ORIGINAL ARTICLE – 2017
REVISION – 2026 (June)

TRAINING SENSEMAKING FOR THE MODERN FLIGHT DECK: A NEW PERSPECTIVE

Frederik Mohrmann[1]

[1] Delft University of Technology, Delft, the Netherlands

Joris Field[2]

[2] Royal Netherlands Aerospace Centre, Amsterdam, the Netherlands

ABSTRACT: This study evaluates a novel problem-solving strategy for flight crews experiencing complex, ambiguous and opaque situations in modern, highly automated (fourth generation) aircraft. This strategy has been developed and evaluated in the EU research project Man4Gen in 2015. An experiment was performed with 14 flight crews type-rated for either the Airbus A330 or A320 from various airlines, presenting flight crews an ambiguous and opaque situation in a carefully refined simulated scenario featuring a complex energy management issue combined with a diversion in poor weather. Crew performance was evaluated on 30 distinct challenges and risks that could present themselves throughout the scenario. These performance metrics were compared to crew behavioural indicators concerning the use of the novel strategy and supporting risk information display innovations. The experiment features three sets of crews: baseline crews (N=3 crews) did not receive training nor displays, one set of crews (N=7 crew) received only training in the strategy and were provided with a quick reference card, and a third set of crews (N=4 crews) received training in the strategy as well as access to the supplementary Risk Information System (RIS) designed to work in tandem with the strategy. This study evaluates both the strategy training and RIS impact on behaviours (through a between-group analysis), as well as the relationships between behaviours and safety performance, by exploring correlations in a within-group analysis of the entire population (N=14). This allowed for a more robust experiment setup, with less sensitivity to random effects in small group sizes. Results on training effectiveness indicate that the training alone did not induce the desired behaviours as hypothesized (and sometimes the opposite). The RIS was effective in increasing behaviours related to contingency management. The second analysis confirmed that performance increase correlates with several hypothesized behaviours, but results were not powerful enough above post False Discovery Rate *p*-adjustments. Behaviours related to increased performance included efficient short-term time management before problem solving and increased time spent managing uncertainties and contingencies in a sensemaking cycle. Although the experiment featured several limitations and can be improved upon in terms of training design and possibly deploy a longitudinal study design, it presents a concrete first attempt to both operationalize and visualize resilience theory by engaging human sensemaking abilities in opaque and ambiguous situations.

KEYWORDS: Resilience, Sensemaking, Aviation, Ambiguity, Complexity, Behaviour, Training

1 INTRODUCTION

The past two decades have shown a shift in aviation safety and accident causation. While the safety performance of the aviation transport system is unprecedented, accidents are increasingly characterized by difficulty of flight crews asserting manual control, maintaining situation awareness and demonstrating effective decision making (Sarter & Woods; 1994; Saurin & Carim, 2012; IATA, 2014; Man4Gen Consortium, 2015; Mohrmann et al., 2015; Stoop & Van Kleef, 2015; Wolter et al., 2015; Strauch, 2017; Kharoufah et al., 2018; Proctor & Van Zandt, 2018; Kelly & Efthymiou, 2019; Prinzel et al., 2024; Gago et al., 2025). This is in part attributable to the increased reliability of systems which erode crew familiarity with system intricacies and malfunctions, in addition to an increased focus on (procedural) compliance in crew training and operations. While the emphasis on procedural compliance is understandable, it may also be ossifying crew competencies to act effectively in non-standard situations (Landry, 2009; Mohrmann et al., 2015; Rankin et al., 2016; Abbot, 2017; Niedermeier et al., 2018; Boy, 2020; Clark & Wilson, 2024), expanding the chasm of control between crew and systems.

To better understand the nature of this shift in accident causation and how it can be mitigated, the EU Man4Gen project ("Manual Operations of Fourth Generation Airliners", 2012-2016) investigated human-machine interactions in modern, highly automated fourth generation aircraft (Man4Gen Consortium, 2015). The need for such an exploration is also reiterated by DeSalvo & Fogarty (2016) who underscore, from a system verification and validation point of view, the necessity for "aircraft-level operation" validation of systems, instead of only single system-level and functional-level validation. Such limited scopes of validation will too often lead to a myopic set of assumptions of system functionality, when the "cockpit operational ecosystem" may suffer from clutter, poor integration and (most notably) ineffective human-machine interactions. Furthermore, DeSalvo & Fogarty (2016) indicate their concern with the assumptions about the real-world accuracy of system simulation models, as well as the safety assessment of changes in highly integrated, complex operational systems. In the context of such complex systems, verification and validation methods must shift from isolated system validation, to integrated and operationally representative validation of the collective system, including human actors. It is only through such holistic approaches that sufficient insight can be provided into cascading effects and dynamic relations within increasingly coupled system architectures. Such a perspective on system validation also meshes well

with the Joint Cognitive System (JCS) perspective (Hollnagel, 2005). The necessity for such a holistic approach was further underlined in the US Senate's Committee Investigation Report concerning the certification efforts of the Boeing 737MAX (Senate Committee, 2020), in which long-held human factors assumptions were left unchallenged, despite growing system complexity in primary flight control systems.

This study seeks to validate whether a novel problem-solving strategy designed toward managing complex and ambiguous situations can induce effective crew behaviours and improve flight safety in such opaque and ambiguous situations. The strategy and associated display concepts were developed, implemented and tested by the Netherlands Aerospace Centre (NLR) and German Aerospace Centre (DLR) in the Man4Gen project. Validating this strategy, given the complexity, human variability and scenario flexibility, calls for a novel, holistic and data-driven approach to compare flight crews in behaviour and flight safety performance indicators. This study will therefore evaluate the effectiveness of the proposed resolution strategy as well as the application of new behavioural and performance analysis methods suited for complex and unpredictable research contexts.

Section 2 will provide research background of the Man4Gen project, introducing both the proposed strategy and risk display support solutions and validation approach. Section 3 describes the research method, evaluation context (scenario) and behavioural and performance measures. Sections 4 and 5 present study results and conclusions, and Section 6 discusses the larger operational and research implications for human performance in complex and ambiguous systems and operations.

2 MAN4GEN PROJECT BACKGROUND

2.1 Summary of the Man4Gen strategy design

The EU research project Man4Gen explored the difficulties that flight crews experience in asserting manual control, both physically and cognitively, in modern fourth generation aircraft that feature high levels of automation. Based on the initial Man4Gen exploratory simulation experiments studying these control challenges (Mohrmann et al., 2015; Rankin et al., 2016; Niedermaier et al., 2018), the NLR and DLR teams in the Man4Gen project articulated training, operational and cockpit display innovations to support flight crew recovery of situation awareness and cognitive control (i.e., understanding the state of the

aircraft and its systems, recovery options and mitigation options) in the face of opaque and ambiguous situations that modern, complex aircraft may provide (Field et al., 2017; Buch et al., 2017). These innovations distinguished themselves from existing problem-solving guidance and methods in three distinct areas.

The first is a differentiation from common industry standard approaches such as T-DODAR, FORDEC and DESIDE[1] (Banks et al., 2020) which emphasize "identify the failure" or "diagnose" and proceed linearly to flight continuation planning. While such a linear process suffices when failures are singular, distinct and recognizable, complex failures in modern airlines do not always present themselves in such a transparent way and may feature significant opacity and ambiguity. Rather, a *cyclic* learning process which alternates between sensemaking and execution was observed as a more effective approach to understand and resolve complex, multilateral and/or diffuse non-normal situations.

Second, existing problem-solving strategies are execution-focused and feature little guidance toward stabilizing the situation or buying time (which the exception of *Time*-DODAR). In the exploratory studies, poor-performing crews were observed to experience high temporal pressure and often a strong drive to land the aircraft as soon as possible. While not a bad objective in and of itself, when returning to land *becomes* the short-term plan, additional complexity and risk is introduced when requisite sensemaking occurs simultaneously with high workload flight phases that also feature system reconfigurations (i.e., approach and landing). As such, effectively managing time and "taking a moment" for sensemaking was shown to positively contribute to a crew's ability to manage these complex situations. The necessity for such "self-recovery" is echoed in findings from more recent studies in startle and surprise management techniques (EASA, 2018; Landman et al., 2020).

Third, the interaction between crew members and systems within the JCS may also feature innovations that can make (or break) such sensemaking for opacity. Existing failure management systems such as the Airbus Electronic Centralized Aircraft Monitor (ECAM) and the Boeing Engine Indicating and Crew Alerting System (EICAS) are useful interfaces to present a system's current state but are poor

[1] T-DODAR: Time, Diagnose, Options, Decide, Assign, and Review
FORDEC: Facts, Options, Risks and benefits, Decide, Execute, Check
DESIDE: Detect, Estimate, Set safety objectives, Identify, Do, Evaluate

substitutes for diagnostics (for which they are often used). By explicitly decoupling system state information and executive guidance in system presentations, cognitive space can be created for the flight crew to aggregate a variety of contextual and state information and engage in more effective Threat and Error Management (TEM) techniques (Helmreich et al., 1999). Such a revision in systems allows a broader transition in the cockpit JCS to support sensemaking.

The Man4Gen project developed a novel problem-solving strategy and associated Risk Information System (RIS) to evaluate whether flight crews can be trained and guided towards effective problem solving in opaque and ambiguous situations. This study focuses on the evaluation of the effectiveness of this problem-solving strategy and associated systems in the context of a fourth-generation aircraft scenario featuring opacity and ambiguity. Additional details in the design and development of this strategy may be found in a sister publication (Field et al., 2017). The underlying philosophy behind this strategy consists of three key principles, in the following order:

- Time Management (TM),
- Uncertainty Management (UM), and
- Contingency Management (CM)

These principles were operationalized in six phases, with each phase actively engaging the crew cognitively, in contrast to merely prescribing actions as traditional checklists and procedures often do. The phases are listed in Table 1. Phases 1, 2 and 3 were intended to manage time criticality; Phases 4 and 5 manage uncertainty; Phase 6 supports planning for contingencies and changes.

Table 1. Overview of strategy phases and underlying philosophy principles

Principle	**Phase No.**	**Phase**
Time Management (TM)	1	Stabilize flight path
	2	Mitigate immediate threats
	3	Short-term planning
Uncertainty Management (UM)	4	Identify situation
	5	Perform appropriate actions
Contingency Management (CM)	6	Long-term planning

Crews are expected to sequence the strategy from Phase 1 to Phase 6, and then cycle between Phases 4, 5 and 6 to engage in a continuous sensemaking process. This is in line with the recursive sensemaking concept proposed by Hollnagel (2005). As complex, ambiguous, opaque situations are often unclear when they first present, they clarify as crews repeatedly hypothesize, experiment and learn about the situation. The initial Phases of the strategy have their primary purpose to create time and space for sensemaking to occur (i.e., providing a temporary, stable situation and removing temporal stress on the crew). These initial phases are critical yet intended to be performed usually only once. Crews trained in following this strategy are provided preparation material, classroom training and a quick reference card (Appendix A) as a mental aid containing the most important considerations for each phase. More information about the strategy development and design process can be found in (Field et al., 2017).

2.2 Extending the strategy with the Man4Gen Risk Information System (RIS)

Alongside the development of the strategy, the Man4Gen project also investigated a new cockpit display concept to support pilots in regaining manual control. This concept described in detail by Buch et al. (2017) is referred to as the Risk Information System (RIS) which aims to provide flight crews more context-specific risk information, to rebuild and maintain situation awareness and cognitively engage with the situation. The RIS was developed by the DLR in collaboration with NLR and implemented in DLR's A320 Air Vehicle Simulator (AVES). The system emulates an A320 ECAM display by providing extra ECAM "risk pages" illustrated in Appendix B (adapted from Buch et al., 2017). The RIS aims to support the three principles in the following ways:

- Time management: Risk overview pages (flight phases, aircraft characteristics) show the crew whether immediate action is required (which is not often) and that there is time to engage cognitively (which is more often the case).
- Uncertainty management: Risk detail pages provide causal descriptions as well as action considerations that support crews in further diagnosing the situation and understanding the impact on the aircraft.

- Contingency management: Risk overview pages allow flight crews to plan and identify risks in alignment with industry Threat & Error Management (TEM) concepts, developing more effective options and long-term plans.

Crews trained in the strategy and display solutions were provided similar strategy training, briefing and quick reference cards as the strategy-only crews, although these resources were all adjusted to include and explain the RIS as an integral part of the strategy execution. This included a modified quick reference card indicating which RIS pages may be most helpful in certain strategy phases (Appendix A).

2.3 Man4Gen validation study

The study hypothesized that the more training and support was provided to a crew in the form of strategy (training) and the RIS, the better their performance would be in managing safety-critical events. To investigate this, the study featured three test groups:

- A baseline group (BSL) which did not receive any training or RIS;
- A strategy only group (STG) which was trained in the new problem-solving strategy; and
- A strategy with display group (DIS) which received both strategy training, as well as the RIS.

The groups were divided across two experiment locations at NLR in Amsterdam and DLR in Braunschweig. At this point it must be mentioned that the study suffered from pilot recruitment, data collection and integration complexities which reduced the effective sample size considerably (discussed in Section 6). A Kruskal-Wallis paired ANOVA test indicated no clear differentiation between groups in terms of safety performance, likely attributable to the (very) small group sample sizes. However, reviewing performance data across the entire population (N = 14) indicated that there was considerable *within-group* variation[2]. For the current study, this resulted in a segmentation of the initial research hypothesis into two underlying causal pathways which potentially would allow for a stronger study design. The two resulting hypotheses are:

1. Crews trained in the strategy—with and without the supplementary RIS— (i.e., STG and RIS groups) will feature behaviours more closely aligned with the strategy, as compared to

[2] Coefficient of quartile variation >30%

crews without training or RIS (i.e., BSL group). This will be evaluated by analysing *between-group* differences in behavioural data.

