 1 or more sub-categories, grouped into more general categories. Figure 1 shows the dominant challenge was Operations, specifically, Scheduling/Delays/Timing Out, with Fatigue a close second. It is striking that CRM was identified as a management method in almost 3/4's of the reports, with the proactive strategy of extending a briefing beyond the usual the most prevalent CRM method in almost 1/4 of reports.

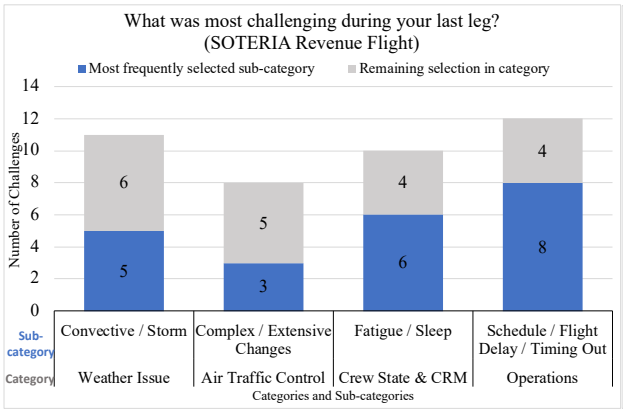
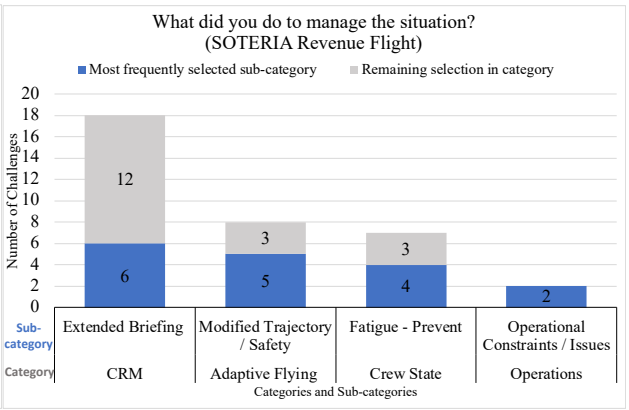


Figure 1A: Most challenging aspects.



1B: Method for management.

Example 1B: If a pilot said they had learned something that might help on a future flight (32 of 65 did), they described what that was. Their responses were classified into one of 9 categories (see Figure 2). Choices were diverse, but the most common (1/5 of the group) addressed communication in the cockpit, again highlighting the prominence of CRM in pilot experience.

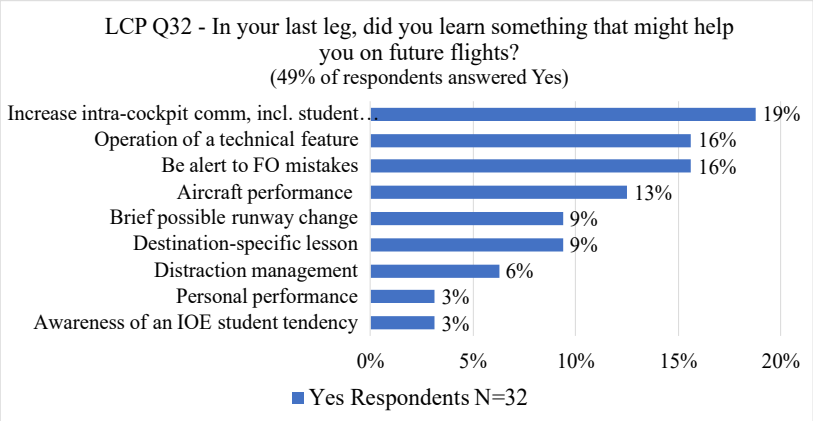


Figure 2. Categories of what pilots learned.(LCP data)

Example 1C: Several checkbox questions asked about what ATC did, the weather, and other aspects of the operational environment. As Figure 3A shows, Q16 asks about ways ATC might modify an arrival, plus a "none" and other option (as on all LCP checkboxes). Strikingly, only 13.9% of arrivals were not modified by ATC. Thus, it is a small minority of arrivals where STARs are flown as programmed (and the large majority where pilot response is required). Also of interest is the rather high proportion of runway changes.