2. Crews exhibiting behaviour more closely aligned to the strategy will feature higher performance scores, and vice versa. As all crews (regardless of their grouping) provide both behaviour and performance data, this can be investigated through correlation analysis considering all crews *as a single population*.

Both analyses will be done at two levels of abstraction: the strategy phase level, and the strategy principles level. By investigating the strategy's effectiveness at both phase- and principle-levels, this experiment may be able to discriminate whether the strategy's phases provide useful guidance to flight crew or—in the event they are not clearly effective—whether the underlying principles remain valid.

3 METHOD

3.1 Setup and scenario

The experiment featured 14 crews in total[3], each crew consisted of two pilots type-rated for either the Airbus A330 (NLR setup) or A320 (DLR setup). Three crews were not provided any Man4Gen training (BSL group), seven crews were trained in the Man4Gen strategy only (STG group), and four crews were trained in both the strategy as well as being provided the RIS (DIS group). As a general remark, the small number of crews warrants careful exploration of the study's results, both statistically and qualitatively.

All crews were provided a simulated flight scenario either at NLR or DLR, which was designed to provide similar challenges as two other scenarios explored in previous Man4Gen studies (Rankin et al., 2016). The scenario provided crews with a complex, ambiguous situation which challenged their ability make sense of the situation and regain cognitive control. This paper will not detail the method used to design this scenario; it only provides a synopsis of the scenario. Readers are referred to (Niedermaier et al., 2018) for a detailed account of the scenario design method. Below is a synopsis of the scenario, a visual schematic of the scenario is provided in Appendix C.

[3] Originally there were 17 crews, but three crews (all in the BSL group) did not feature complete behavioural datasets to be included in this evaluation. This reduced the BSL group from six crews to three crews.

"The scenario entails an Airbus A320 or A330 (depending on the simulator), and begins with a take-off, climb out and departure with bad weather approaching in the vicinity, with the redundant FCDC2 and SFCC2 flight control computers faulted but reported as (legal) MMEL conditions. These flight control computer faults have no effect on the scenario, other than to potentially misguide crews at the start of the scenario. During climb-out the Engine #2 (ENG2) oil temperature begins to rise. Upon reaching advisory temperature, the crew may elect to reduce thrust or (if referred to the FCOM) disconnect the GEN2 or IDG2 electrical systems, doing so will be rewarded with a lower/stable oil temperature. However, maintaining climb thrust will further increase oil temperature and trigger an ECAM warning requesting an ENG2 shutdown. 14.5 minutes into the flight the aircraft is struck by lightning which has two effects. One effect is a Thrust Lever Fault of Engine #1 (ENG1), which reduces ENG1 thrust control to binary control: either Max Continuous Thrust (MCT, ~96% N1) or, when either slats or gear are extended, idle thrust. Normally this is not a problem because a functioning autothrust system will be able to still control the engine at all power levels. However, this option is not readily available, as the lightning strike also affects the autothrust system whereby the autothrust "speed mode" does not function (and disengages autothrust), however autothrust "thrust mode" does work. Speed mode is activated in the following common autoflight modes (and a few others): vertical-speed, ALT, GS, FINAL. If any of these modes are selected by the crew, autothrust disengages and ENG1 control returns to binary manual control, controllable solely by slat and gear positions. Crews experience the critical lightning strike about two minutes before they reach a point equidistant from three diversion fields, including their departure (home) field which has become relatively unsafe due to severely deteriorating weather (approaching thunderstorms). They are subsequently left to decide about a return or diversion and manage the energy challenges for the descent and approach."

As this was a departure scenario, crews had enough fuel and time (several hours) to address the problem and were briefed as such to take as much time as they need. The scenario was developed to work for both the DLR A320 AVES simulator and the NLR A330-configured Generic Research Aircraft Cockpit Environment (GRACE) simulator. For this reason, the scenarios in these sims featured slightly different

navigation and endurance settings such that the challenges would be comparable for both an A320 and A330 setup. Except for a few differences in temperature limits, the thrust control systems and failure modes are identical between these two aircraft types.

The scenario was designed to challenge the crews around themes that are relevant for fourth generation aircraft operations and technology and provide the context for which the strategy and RIS has been designed. A comprehensive explanation of the design of this specific scenario can be found in (Niedermaier, 2018), Table 2 provides a summary of the designed effects.

Table 2. Scenario characteristics for specific effects on the crew (adapted from Niedermaier, 2018, Table 3)

Crew effect	Scenario characteristic
Ambiguous decision-making (uncertainty management)	▪ Return to the departure airport to facilitate maintenance vs. risk of airport closure due to approaching thunderstorms ▪ Most suitable diversion airports causing longer flight time with unreliable right-hand engine and increased fuel consumption ▪ Restart/keep right-hand engine after left-hand engine thrust lever fault although violating SOPs on ECAM
Surprise (manage immediate threats)	▪ Thrust lever fault usually not part of standard airline training ▪ Autothrust failure causing sudden disengagement of autothrust
Enhanced monitoring (uncertainty and contingency management)	▪ Detection of high oil temperature ECAM advisory on right-hand engine
Complex long-term planning (contingency management)	▪ Consideration of reduced go-around capabilities due to FADEC behaviour after slat extension in case of thrust lever fault ▪ Consideration of strong crosswind at some of the alternate airports ▪ Consideration of thunderstorm movement to departure airport ▪ Energy management with one engine in IDLE and one in MCT during cruise and descent and both engines almost in IDLE during approach
Uncertainty about system status / confusion (uncertainty management)	▪ Reliability of right-hand engine due to increased oil temperature ▪ Unreliable autothrust function due to autothrust failure
Deep system knowledge (uncertainty and contingency management)	▪ Knowledge of the FADEC behaviour after ENG 1 thrust lever failure

Crews in the two treatment groups (STG and DIS) were trained in the strategy. Training consisted of a briefing document explaining the philosophy principles and the strategy sent per email prior to the session, a two-hour training session to explain the strategy, applying it in several case studies, and practicing it in the simulator with a simple approach scenario featuring an extreme fuel leak. In this practice scenario, the aircraft remained well within range of its destination airfield, despite heavy fuel leakage, and provided crews with an opportunity to integrate the strategy into their operational routines. This allowed the crews from the two treatment groups to also use their quick reference card and familiarize with it in practice. The training and quick reference card provided to the DIS group was slightly augmented compared to the STG group, to highlight and reinforce the combined use of the RIS and strategy.

BSL crews did not receive any training in problem resolution, to allow for natural behaviour. In order to mitigate experimental effects, the BSL crews still flew the fuel leak approach practice scenario as an "acclimatization scenario" to ensure all crews in all groups had approximately the same amount of (familiarisation) time in the simulator. All crews were also invited for a short free-flight session before the fuel leak scenario to become accustomed to the research simulators' peculiarities. Figure 1 illustrates the sequence of simulation scenarios in this study.

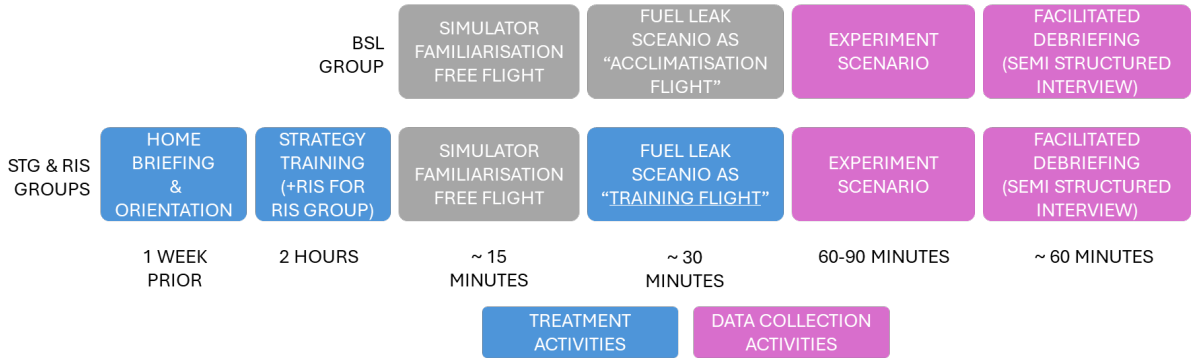


Figure 1. Illustration of the study setup

3.2 Performance and behaviour dependent measures

All data collection was done according to NLR and DLR experiment ethics policies, including a data protection and privacy policy form signed by both subjects and a research team member, and all data being stored responsibly and securely. Participants were allowed to revoke their data being collected or used and could also request their data in accordance with NLR and DLR privacy policies. Experimenters took great

care to avoid any negative training, and to provide participating crew members with ample opportunity to discuss any negative impact the experiment may have had on their professional confidence (no negative impact was established for any participant). Two main datasets were collected from the scenario runs: crew performance data and crew behavioural data.

3.2.1 Crew performance measures

The first dataset, crew performance, was defined as a crew's ability to maintain a high level of safety throughout the scenario. Crews were not explicitly instructed to maximize safety, as the study was semi-naturalistic and aimed to replicate authentic behaviours and prioritizations, except for the strategy and display treatments. Crew performance was measured using the Desired Flight Crew Performance (DFCP) method (Field et al., 2016). The DFCP method was developed earlier in the Man4Gen project and essentially provides a useable performance benchmark to compare crews in a scenario in which there are multiple solution pathways. The DFCP method consists of a list of very scenario-specific desired decisions and actions which are marked as observed (performed) or not. A perfect score is not possible as a crew cannot simultaneously engage in all different solution pathways. In this study the DFCP score was used as a relative measure, and not as an absolute indication of safety performance. The reader is directed to (Field, 2016) for a detailed explanation of the DFCP method.

All DFCP items (30 for this scenario, see Appendix D) were weighted equally with the final performance score being the sum of all correctly executed items. The DFCP items were scored by the main author post-hoc with the use of multi-source video recordings (including display recordings), supported by cross-examination of the crew decisions, thoughts and considerations in post-scenario interviews (also recorded). Performance scores were validated with an active-duty type-rating instructor and examiner to reduce observer bias. Next to the total DFCP score variable, four sub-DFCP variables were defined, based on specific challenges and complexity factors in the scenario. The reason for this is that this accounts for different constellations of local high and low performance which may otherwise cancel out in a total DFCP score (e.g., one crew is effective in challenge A and weak in challenge B, and another crew vice versa). Sub-performance indicators (sub-PIs) were divided according to four main complexity factors and challenges in the scenario:

- ENG2 DFCP: Engine 2 management (oil temperature rising) (11 DFCP items),
- ENG1 DFCP: Engine 1 management (thrust lever fault) (7 DFCP items),
- ROUTE DFCP: Route management (weather, destination choice, descent) (8 DFCP items), and
- COMM DFCP: Communication (with ATC, dispatch) (7 DFCP items).

Note that a few DFCP items overlap in challenges and contribute to two sub-PI's. Two remaining DFCP items could not be readily classified into the above sub-PIs but were still part of the total DFCP score.

3.2.2 Crew behavioural measures

The second dataset consists of crew behavioural measures, which were generated by an observation form that followed the structure of the strategy phases and associated activities. The observation form included a total of 23 behaviours, distributed across the six phases. These were supplemented with six additional indicators whether a phase was verbalized (i.e., did the crew explicitly mention the use of any part of a phase such as "*let's first make a short-term plan*" or "*we've done the first phase, the aircraft is stable*"). The verbalization indicators were used in the Man4Gen study to evaluate the training and investigate how often strategy-trained crews retrieved and used the quick reference card. For this study, those verbalization indicators are not used as performance indicators, as the baseline crews would then suffer an unfair disadvantage because they were neither trained in the strategy, nor were provided the quick reference card to verbalize. The other 23 behavioural indicators were generic enough, however, to observe in all crews regardless of their training. All behaviours were marked along a timeline for all crews. Strategy behaviour markers were recorded and timestamped by the main author post-hoc with the use of the same video recordings as used for the performance data using Noldus© software, generating data as illustrated in Figure 2. Behavioural analysis was validated by a secondary observation performed by a human factors specialist. Note that most behaviour marking occurs after about 14-15 minutes of runtime. This is because scenario non-normal operations, malfunctions and complexities only start developing around this time.

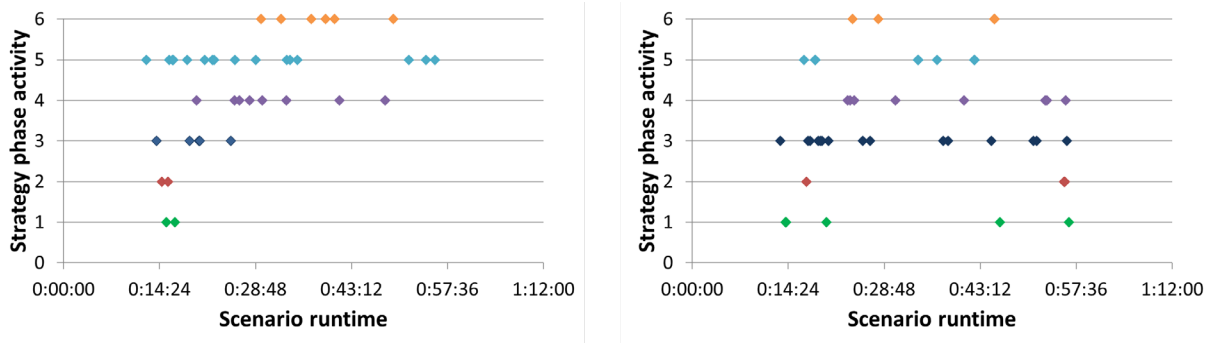


Figure 2. Examples of behavioural mapping of the strategy phases over time, for two different crews.

Behavioural patterns are investigated at two abstraction levels, pertaining to the six-phase strategy as well as the three underlying principles. For the principles analysis, the behavioural indicators among the six phases are clustered as follows: Behavioural observations in strategy Phases 1, 2 and 3 are all considered TM, Phase 4 is UM and Phases 5 and 6 are considered CM. The purpose of this a-priori composite variable aggregation is to reduce noise that may stem from nuances in applying the strategy phases (e.g., P2 "manage immediate threats" may or may not occur if the crew does not have or see an *immediate* threat, but general stabilization/TM may well occur). Composing a smaller set of aggregate indicators from the raw observations also reduces Type II errors that may be present in the large number of measures at the strategy phase level.

After mapping and classifying the basic instances of observed behaviours, the data were processed to distil two subsets of behavioural data:

- phase/principle *duration* (how much time was spent on a particular phase/principle), and
- phase/principle *sequence* (in which order were phases/principles demonstrated).

The methods employed to process the raw behavioural data into these two characteristics are explained in the next two subsections.

3.2.2.1 Behavior duration analysis method

For both duration and sequence analyses, the instance-based behaviour data (as seen in Figure 2) had to be converted into segments of activity for each of the phases. As the exact start and end timing of activity in a specific phase was (obviously) not declared by pilots, nor is it easily identifiable from videos or debriefings, it was chosen to define phase activity time blocks using a mathematical heuristic method dubbed the "blocking algorithm".

The blocking algorithm defines phase activity segment start- and end-times by working chronologically along a given phase, using a time criterion to group observations. The time criterion for this experiment was set at three minutes, as will be explained below. This process begins at the first instance of observing a phase-related behaviour, and from that timestamp on determines if the next observed behaviour of that phase occurs within the subsequent three minutes. If it does, the block is extended to this next point, and the process checks again if there is a subsequent third observation of the same phase within three minutes of this second point, and so on. If no observation of this phase's activity occurs within the next three minutes, the timestamp of the last observation point *plus three minutes* is set as the end time for this block. Figure 3 illustrates the blocking algorithm.

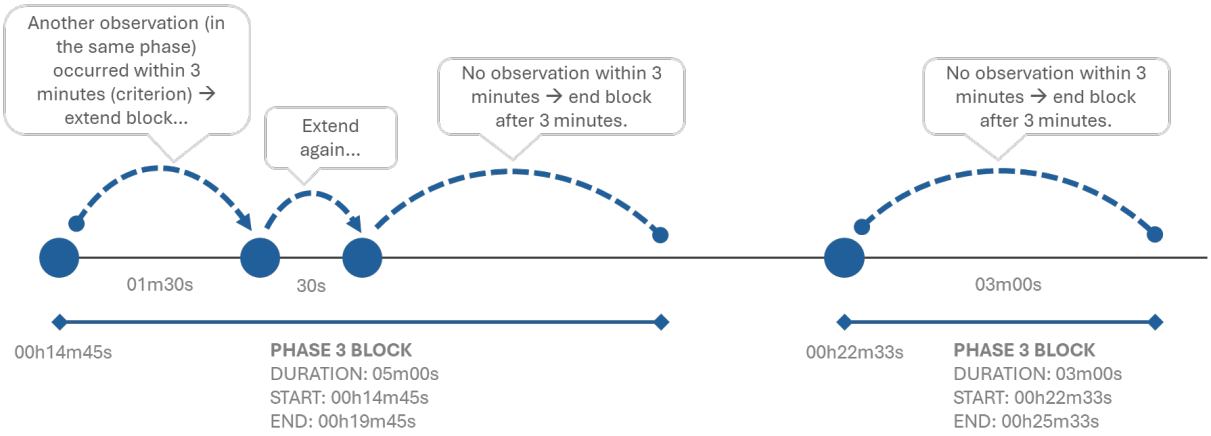


Figure 3. Blocking-algorithm to cluster phase behaviour observation instances (3-minute time criterion)

The time criterion is determined using a multitasking index. This is calculated by summing all blocks across all phases (total "modelled" behaviour) and dividing in by each sample's total scenario runtime, providing a multitasking measure for each sample. A shorter time criterion will result in smaller, erratic time blocking which reduces phase concurrence (multitasking), but may not account for continued cognitive engagement in a phase despite sparse observations. A longer time criterion will result in larger, more continuous time blocking which may increase phase concurrence beyond what may be assumed as cognitively reasonable for a two-person crew. These metrics are calculated from the behavioural data and are unique to the dataset collected for these crews in this scenario. Different time criteria were tested as shown in Table 3.

Table 3. Comparison of several time criteria for the blocking-algorithm

Time Criterion	**1.5 MIN**	**2 MIN**	**2.5 MIN**	**3 MIN**	**4 MIN**	**5 MIN**
Average multitasking	0.92	1.14	1.33	1.50	1.80	2.06
Highest multitasking	1.25	1.53	1.75	1.97	2.34	2.70
Lowest multitasking	0.61	0.75	0.87	0.99	1.17	1.33
σ multitasking	0.197	0.223	0.245	0.268	0.318	0.369

Based on the assumption that each crew member can only actively be engaged in one task rather than truly multitasking multiple tasks concurrently (Loukopoulos et al., 2009), the three-minute criterion is selected as the average multitasking index lies between one and two. Smaller criteria show more gaps in the blocking models (multitasking indices less than one), while longer criteria begin to exceed a multitasking limit of two. The multitasking data presented below are calculated across the entire population (N=14). A Kruskal-Wallis between group analysis of multitasking indices indicated no significant differences between the BSL, STG and RIS groups in terms of their multitasking index, indicating that this three-minute blocking model may be used similarly for the entire study population.

This blocking model provides duration measures for each of the six phases (the runtime sum of all blocks per phase). This is the raw (observed) duration in each phase. These raw measures are further transformed into two other sets of phase duration measures. The first is correcting the duration for multitasking to account for finite cognitive capacity, by dividing the duration of overlap across the number of phases concurrently active, as illustrated in Figure 4.

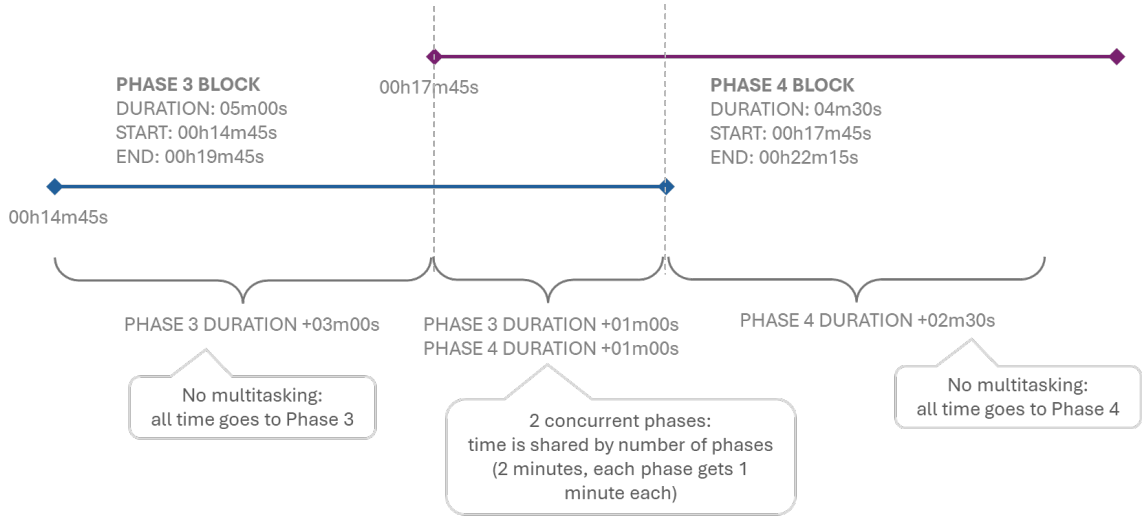


Figure 4. Example of phase duration multitasking correction

This avoids an over-valuation of time "active" in a phase if that phase is constantly run parallel to other activities, thus providing a proxy for a crew's actual division of their attention across the phases. The other correction is subsequent expressing these multitasking-corrected durations as a percentage ratio. This provides a proxy for crew phase preference and selection, as well as allows for a comparison of crew duration patterns despite different runtimes. The six observed, six corrected and six ratio duration measures provide different *lenses* to observe relations of behavioural measures to each other and to performance scores. By observing how correlations may disappear (e.g., when correcting for multitasking) or become stronger (e.g., when correcting for time with ratios), the results can segregate between particularly strong and weak corrections[4]. It is expected that (relatively) more attention spent on later phases (Phase 4, 5 and 6) and (relatively) less on earlier phases (notably Phases 1 and 2) will correlate positively with DFCP performance. Figure 5 illustrates the transformation of these correlation matrices. Note that the study will not correlate different transformations of the same data (e.g., correlating observed data with ratios).

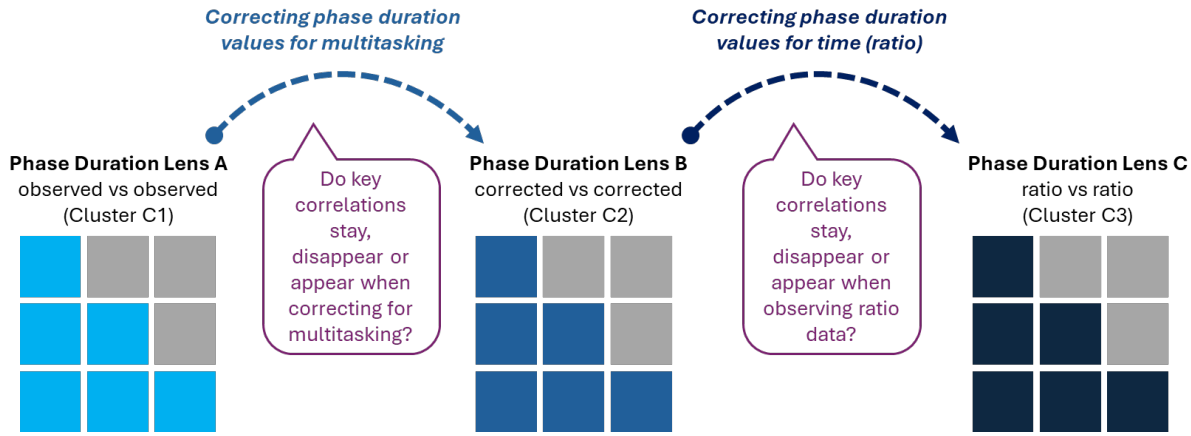


Figure 5. Illustration of the two measures transformations (lenses)

At a higher level of abstraction, the three philosophy principles were also characterized using the same time criterion of three minutes, resulting in three duration measures (one for each principle), which are also subject to the multitasking and ratio transformations. It is expected that (relatively) more attention spent on UM and CM, and (relatively) less on TM will relate to higher DFCP performance.

[4] As a general observation: the transformation to *ratio* will expect to dampen positive correlations as the data is now constrained to 100%. Conversely, negative correlations may become stronger. That is the value of such a relational lens.

3.2.2.2 Behavior sequence analysis method

Next to phase/principle duration, the study also investigated phase/principle sequencing. The sequence analysis characterisation uses the same data structured by the blocking algorithm, and tracks when crews switch their activity from one strategy phase to another, and to which phase they switch. The focus lies on the transitions between phases, to be able to model the order of a crew's activity. This method also assumes that cognitive load and attention is divided equally between all concurrently active phases and therefore can indicate how much attention is being shifted from one phase to another when a given phase time block activates or terminates. Characteristic behavioural patterns in attention switching between phases can then be calculated along the scenario runtime.

Phase switching is evaluated numerically by modelling the amount of crew attention transferred from one phase to another. Figure 6 illustrates this concept for a given constellation of phase time blocks. Observe in Figure 6 the small arrow on the left which indicates that when the Phase 3 time block activates (with Phase 2 still active), the sequence algorithm assumes that 50% of the attention is now shifted from Phase 2 to Phase 3, numerically indicated as J_{23}=50%. The diagram at the bottom of Figure 6 shows how the attention is distributed between different activation and termination events. Similarly, when Phase 5 is activated while Phases 2 and 3 remain active, attention must be redivided from two phases to three phases, resulting in two equal attention switches of about 17% from Phases 2 and 3 to Phase 5.

Figure 6 also illustrates how the sequence algorithm redistributes attention when a phase block terminates. When Phase 5 terminates, this results in a redistribution of attention to the three active phase (2,3 and 4), which increase from 25% to about 33% attention demand per phase. Similarly, when the Phase 2 block terminates, 50% of attention is moved from Phase 2 to Phase 4, which then has 100% of the crew's attention. When no phases are active, attention is set to 0% (the next phase to activate receives 100% attention). However, the last known active phase is still used as the algorithmic departure point if a different phase activates after a period of zero phase activity. It is recognized that the phase switching method assumes that multiple tasks may be concurrently active, at times more than the maximum assumed multitasking index of two for a crew of two persons. However, given that the average multitasking index was found to be 1.5 for this choice of time criterion (see Table 3), such instances will be limited and of relatively short duration.