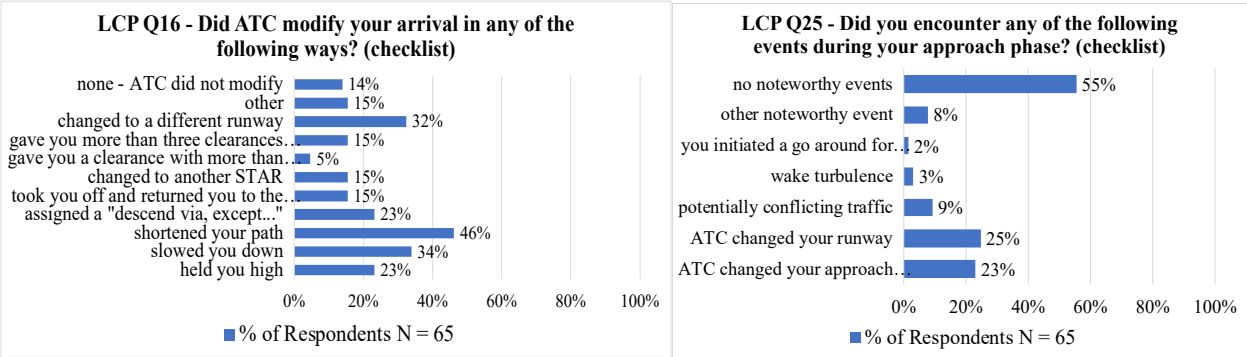


Figure 3A: Arrival Phase - ATC modifications (LCP). 3B: Approach Phase - events encountered (LCP).

Approach 2: consistency across related questions. The relations between responses to different questions may be associated in expected or in surprising ways; a surprise may challenge assumptions about the world or about the basis for answering the question. We give examples of findings consistent with an expected pattern, of surprising patterns, and of absence of clear relations where we thought they might occur.

Example 2A: we thought flights judged more challenging than normal might be more likely to provide something to learn. Of the LCP pilots who judged the flight more challenging than normal, 60% said they learned something (and 40% did not), while of the pilots who said it was a normal or less level of challenge, 40% said they learned something (and 60% did not). The correlation between challenge rating 5 pt scale) and pilot learning was r(65)=.69. Thus, the pattern of responses to these two items was consistent with the expected relation.

It may be hard to tell whether a surprising finding is accurate or an artifact of the question design. Consider the reports of runway changes during shown in Figure 3. If the percentage of runway changes on Arrival (Fig. 3A) and on Approach (Fig. 3B) are summed, the total is over 50% (33% +25%). Looking at individual responses shows everyone who checked the 'runway on approach' response also checked 'runway on arrival' response. Possibly there were two runway changes. Alternatively, the respondents counted the same change twice. Using *Approach 5* on the SOTERIA simulation data provides additional hints, below.

Example 2B: we asked pilots about the percentage of time spent on different activities, as shown in Table 2. Items a and b in Table 2 are not explicitly reverse coded, but we expected these two would sum to about 100%, which they do. However, it is highly likely that when working on systems (44%>> 16%) one is not also specifically attending to the progress of the flight (44 + 84>> 100). This apparent inconsistency may suggest difficulty of reporting about interleaved tasks or alternatively, a strong belief in the ability to truly multitask.

Table 2. Percents of flight time pilots judged as allocated to different activities. (LCP data)

During the descent phases, estimate the % of time in which... *[respondent clicked on a 1-100 timeline]*	Mean %
a) I "mentally flew" the aircraft, even when the autopilot, or the other pilot, was controlling it.	84
b) I was NOT specifically attending to the progress of the flight.	15
c) I was working on systems management (e.g., entering values in FMS) or communications (e.g., radio settings, talk with ATC).	44

Example 2C: We had hypothesized we might see clear associations between events (e.g., ATC clearances) and pilot actions (e.g., input to the autopilot). However, the complexity of possible relations was not easy to trace out in relations among these responses. This may suggest that the combined

operational complexity of how pilots adapt will benefit from a more structured inquiry; this might be asking whether an event occurred, such as being held high by ATC, and if so, how it was managed.

Approach 3: compare response patterns across different groups. Comparing frequency of responses across different groups provides some indicators of stability. For example, in the LCP group the proportion of flights where ATC did not modify descent was low (14%); turning to the SOTERIA revenue flights, 20% did not have an ATC modification, a similar though somewhat higher percent. Of course, differences may reflect actual differences between groups as well as less meaningful variability. Turning to the pilot monitoring (PM) versus pilot flying (PF) within the LCP group, we set a heuristic criterion of 20% difference between the two roles to consider noteworthy. None of the responses to any of the 6 items about what happened and only 4 responses in the more than 75 responses across the 21 items about pilot action differed by this criterion. These broad patterns are not particularly diagnostic but suggest that findings do not differ majorly when a flight is reported by PM or PF.

Approach 4 & 5 compare ratings of the same situations. These are feasible for SOTERIA crews in simulator events, for responses to checkbox items. In *Approach 4*, ratings of same-crew PM and PF can be compared using standard reliability measures; we explored several and settled on percent agreement. We looked at the agreement between PM and PF on whether they selected a particular response on checkbox questions. We scored whether a given crew agreed on a given response and averaged these to get a percent agreement a) across crews for a response and b) across responses for a crew. Agreement scores for individual crews ranged from 72% to 86%. Agreement scores for individual responses ranged from 36% to 100%. The overall agreement level averaged 77%.