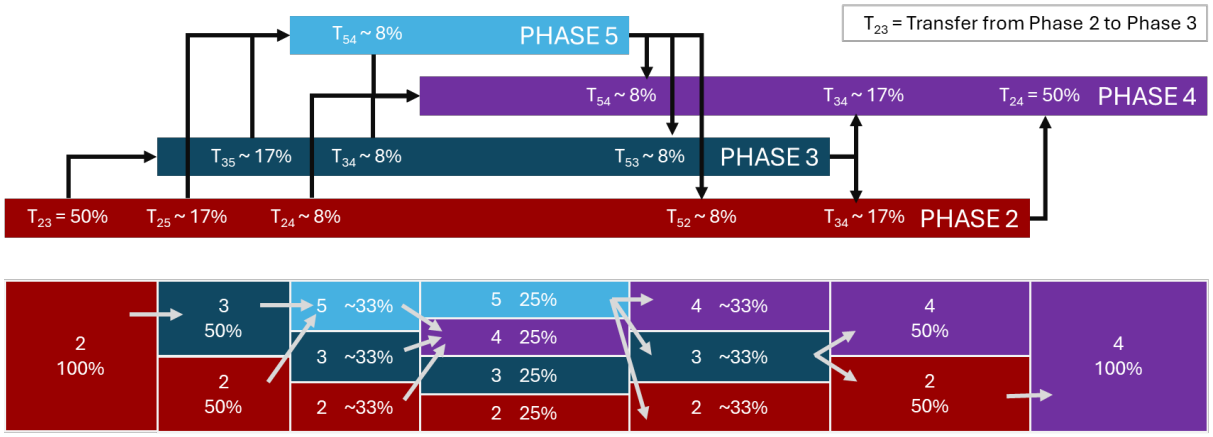


Figure 6. Phase switching example – illustration of attention redistribution (%) between phases

Results of the sequencing algorithm are then organized in a matrix that tracks the runtime sum of all 30 possible attention transfers between phases. The final variable value for each transfer type is the cumulative runtime sum of attention switched as such. Table 4 illustrates how the example block sequence shown in Figure 6 is translated to cumulative sequence data. As Figure 6 only depicts a basic, limited set, a complete scenario may feature cumulative attention switching values exceeding 100% for one or multiple switches.

Table 4. Example of cumulative phase switching based in the illustration in Figure 6.

Figure 6 Example	**From P1 (TM)**	**From P2 (TM)**	**From P3 (TM)**	**From P4 (UM)**	**From P5 (CM)**	**From P6 (CM)**
To P1 (TM)						
To P2 (TM)			**17%**		**8%**	
To P3 (TM)		**50%**			**8%**	
To P4 (UM)		**58%**	**23%**		**16%**	
To P5 (CM)		**17%**	**17%**			
To P6 (CM)						

The 30 phase transfers were further grouped into five characteristic strategy phase sequencing patterns defined with respect to the Man4Gen strategy phase utility, and three characteristic principles sequencing patterns for strategy principles utility. These eight characteristics are illustrated in Figure 7 and described in more detail below.

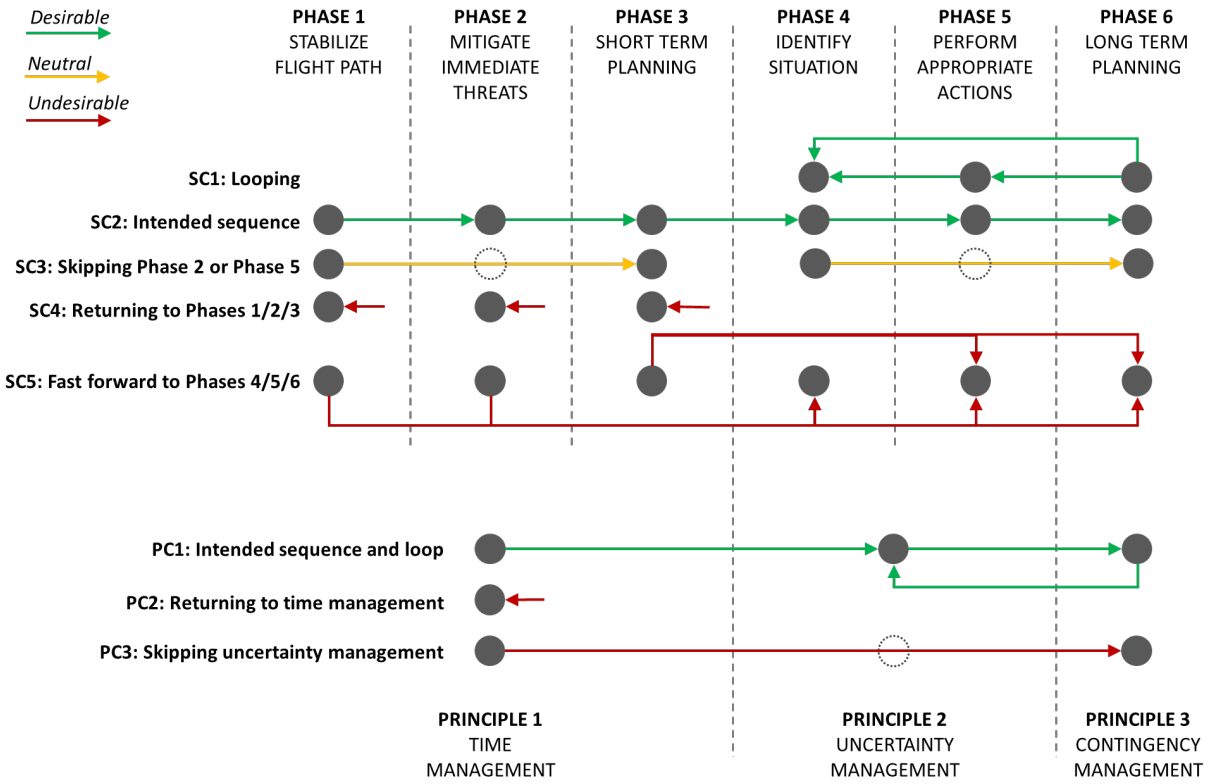


Figure 7. Overview of strategy and principle sequence characterisations

Strategy Phase Characteristic 1 (coded as SC1) refers to the crew repeatedly engaging in cyclic sensemaking with contingency management, by cycling through Phases 4, 5 and 6. While this is chronologically not the first behaviour to observe, it is hypothesized to have the strongest effect on performance. SC2 refer to following the sequence of phases as intended, and monitors how often a crew moves from one phase to the subsequent phase of the strategy. It is expected that these two characteristics (SC1 and SC2) are positively correlated to higher DFCP performance scores. This is based on the expectation that more time spent in later phases (i.e., Phases 4, 5 and 6) will improve crew sensemaking of ambiguous situations and identify better/safer contingency plans. Note that the intended sequence is not to cycle back to Phase 1 as that would indicate ineffective short-term time management. Rather, the intention is that after reaching Phase 5 or 6, the crew returns to Phase 4 for continued monitoring and management of uncertainties and risks (SC1).

SC3 measures whether a crew skips short-term actions (Phase 3) or appropriate recovery actions (Phase 5). By skipping such phases, crews forfeit an opportunity to learn and engage with the system. On the other hand, sometimes there may not be any immediate or recovery actions applicable. This behaviour is expected to correlate slightly negatively with DFCP performance scores. SC4 and SC5 refer to skipping

the initial stabilization phases or returning to them from later phases, respectively. As the Man4Gen strategy was designed to creating stability and time at the start of a developing situation to provide problem-solving space, it is expected that the behaviour of skipping either these initial phases or a necessity to return to them later would correlate negatively with DFCP performance scores

Sample Characterisation Matrix	From P1 (TM)	From P2 (TM)	From P3 (TM)	From P4 (UM)	From P5 (CM)	From P6 (CM)
To P1 (TM)		4	4	4	4	4
To P2 (TM)	2		4	4	4	4
To P3 (TM)	3	2		4	4	4
To P4 (UM)	5	5	2		1	1
To P5 (CM)	5	5	5	2		1
To P6 (CM)	5	5	5	3	2	

SC1 = Looping between Phases 4,5 or 6 (SCSC1) - Desirable
SC2 = Intended sequence from P1 to P6 (SCSC2) - Desirable
SC3 = Skipping Phase 2 or Phase 5 (SCSC3) – Neutral
SC4 = Going back to Phases 1,2 or 3 (SCSC4) – Not Desirable
SC5 = Fast forwarding to Phases 4, 5 or 6 (SCSC5) – Not Desirable

Figure 8. Strategy phase-switching matrix showing the five characteristic behavioural patterns

Each characteristic is numerically expressed (for each crew) as the total scenario runtime sum of all transfers allocated to that SC group (as illustrated by Figure 8). Each SC sum is normalized by dividing that sum by the number of transfers in each SC group to account for equal distribution likelihoods. The total runtime sum for each SC is expected to provide sufficient variation to validate whether desired and undesired characteristic sequences have positive and negative effects on DFCP performance, respectively. These normalized runtime sums are also transformed to an additional set of characteristic ratios as a proxy to crew preference and selection of certain sequencing strategies and allow comparison of behavioural distributions despite different runtimes.

The total scenario runtime sum of a characteristic sequence may feature different underlying patterns. Equivalent total runtime sums may consist of many smaller transfers (i.e., frequent task switching)

or fewer but larger transfers (i.e., more task focus and less switching). illustrates four possible crew modalities in characterisation sequencing.

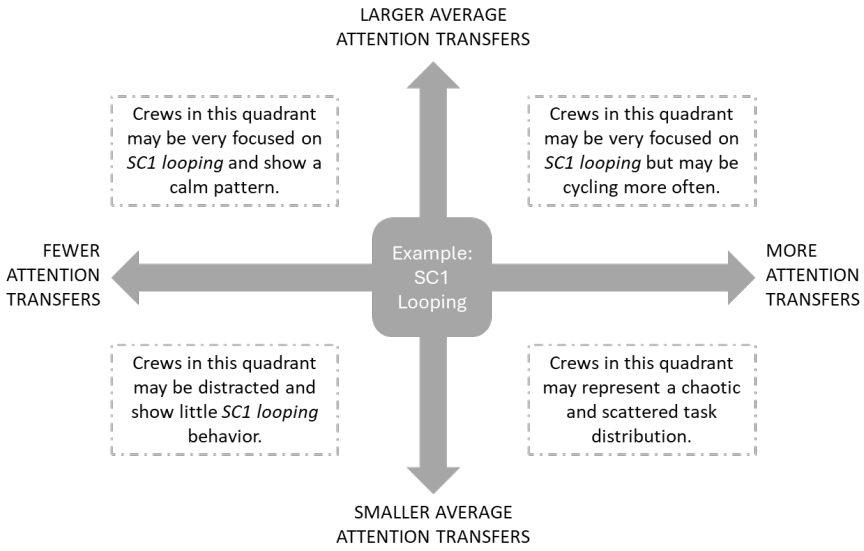


Figure 9. Description of four characterisation modalities

However, given the limited number of samples and highly granular nature of such sub-characterisations (i.e., number of attention transfers and average size of attention transfers), this study will only broadly and qualitatively explore whether crews show large differences in modalities and if sequencing measures should account for this. Subsection 4.2 will provide such a characterisation modality plot for each SC measure for exploratory discussion. For the purpose of hypotheses, correlations and regression analyses, only the total sum and respective ratio measures will be used. This is also to avoid possible strong multicollinearity between the size and number of transfers (i.e., smaller sized transfers will often coincide with more frequent transfers as a result of the phase switching algorithm). Table 5 illustrates a basic example of these variable transformations and normalisations from the blocking described in Figure 6 and Table 4. Figure 10. shows an example of different strategy phase characterisations over the runtime of a scenario.

Table 5. SC measures extracted from the example in Figure 6

SC Measure	Normalized cumulative transfer	Normalized total transfer count	Average Transfer size [5]
SC1	16% ÷ 3 = 5.3%	2 ÷ 3 = 0.67	16% ÷ 2 = 8%
SC2	73% ÷ 5 = 14.6%	3 ÷ 5 = 0.6	73% ÷3 = 24.3%
SC3	0% ÷ 2 = 0%	0 ÷ 2 = 0	0% ÷ 0 = 0%
SC4	33% ÷ 12 = 2.75%	3 ÷ 12 = 0.25	33% ÷ 3 = 11%
SC5	92% ÷ 8 = 11.5%	4 ÷ 8 = 0.5	92% ÷ 4 = 23%

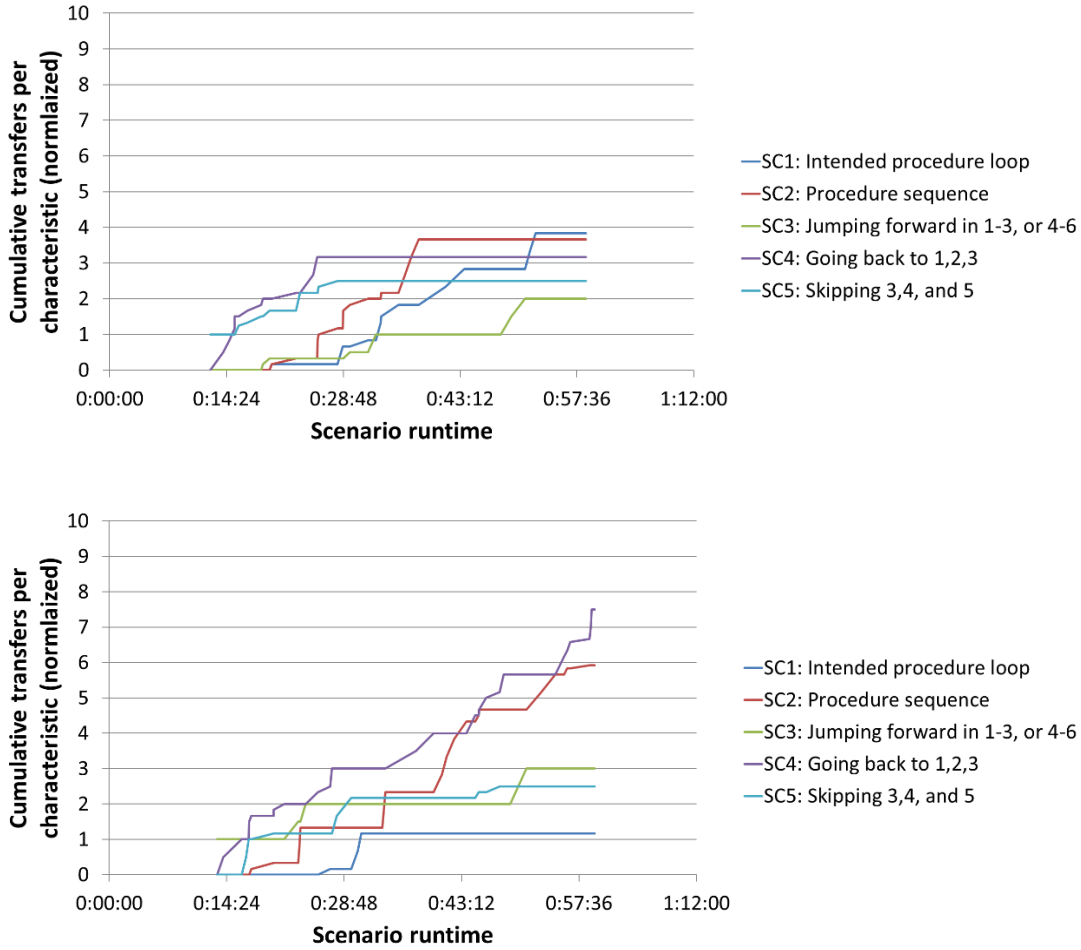


Figure 10. Examples of output of the characterisation analysis, clearly illustrating different behavioural patterns

At a higher abstraction level, crew behaviours were also characterized into the three philosophy principle characterisations. This was performed using a philosophy principles characterisation algorithm and featured a smaller characterization matrix with only six transfers, as illustrated in Figure 11. Three behavioural characterizations are modelled: Principle Characteristic 1 (PC1) describes the desired sequence and looping of principles (similar to a combination of SC1 and SC2) and was expected to correlate positively

[5] As a point of clarification: The average transfer size is not normalized. Dividing the normalized total runtime sum by the normalized total transfer counts provides the average transfer size (normalisation factors cancel out).

with DFCP performance scores. PC2 and PC3 are undesired characteristics, respectively describing a return to TM or skipping UM, both of which were expected to correlate negatively with DFCP performance scores. All PC measures are also transformed to relative ratios to indicate crew preference for sequence patterns.

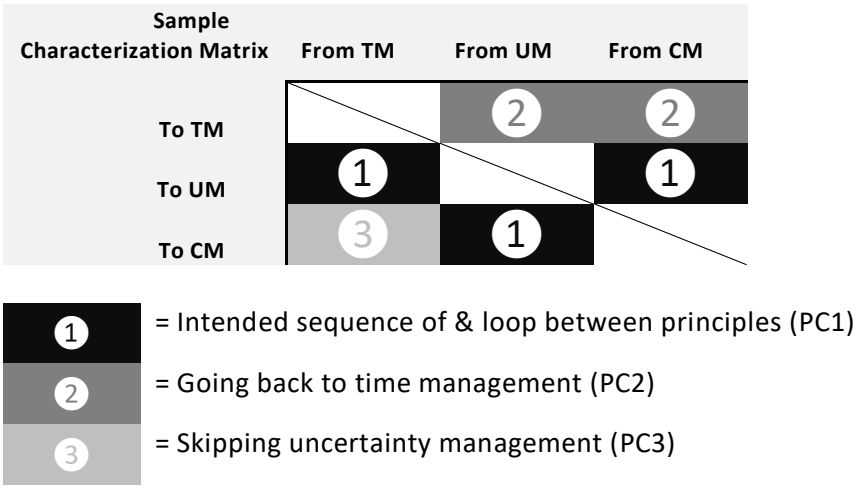


Figure 11. Principles-switching characterization matrix

At both levels of abstraction, the phase switch characterisation approach is limited to analysing the amount of switching between any two phases. While it may be possible to also model and track all 64 unique states (i.e., all possible combinations of phases co-activation) that a crew can be in, the derivative nature of such a behavioural model, the increasing number of assumptions combined with the low number of samples (N=14) would quickly render any analysis invalid.

3.2.3 Summary of measures

The study will feature five performance measures: their total DFCP score as well as the four sub-PI's for Engine 2 management (ENG2 DFCP), Engine 1 management (ENG1 DFCP), route management (ROUTE DFCP) and communication (COMM DFCP). Each (sub-)PI score is expressed as percentage of the total achievable DFCP (sub-) PI. The duration and sequence measures at both the strategy and principle levels result in ten distinct *clusters* of behavioural measures. Each cluster can be regarded as a behavioural *lens* through which interactions may be studied. Table 6 provides an overview of these clusters, their rationale and their ID (to facilitate results reporting).

Table 6. Overview of and rationale for behavioural measures clusters

Cluster ID	Name	Variables	Rationale
C1	Strategy Phase Duration - observed	6 measures *P1-P6 Duration (observed)*	Ability to contrast with multi-tasking corrected durations to observe phase concurrence
C2	Strategy Phase Duration - multitasking corrected	6 measures *P1-P6 Duration (corrected)*	Measures that account for the reality of finite cognitive capacity
C3	Strategy Phase Duration - multitasking corrected (% ratio)	6 measures *P1-P6 Duration (%corrected)*	Ability to contrast patterns in durations and account for runtimes (e.g., crew preferences)
C4	Strategy Characterization – normalized	5 measures *SC1-SC5*	Fair comparison of key sequence patterns (normalized to category number of possible transitions)
C5	Strategy Characterization – normalized (% Ratio)	5 measures *SC1-SC5 (%)*	Ability to contrast patterns in durations and account for runtimes (e.g., crew preferences for sequences)
C6	Principle Phase Duration - observed	3 measures *TM / UM / CM duration (observed)*	Same as C1, but for principle level
C7	Principle Phase Duration – multitasking corrected	3 measures *TM / UM / CM duration (corrected)*	Same as C2, but for principle level
C8	Principle Phase Duration – multitasking corrected (% Ratio)	3 measures *TM / UM / CM duration (%corrected)*	Same as C3, but for principle level
C9	Principle Characterization - normalized	3 measures *PC1-PC3*	Same as C4, but for principle level
C10	Principle Characterization - normalized (% ratio)	3 measures *PC1-PC3 (%)*	Same as C5, but for principle level

3.3 Study hypotheses

Two main hypotheses will be investigated:

Hypothesis H_1: Between groups analysis: Crews trained in the strategy both with and without the RIS (STG and RIS groups) demonstrate behaviour more aligned with the Man4Gen strategy, than control group crews (BSL group). This means (relatively) more attention spent on Phases 4-6 (Principles: UM and CM) than Phases 1-3 (Principles: TM), a (relative) increase in SC1 and SC2 patterns (Principles: PC1) and a (relative) decrease in SC3, SC4 and SC5 patterns (Principles PC2 and PC3).

Hypothesis H_2: Within-group analysis: Crews demonstrating (relatively) more behaviour in line with the Man4Gen strategy (as described in H_1), will feature higher DFCP/safety performance in the test scenario.

3.4 Data analysis methods

The study will analyse data in four ways:

Analysis I Relations between performance measures to confirm sub-DFCP validity (stronger measures to evaluate H_2);

Analysis II Relations with and between duration and sequence measures at the strategy phase- and principle-level, respectively (stronger measures to evaluate both H_1 and H_2);

Analysis III Effectiveness of training in changing STG and RIS group behaviours at both the strategy phase- and principle-level (testing H_1);

Analysis IV Relations between behaviour and performance measures at the strategy phase- and principle-level (testing H_2);

Analysis I investigates DFCP measures for intra-DFCP correlations and sub-DFCP predictor validity (i.e., can the sub-measures adequately proxy for the total score). This is done to ensure the set of performance markers are mutually independent while representing the full scope of performance. Bivariate correlations (Kendall Tau) will identify whether the sub-DFCP measures relate to Total DFCP and/or to one another. A linear regression analysis will be performed to validate the sub-DFCP's as valid constituents for Total DFCP performance. The limitations of linear regression modelling and the risk of overfitting and Type I errors are appreciated. As the initial assumption is that all sub-DFCP variables have predictive power, a backwards stepwise regression is considered the preferred approach, reducing overfitting by removing those measures that contribute the least to the model's prediction power. The measure removal criterion is set at p=0.1[6], monitored for a low Variable Inflation Factor (VIF) (less than 10 indicates low multicollinearity) and highest Adjusted R^2 to prevent underfitting and possible Type II errors (i.e., removal of measures with predictive value).

Analysis II features a comprehensive review to identify and validate correlations between various behavioural measures. This is done to reinforce the behavioural measures model before Analysis III and IV test the two study hypotheses. Duration and sequence behavioural measure sets are initially examined independently (i.e., duration measures correlated with other duration measures), across the different

[6] Twice the minimum significant level of 0.05 is customary (Hosmer & Lemeshow, 2000)

lenses. Robust correlations are those that appear in multiple correlation lenses (e.g., observed, corrected and ratio measures). Subsequently, correlations between duration and sequence clusters are identified and evaluated for their robustness across the data transformations. Awareness of strong relations between behavioural measures and possible covariation between measures will provide additional contrast when evaluating training effectiveness and behaviour-performance correlations.

Analysis III tests H_1 by evaluating differences between the three groups, across all ten behavioural measures clusters. This is done by a Kruskal-Wallis (KW) non-parametric test (BH corrected) to identify significant between-group differences. Significant differences will feature post hoc pairwise testing using Dunn's test (also BH corrected) to identify which groups differed significantly.

Analysis IV tests H_2 by investigating the effectiveness of the Man4Gen strategy by investigating the correlation between behaviours and performance across all crews as a single population. Analysing all crews as a single group for H_2 is appropriate as the variation of crew behaviour and performance (regardless of whether this is random or due to training) provides better data stratification for correlation analysis, as well as a larger sample size. Each performance (sub-)PI will be compared to each behavioural variable by means of a linear bivariate correlation analysis using Kendall Tau to account for non-parametric datasets. BH correction is applied to each analysis cluster (e.g., DFCP vs.C1…C5…C10 etc.).

As a general caution given the low sample size, normality is not assumed given the low power of Shapiro-Wilk tests for N=14. Non-parametric ranking tests will be used to reduce sensitivity to outliers and possibly unequal variances, particularly in small group sizes (e.g., N=3). Furthermore, given the large number of behavioural measures, *p*-adjustment will be consistently applied to account for Type 1 errors (false positives). Adjustment will occur at the cluster level as different cluster of measures represent distinct relations with performance or other behavioural measures. As single cluster-level analysis can feature up to 30 correlations (e.g., 6 duration x 5 SC measures), a Benjamini-Hochberg (BH) False Discovery Rate (FDR) correction is applied to provide a progressive *p*-adjustment that is more appropriate for explorative studies with multiple independent variables[7]. Significance is tested at the p=0.05 level, BH-

[7] Where the alternative Bonferroni correction aims to reduce the family-wide error rate, BH aims to reduce the false discovery rate, which is generally more relevant and applicable for explorative research.

corrected *p*-values are reported as *q*-values and also tested at the *q*=0.05 level, with the following effects strengths being reported:

- $q \leq 0.05$: significant effect (***);
- $p \leq 0.05$ but $0.05 < q \leq 0.1$: near-significant effect (**);
- $p \leq 0.05$ but $q > 0.1$: trend (*); and
- $0.05 < p \leq 0.1$ but with large effect size or correlation coefficient is large[8]: strong effect (Δ).

4 RESULTS

This section presents the performance and behavioural measures as collected and their analysis. Subsection 4.1 inspects the data collected in terms of normality and validates whether the DFCP Sub-PI's are valid segmentations of the total DCP performance (Analysis I). Subsections 4.2 and 4.3 present behavioural data for the strategy phase and principle level respectively, where Subsection 4.4 presents the relations between behavioural measure through multiple correlation lenses (Analysis II). Subsection 4.5 presents the evaluation of training effectiveness (Analysis III, for Hypothesis H_1) and Subsection 4.6 will present the relationships between behavioural and performance measures (Analysis IV, for Hypothesis H_2). The original behavioural and performance data for all crews can be found in Appendices E and F.

4.1 Performance measures results and Analysis I

Figure 12 provides an overview of the DFCP scores. In general, crews have more difficulty (lower DFCP scores) in managing Engine 2 than Engine 1 and showed a large variation in Route Management scores. The Total DFCP scores (leftmost in Figure 12) showed less variation than the other four Sub-PI's. This may be explained as different scenario resolution strategies may vary in their prioritisation of various sub-performance elements (e.g., engine versus route optimisation), while still achieving a similar Total DFCP score. This was in part the reason to recognize sub-PI's.

[8] The following is considered as large effect sizes for tests used: Kruskal-Wallis 0.14; Dunn post-hoc test > 0.5; Kendall $\tau > 0.5$ ($\tau > 0.3$ is considered moderate).