Factors that seem to contribute to high reliability of a response include being highly standard actions or SOPs (Table 3 #1) and being highly salient, observable events (Table 3 #2). Factors contributing to low reliability include reference to standards SOP; it may be unclear what is the standard level of automation, or SOP (Table 3#3, #4), and actions which may fall close to such a boundary (Table 3 #4); a response may have low reliability both because it is hard to decide what category the question refers to, and to decide if the actual events fit in that category. Table 3 shows examples.

Table 3. Responses With High and Low Agreement (SOTERIA -sim data).

Highest Agreement	#1 What did you do to assess how your autoflight system would handle your STAR? --checked that the values in the flight management computer matched values on the chart--	91%
	#2 Did you encounter any of the following events during your arrival? --ATC changed your runway--	100%
Lowest Agreement	#3 Did you fly any part of the approach manually, or at lower levels of automation than standard for your airline? --No/Not Applicable--	36%
	#4 During descent, the PM: --provided positive confirmation of expected actions or states, beyond SOP--	36%

Approach 5: comparison to an observer. Observers are given the best feasible way to review and rate the crew's flight using video of the sim session. We hope to be able to conduct *Approach 5* assessment in the future. Nevertheless, we can gain some clues about validity without an extensive review of simulator events. ATC clearances were scripted elements of the scenarios, delivered by a member of the research team. Two of the event scenarios, seen by 6 total crews, included a single, scripted runway change. Although only one runway change occurred, 11 of the 12 pilots reported two, one during arrival, one during approach. This suggests that in these scenarios, pilots were not distinguishing when a runway

change occurred, and that the question might be better framed by asking about whether any runway change(s) occured, and then asking in what phase of flight.

Discussion & Conclusions

The primary purpose of the research was to develop and assess surveys, as a little-used method for assessing crews' activities in normal flights and the operational perturbations routinely introduced. The assessment provided both information about the flights and information about what questions might merit revision. For example, despite much iteration on this topic it was hard to ask pilots questions involving behavior that went beyond standard performance, one of the ways we tried to communicate resilience. How much difficulty comes from understanding the question intent or from assessing the behavior is hard to determine. We were extremely fortunate to have data from two groups of pilots, and from simulated as well as revenue and flights, including pilot pairs crewing the same simulated flight. This gave us the opportunity to use the data to assess the survey using several *approaches*. *Approaches 1-3* depend on making sense of how responses fit in with, yet extend, what we know, broadly, its validity. This can be done by looking at individual items and responses, by looking for patterns of coherence between items, and by looking for consistency or meaningful differences between groups replying to the survey. *Approaches 4* and *5* measure agreement between pilots in the same crew or compare crew responses to observers equipped to make a best estimate of 'ground truth.' This agreement measure would be a further measure of validity. We are not aware of this style or degree of assessment of surveys in the aviation domain.

As the presented examples suggest, responses also provided sensible and interesting information about prevalence of situations or behaviors, for example, the pervasiveness of ATC changes during descent and the association between how challenging the flight was and learning something new. Future reports will provide more comprehensive coverage of findings. We also plan to summarize suggestions about survey design relevant to understanding resilience, pilot activity, and its context. We hope that survey assessment will result in useful surveys for measuring pilot contributions to resilience.

Acknowledgements

Our thanks to the airline pilots who gave their time to complete the surveys. Thanks also to the several pilots who provided valuable input to survey development, particularly Capt Barth Baron, Capt Dan Kiggins, and Capt Rob Kotesky. Thanks to Immanuel Barshi for critical networking, to Jon Krosnick for discussion, and to Jon Holbrook, Lawrence Prinzel & Chad Stephens for support of the SOTERIA project. This research was funded by NASA's System Wide Safety Project.

References

American Airlines' Department of Flight Safety (AA DFS). (2020). Trailblazers into Safety-II: American Airlines' learning and improvement team, a white paper outlining AA's beginnings of a Safety-II journey. Retrieved from https://www.skybrary.aero/sites/default/files/bookshelf/5964.pdf

Hollnagel, E. (2015). RAG—Resilience Analysis Grid. Retrieved from http://erikhollnagel.com/onewebmedia/RAG%20Outline%20V2.pdf

Stephens, C., Prinzel, L., Kiggins, D. Ballard, K., & Holbrook, J. (2021). Evaluating the Use of High-Fidelity Simulator Research Methods to Study Airline Flight Crew Resilience. 21st International Symposium on Aviation Psychology, 140-145.