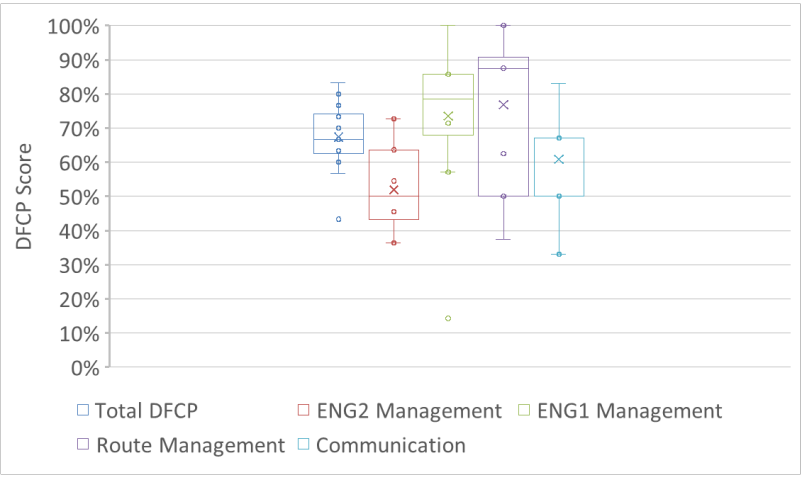


Figure 12. DFCP performance measure box plots

Despite its weak power for low samples sizes, Shapiro-Wilks testing for DFCP data normality indicated that ENG1 (p=0.005), ROUTE (p=0.01) and COMM (p=0.002) data are not normally distributed, confirming the use of non-parametric testing for all performance-related comparisons. Kendall Tau non-parametric correlation test statistics in Table 7 indicate that the performance metrics ENG2, ENG1 and ROUTE sub-PIs are each correlated with the TOTAL DFCP score while also featuring no significant correlations between themselves.

Table 7. DFCP variable correlation analysis - Kendall Tau with p and (q)

	TOTAL DFCP	**ENG2 DFCP**	**ENG1 DFCP**	**ROUTE DFCP**
ENG2 DFCP	0.597*** *0.006 (0.03)*			
ENG1 DFCP	0.473* *0.034 (0.113)*	0.288 0.211		
ROUTE DFCP	0.665*** *0.003 (0.03)*	0.327 *0.154*	0.187 *0.425*	
COMM DFCP	-0.120 *0.601*	-0.0626 *0.791*	-0.314 *0.195*	-0.0657 *0.786*

*** Corrected p (q) is significant (q ≤ 0.05) (two-tailed)
* Uncorrected p is significant, trend (p ≤ 0.05; q > 0.1) (two-tailed)

A further backward stepwise regression[9] analysis (exit criterion p=0.1) indicated that, together, these same sub-PI's can indeed account for the majority of the TOTAL DFCP score variance (adjusted R^2 = 0.951), while featuring limited multicollinearity with all independent variables' VIFs close to one.

4.2 Behavioral measures results – Strategy phase level

The distributions of strategy phase duration measures (observed, corrected, and ratio) are shown in Figure 13. Overall, crews spent more time in later phases than in the initial phases. This is in general alignment with the scenario challenges with were more focused on uncertainty and contingency management, rather than immediate time management. It is notable that all crews spent very little time in Phase 2 ("Manage Immediate Threats"), and that there is considerable variation in the time spent in later phases, notably Phases 5 ("Perform Actions") and Phase 6 ("Long-Term Planning"). Correction of multitasking reduces the total durations (which is expected), but the general pattern appears consistent, also for the transformation to ratios.

[9] Ordinary Least Square (OLS) linear regression is permitted for non-normal variables as inspection of regression residuals shows a normal residual distribution, which is the case for this particular regression.

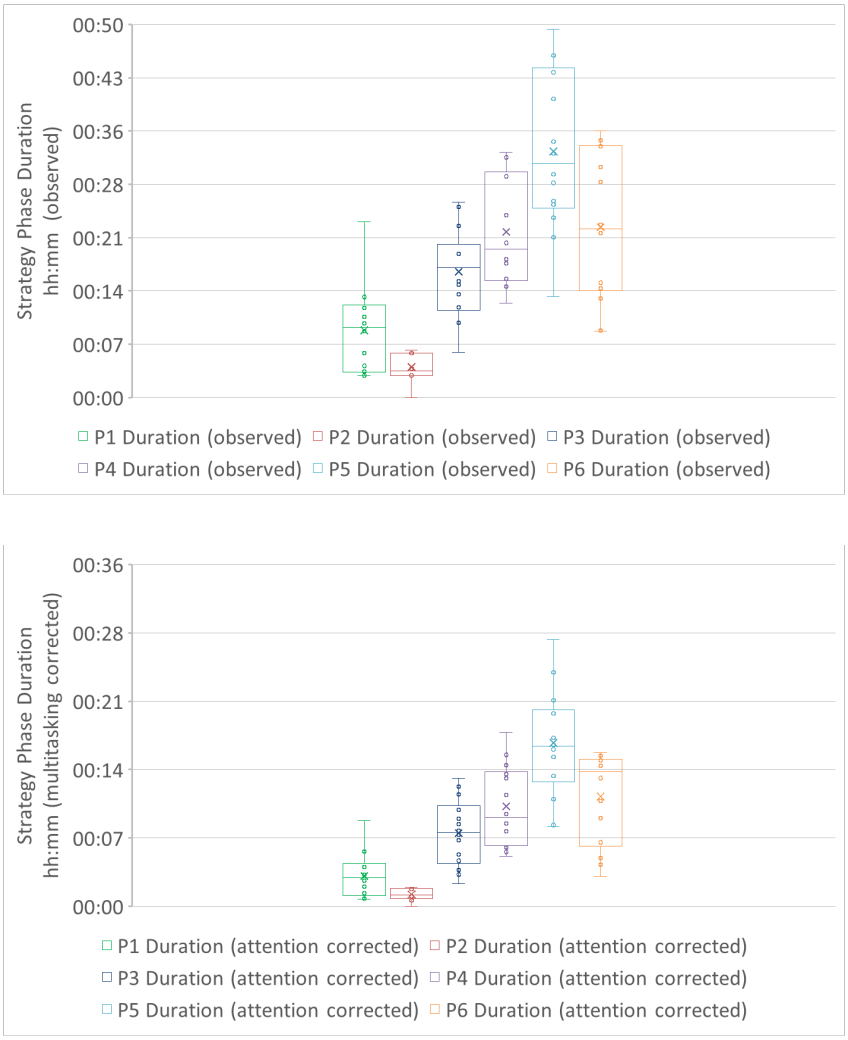


Figure 13. Strategy phase duration measures box plots (observed, corrected & ratio)

The distributions of strategy characterisation measures are shown in Figure 14 (normalized runtime cumulative transfers, and ratio variant) and indicate that crews feature considerably more SC1 (desired) and considerably less SC4 and SC5 (undesired) behaviour sequences.

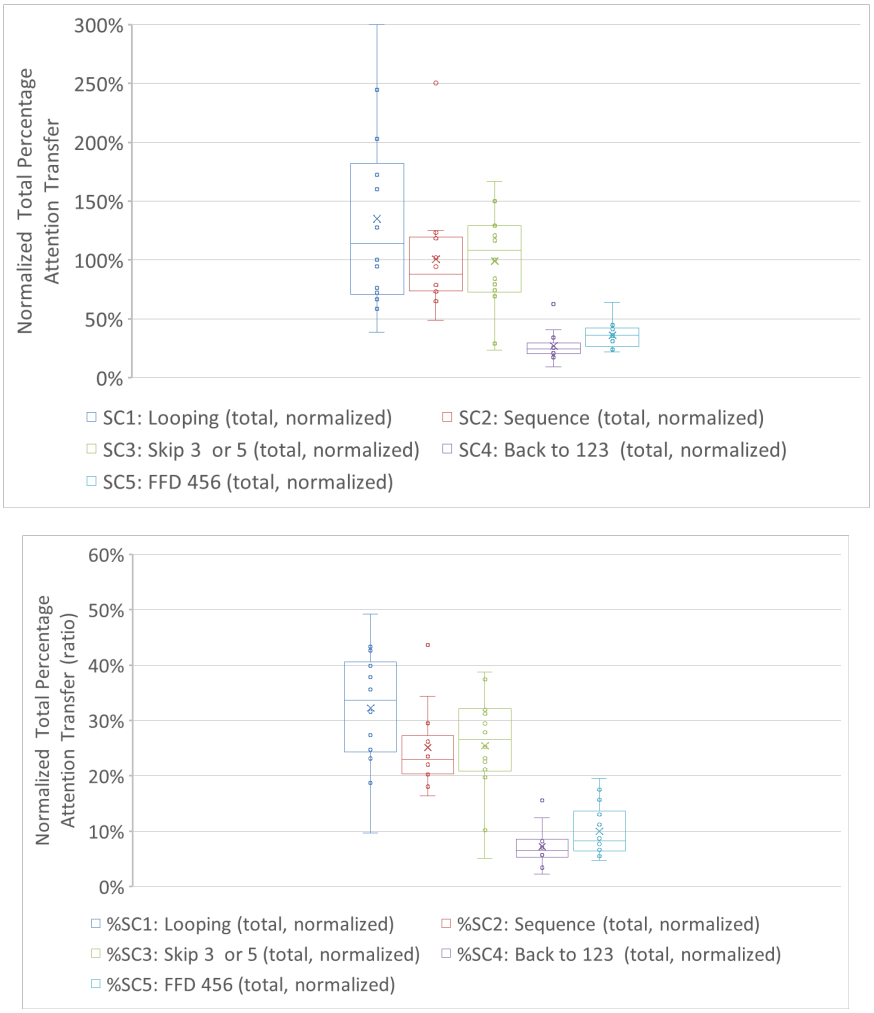


Figure 14. Strategy characterisation measures box plots – runtime cumulative attention transfer and ratios

Figure 15 presents the distribution of sub-characterisation measures (total count of transfers and average transfer size) used for a qualitative exploration of characterisation modalities. A greater number of transfers for SC1 in Figure 15 can be attributed to the fact that it is a repeating action, rather than a one-time action such as SC2. SC1 also shows a greater variability between crews, which may be related to variations in performance. Average attention transfers seem more consistent across the SC measures.

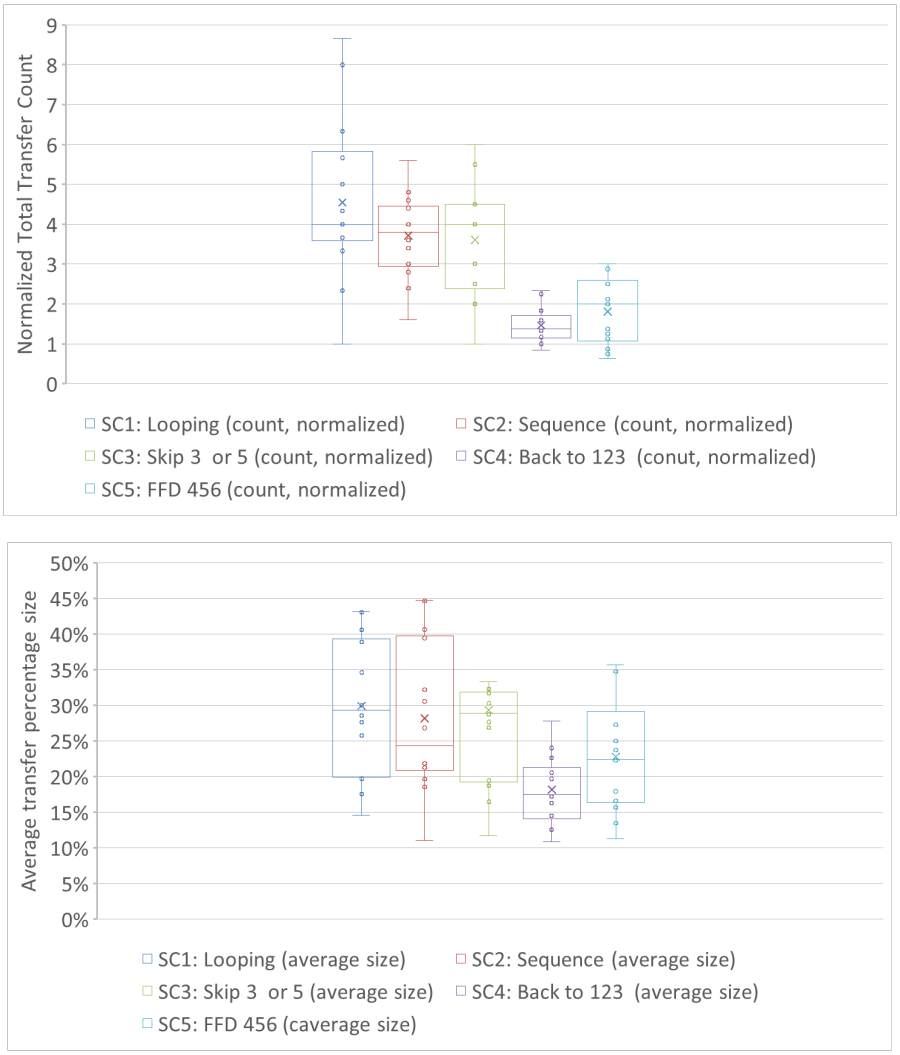


Figure 15. Strategy characterisation measures box plots – transfer count and average transfer sizes

A review of the characterisation modalities (i.e., whether attention transfers were notably large or frequent), did not show clear differences between the top and bottom half of performers (based on Total DFCP) for any of the SC measures, with only SC1 (desirable) possibly indicating that higher performers featured more frequent attention transfers, which would be consistent with increased SC1 cycling behaviour. Given these findings, the paper will proceed to only consider the total transfer variables for the SC measures. The results of the modalities qualitative review can be found in Appendix G.

4.3 Behavioral measures results – Principle level

In addition to the strategy phase-level results, Figure 16 through Figure 18 present the box plots for principle level duration measures (observed, corrected & ratio), principle sequence measures (normalized cumulative total transfers, and ratios) and the sub-characterisation measures (transfer count and average attention transfer size). Multitasking correction reduces TM and UM activities considerably,

possibly indicating that these activities are practiced more concurrently than CM activities. Ratio analysis clearly shows that crews focus on TM and CM more than UM. Crews exhibit (relatively) more PC1 (sequence & looping, desirable) but also some PC3 (Skipping UM, undesirable), which echo's the duration results showing less time spent in UM. Average transfer and transfer count sub-characterisations follow a similar pattern as the total attention measures.

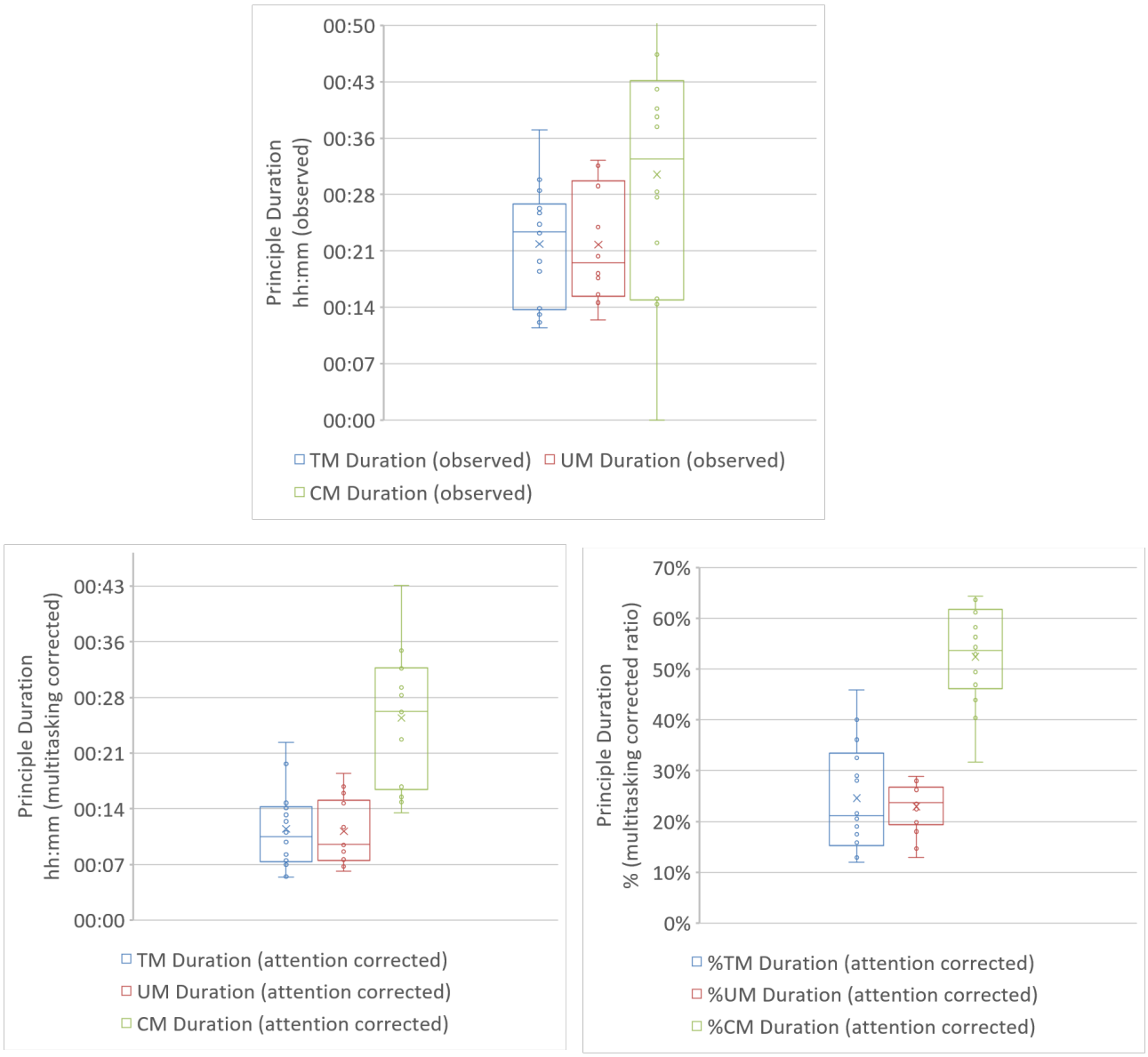


Figure 16. Principle duration measure box plots (observed, corrected & ratio)

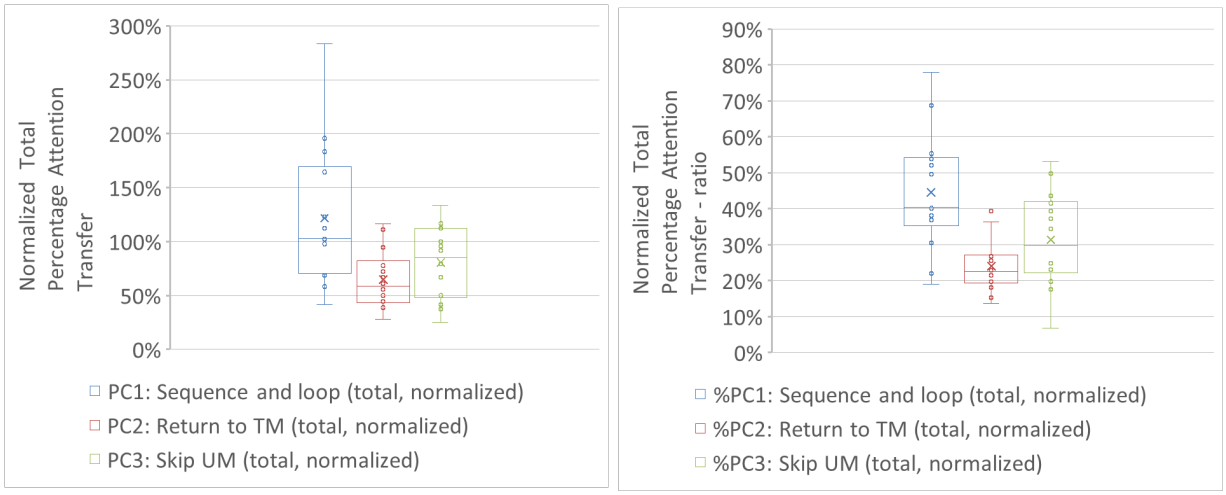


Figure 17. Principle sequence measure box plots –runtime cumulative attention transfer and ratios

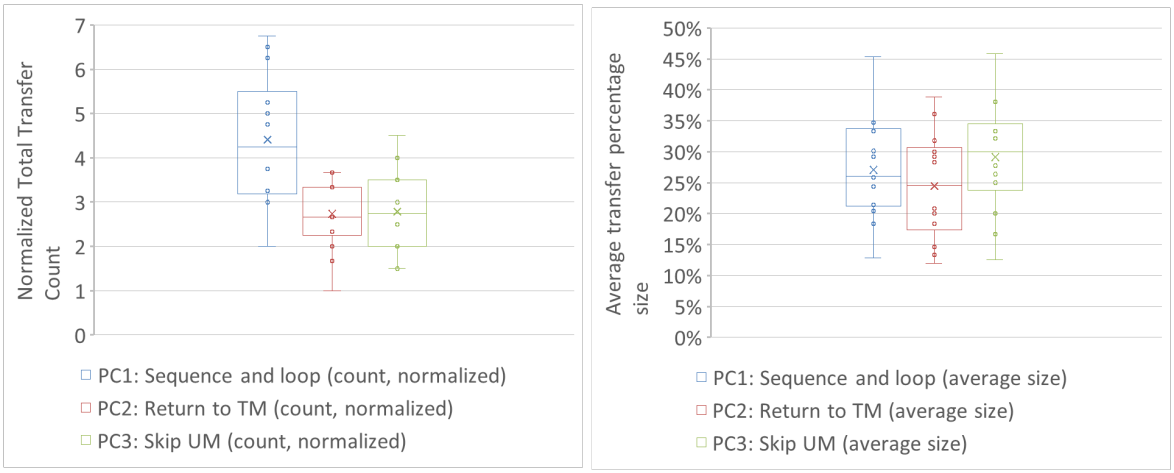


Figure 18. Principle sequence measure box plots – transfer count and average attention transfer size

Similar to the procedure phase level analysis, characterisation modalities for principle characteristic sequences did not present clear differences between the top and bottom half of performers. Principle level modality results can be found in Appendix G as well.

4.4 Behavioural measures correlation – Analysis II

This subsection investigates relations within the behavioural model. As several measures are non-parametric, correlation tests will use Kendall's Tau and with BH *p*-adjustments per measures cluster. Three main correlations are presented in the subsections below: duration versus duration measures, sequence versus sequence measures and duration versus sequence measures. Within each of these correlations, the same set of measures (e.g., six Phase Duration measures) are correlated through different lenses, which provide different insights into behavioural patterns. Duration measures have three correlation lenses

(observed; corrected; ratio), characterisation measures have two lenses (normalized total sum; ratio) and their cross-correlation will feature six lenses (2x3 lenses). Correlations that persist across different lenses are particularly strong, which is reported as a robustness indicator indicating how often the correlation is (near) significant. Strategy phase and principle-level measures are kept isolated from each other as they are two different abstraction levels of the same data.

4.4.1.1 Between-duration measures correlation

Duration measures feature three lenses: observed duration (co-occurrence lens), multi-tasking corrected duration (attention lens) and the ratio of corrected durations (pattern lens).

Table 8 only presents the duration measure correlations which are (nearly) statistically significant (including BH correction separate per lens) within its own matrix/lens, and how they fare in adjacent lenses. Both strategy phase level and principle level correlations are presented (ratios are calculated separate for strategy phase duration measures and principle duration measures).

Table 8. Overview of between-duration measure correlations (strategy phase and principle level)

Duration Correlation	Observed τ (*p*)	Corrected τ (*p*)	Corrected Ratio τ (*p*)	Robustness	Interpretation
P2-P5	-0.389 (0.055)	-0.495* (0.014)	-0.077 (0.702)	1/3	Trend: more attention provided to P2 relates to less attention to P5
P3-P5	-0.319 (0.112)	-0.275 (0.171)	-0.538* (0.007)	1/3	Trend: P3 and P5 proportions vary inverse with each other
P4-P5	0.407* (0.043)	0.407* (0.043)	-0.143 (0.477)	2/3	Trend: P4 and P5 co-occur and also correlate in attention
P5-P6	0.538* (0.007)	0.187 (0.352)	-0.033 (0.87)	1/3	Trend: P5 and P6 co-occur but after discounting them for multitasking they don't correlate in attention.
TM-UM	0.011 (0.956)	-0.077 (0.702)	-0.385** (0.055)	1/3	After correcting for runtimes with ratio, TM and UM feature a strong negative correlation
TM-CM	-0.143 (0.477)	-0.209 (0.298)	-0.648*** (0.001)	1/3	After correcting for runtimes with ratio, TM and CM feature a strong negative correlation
UM-CM	0.143 (0.477)	0.429*** (0.033)	0.033 (0.87)	1/3	After correcting for multitasking, UM and CM show a positive correlation in attention division

*** Corrected p (*q*) is significant ($q \leq 0.05$) (two-tailed)
** Uncorrected *p* is significant, q is near-significant ($p \leq 0.05$; $0.05 < q \leq 0.1$) (two-tailed)
* Uncorrected *p* is significant, trend ($p \leq 0.05$; $q > 0.1$) (two-tailed)

At the strategy-phase level, no significant correlations (passing BH correction) appear but there are trends with moderate ($|\tau| > 0.3$) to strong effect sizes ($|\tau| > 0.5$), which must of course be interpreted with caution. The principle-level analysis provides clearer relations between duration metrics, where UM and CM activities seem to share attention together, and both are negatively related to the proportion of attention spend on TM. These results are supported by the trends at the strategy phase level.

In summary, attention is shared between later phases, but not with initial ones. The inverse may also be true: more time spent in earlier phases reduces time/attention in later phases. Apart from ratio measures[10], the duration values are not zero-sum bounded. Therefore, given the large spread of durations between crews across most phases, this suggests the existence of two distinct crew modalities: 1) being more focused on UM and CM or 2) being more engaged in TM. Correlations with performance should indicate whether this inverse relation has a performance impact.

4.4.1.2 Between-characterisation measures correlation

Characterisation measures feature two lenses: calculated total runtime sum of attention transfers (normalized to number of possible transfers per characterisation) and the ratio between these characterisations (separate for strategy phase and principle-level characterisations). Table 9 summarizes (near) significant correlations for both lenses.

Table 9. Overview of between-characterisation measure correlations (strategy phase and principle-level)

Char. Correlation	Total sum τ (*p*)	Char. Ratio τ (*p*)	Robustness	Interpretation
SC1-SC4	-0.221 (0.273)	-0.824*** (0)	1/2	More looping (SC1) is often seen with less returning to P1/2/3 (SC4)
SC1-SC5	-0.044 (0.826)	-0.407* (0.043)	1/2	Trend: More looping (SC1) is often seen with less skipping ahead (SC5)
SC4-SC5	0.144 (0.475)	0.451* (0.025)	1/2	Trend: More skipping ahead (SC5) is often seen with returning to P1/2/3 (SC4)
PC1-PC3	-0.344 (0.089)	-0.736*** (0)	1/2	More sequence & looping (PC1) is often seen with less skipping of UM (PC3)

*** Corrected p (q) is significant ($q \leq 0.05$) (two-tailed)
* Uncorrected p is significant, trend ($p \leq 0.05$; $q > 0.1$) (two-tailed)

[10] Note: ratio measures do clearly amplify this negative correlation.

Both strategy phase and principle-level correlations indicate that desired sequences (e.g., SC1, PC1) feature a negative correlation to undesired sequences (e.g., SC4, SC5, PC3). While this may seem trivial, this polarisation is an important validation of the model where the characteristic measures seem to be complementary and able to model distinct crew behavioural patterns.

4.4.1.3 Between-characterisation measures correlation

The last behavioural correlation is between duration and characterisation measures. As behavioural measures are subject to three lenses and characterisation measures to two, there are six correlation variations as such, both for the strategy phase and principle-level measures. Table 10 presents all (near) significant correlations. As there are several significant correlations, the focus is on whether the correlation is significant (BH corrected) and whether it has a high robustness score. Only results with a robustness score of 3/6 or higher are presented.

Results show clear positive relations between desired behavioural patterns SC1/SC2/PC1 and increase time and attention for uncertainty management and contingency management, and a negative relation to short-term planning time and attention (P3). Undesired patterns such as SC4/PC2 are positively correlated to short-term planning (P3/TM) and negatively related to contingency management. These results confirm that duration and characterisation measures co-occur in predictable patterns. Desired behaviours are related to more time in UM and CM activities (which is what those desired behaviours were designed for), while undesired behaviours clearly shift the mode of activity to initial phases. Results are similar at both the strategy and principle level analysis and align with the negative relation between early and later phases described in Subsection 4.4.1.1, and the negative relation between desired and undesired characteristics in Subsection 4.4.1.2.

Table 10. Overview of correlation lenses between duration (D-) and characteristics (C-) measures (strategy phase and principle level)

Correlation	D-Observed C-Total sum τ (*p*)	D-Corrected C-Total sum τ (*p*)	D-Ratio C-Total sum τ (*p*)	D-Observed C-Ratio τ (*p*)	D-Corrected C-Ratio τ (*p*)	D-Ratio C-Ratio τ (*p*)	Robustness	Interpretation
SC1-P3	-0.253 (0.208)	-0.319 (0.112)	-0.363 (0.071)	-0.473* (0.019)	-0.451* (0.025)	-0.538** (0.007)	3/6	Trend: Relatively more looping (SC1) is related to less time and attention in short-term planning (P3)
SC4-P3	0.663*** (0.001)	0.685*** (0.001)	0.641*** (0.001)	0.429* (0.033)	0.451* (0.025)	0.538** (0.007)	6/6	More returning to P1/2/3 (SC4) is related to more time and attention in P3
SC1-P4	0.473* (0.019)	0.582*** (0.004)	0.429* (0.033)	0.385 (0.055)	0.407* (0.043)	0.297 (0.139)	4/6	More looping (SC1) is related to more attention to uncertainty management (P4)
SC2-P4	0.575** (0.004)	0.685*** (0.001)	0.663*** (0.001)	0.099 (0.622)	0.033 (0.87)	0.055 (0.784)	3/6	More sequence (SC2) is related to more attention to uncertainty management (P4)
SC1-P5	0.407* (0.043)	0.56*** (0.005)	0.121 (0.547)	0.451* (0.025)	0.560** (0.005)	0.253 (0.208)	4/6	More looping (SC1) is related to more attention to appropriate actions (P5)
SC4-P5	-0.398* (0.048)	-0.309 (0.125)	-0.53** (0.008)	-0.451* (0.025)	-0.560** (0.005)	-0.297 (0.139)	4/6	More returning to P1/2/3 (SC1) is related to relatively less time in appropriate actions (P5)
PC2-TM	0.641*** (0.001)	0.685 (0.001)	0.354*** (0.079)	0.56*** (0.005)	0.604*** (0.003)	0.692*** (0.001)	5/6	More returning to TM (PC2) is related to more time and attention in TM
PC1-UM	0.582** (0.004)	0.846*** (0)	0.538** (0.007)	0.495** (0.014)	0.626*** (0.002)	0.451** (0.025)	6/6	More sequence and looping (PC1) is related to more time and attention in UM
PC2-CM	0.066 (0.742)	-0.088 (0.661)	-0.42* (0.037)	-0.143 (0.477)	-0.429** (0.033)	-0.473** (0.019)	3/6	Relatively more returning to TM (PC2) is related to less attention to CM

*** Corrected p (*q*) is significant ($q \leq 0.05$) (two-tailed)
** Uncorrected *p* is significant, q is near-significant ($p \leq 0.05$; $0.05 < q \leq 0.1$) (two-tailed)
* Uncorrected *p* is significant, trend ($p \leq 0.05$; $q > 0.1$) (two-tailed)

4.5 Training effectiveness results – Analysis III

A between group analysis is performed to evaluate whether different treatments (strategy training with/without RIS) changed behaviour (hypothesis H_1). All tests will be non-parametric Kruskal Wallis and post-hoc Dunn tests due to small sample sizes for the three groups (BSL N=3, STG N=7, RIS N=4). BH *p*-adjustment is performed within each cluster of measures, as well as for Dunn tests (N=3 groups). Near-significant KW results with strong effect sizes are investigated with post-hoc testing but without BH as these results are only presented as indicative trends. All results are summarized in Table 11.

The STG crews spent more time and attention on P2 than other two groups (BSL, RIS), also after correcting for multitasking and runtimes. This may be the result of the STG crews having more awareness of P2 immediate threat management than the untrained BSL crews, while the RIS crews may in turn have had more focus on later phases than the STG crews due to the RIS supporting UM and CM activities. This is counter to the hypothesis that the strategy training would primarily increase time and attention in later phases. The negative correlation between earlier and later phase durations suggests that the training may have missed its mark with the STG crews.

There are non-significant indications that BSL crews spent more time and relative attention on P4/UM (uncertainty management) than STG and RIS crews, and that the BSL crews skipped ahead more (SC3) than the RIS crews. This could be attributed to a less structured approach for UM in BSL crews. In contrast, the STG and RIS groups did not spend more time in UM, which also suggests the training may not have been as effective as thought.

The most profound group effect discovered was that the RIS group showed much more observed behaviour in P5 and P6 than the STG group, possibly due to the addition of the RIS. However, this effect isn't present when correcting for multitasking, which may indicate that while RIS crews have been observed performing P5 and P6 activities more often, they haven't actually allocated more attention to them (i.e., other concurrent tasks are discounting this). That may be the benefit of the display: it offloads some of the P5 and P6 tasks and allows for a wider distribution of cognitive capacity (i.e., to P4 uncertainty diagnostics). In summary, while the results cannot confirm an effective training effect, there are indications that the RIS may have supported more effective management of ambiguity and opacity.

Table 11. Significant and near-significant training effects (strategy phase and principle levels combined)

Cluster & measure	KW effect	KW sig *p* (BH *q*)	Dunn pairs	Dunn effect	Dunn sig *p* (BH *q*)	Median comparison & interpretation
C1 P2 Duration (observed)	0.582***	<0.001 (<0.001)	BSL-STG	0.897***	0.005 (0.015)	STG shows 03m00s more time in P2 manage immediate threats than BSL, and 02m50s more than RIS. BSL and RIS have the same observed P2 durations.
			BSL-RIS	0.193	0.522 (0.522)	
			STG-RIS	0.206***	0.019 (0.029)	
C1 P4 Duration (observed)	0.203Δ	0.083	BSL-STG	0.642*	0.042	Strong KW/Dunn effect: BSL shows +08m46s more time in P4 uncertainty management than STG.
			BSL-RIS	0.189	0.531	
			STG-RIS	0.444	0.141	
C1 P5 Duration (observed)	0.413***	0.008 (0.024)	BSL-STG	0.343	0.255 (0.255)	RIS median shows 20m16s more time in P5 appropriate actions than STG.
			BSL-RIS	0.377	0.211 (0.317)	
			STG-RIS	0.879***	0.005 (0.015)	
C1 P6 Duration (observed)	0.407***	0.009 (0.018)	BSL-STG	0.209	0.488 (0.488)	RIS median shows 18m59s more activity in P6 long-term planning than STG.
			BSL-RIS	0.495	0.1 (0.15)	
			STG-RIS	0.874***	0.006 (0.018)	
C2 P2 Duration (corrected)	0.567***	0.001 (0.006)	BSL-STG	0.662***	0.028 (0.042)	STG shows 00m41s more attention on P2 immediate threats than BSL, and 01m01s more than RIS. RIS and BSL have same P2 duration.
			BSL-RIS	0.110	0.715 (0.715)	
			STG-RIS	0.904***	0.004 (0.012)	
C3 P2 Duration (%corrected)	0.567***	0.001 (0.006)	BSL-STG	0.662***	0.028 (0.042)	STG shows 1.52% more relative attention on P2 manage immediate threats than BSL, and 1.95% more than RIS. BSL and RIS have same P2 duration ratios.
			BSL-RIS	0.110	0.715 (0.715)	
			STG-RIS	0.904***	0.004 (0.012)	
C3 P4 Duration (%corrected)	0.244Δ	0.057	BSL-STG	0.587Δ	0.052	Strong KW/Dunn effect: BSL shows 4.81% more relative attention on P4 uncertainty management than STG, and 7.72% more than RIS. STG and RIS have the same P4 duration ratios.
			BSL-RIS	0.701*	0.027	
			STG-RIS	0.168	0.577	
C4 SC3: *Skip P3 or P5* (undesirable)	0.199Δ	0.067	BSL-STG	0.403	0.181	Strong KW/Dunn effect: BSL median shows 55% more total attention transfers skipping ahead to P4-6 than RIS.
			BSL-RIS	0.718Δ	0.069	
			STG-RIS	0.390	0.195	
C6 UM Duration (observed)	0.203Δ	0.083	BSL-STG	0.642*	0.042	Strong KW/Dunn effect: BSL median shows 8m46s more time in uncertainty management than STG.
			BSL-RIS	0.189	0.531	
			STG-RIS	0.444	0.141	

*** Corrected p (*q*) is significant ($q \leq 0.05$) (two-tailed)

** Uncorrected *p* is significant, q is near-significant ($p \leq 0.05$; $0.05 < q \leq 0.1$) (two-tailed)

* Uncorrected *p* is significant, trend ($p \leq 0.05$; $q > 0.1$) (two-tailed)

Δ Strong effect size, Uncorrected *p* is not significant, trend ($0.05 < p \leq 0.1$) (two-tailed)

4.6 Behaviour versus performance correlation – Analysis IV

This section will investigate whether specific behaviours relate to improved DFCP performance (hypothesis H_2) by evaluating the entire study population in a within-group analysis at both the strategy phase and principle level. Table 12 summarizes all (near-)significant correlations between all behavioural measures and the five DFCP performance indicators. While several correlation results feature a significant *p*, post-BH *p*-adjustments indicate no significant correlations after correcting for false discovery rates. Those results that featured a significant *p* are still presented and discussed for general trends.

In terms of duration measures, more time and attention in short-term planning (P3) is related to lower Route DFCP scores. Conversely, more time and attention toward appropriate actions (P5) is related to higher Route DFCP scores. This also reflects the inverse relationship between P3 and P5 durations and suggests that improved route management requires taking more time for and bringing attention to contingency management (i.e., future threats, also related to routing) rather than spending a lot of time and attention on near-term (possibly reactive) routing decisions. This result is consistent at the principle level which shows a negative relation between TM and Route DFCP, and inversely a positive relation between CM and Route DFCP. Other relations with duration measures indicated that both Total DFCP and ENG1 DFCP performance benefited from more activity in long-term planning (P6). Particularly for the ENG1 DFCP, this is relevant as it presents considerable energy management considerations with an unreliable autothrust system and binary engine control alternatives. Finally, ENG2 DFCP improves with more attention towards P1 stabilize flight path. This is reasonable, considering the impact on the flight path to mitigate and manage ENG1 overheat issues.

In terms of behavioural characteristics, only Route DFCP scores featured correlations with these measures. SC1 (looping P4/5/6) related to increased Route DFCP scores, possibly reflecting a re-appraisal of route options as they appear or close out. Interestingly, skipping to these later phases (SC4, PC3) correlated negatively with Route DFCP. In other words, time spent in and cycling of later phases (P4/5/6) increases Route DFCP but does require some attention in earlier phases (albeit limited, looking at the negative relation between P3 and Route DFCP). This may prove to be an indication that some sequencing is valuable to stabilize the flight and buy time (and cognitive capacity) to engage in the sensemaking done in the later phases.

Table 12. (Near-)significant correlations between behaviour and performance (strategy phase and principle levels combined)

Cluster & Behavioural measure	DFCP Measure	Kendall τ	Sig *p* (BH *q*)	Interpretation
C1 P3 Duration (observed)	Route DFCP	-0.626*	0.004 (>0.1)	Trend: More time spent on short-term planning P3 worsened route management
C1 P5 Duration (observed)	Route DFCP	0.476*	0.028 (>0.1)	Trend: More time spent on appropriate actions P5 improved route management
C1 P6 Duration (observed)	Total DFCP	0.412*	0.046 (>0.1)	Trend: More time spent on long-term planning P5 improved overall performance
C1 P6 Duration (observed)	ENG1 DFCP	0.467*	0.031 (>0.1)	Trend: More time spend in long-term planning P6 improved ENG1 management
C2 P1 Duration (corrected)	ENG2 DFCP	0.418*	0.049 (>0.1)	Trend: More attention to flight path stabilisation P1 improved ENG2 management
C2 P3 Duration (corrected)	Route DFCP	-0.551*	0.011 (>0.1)	Trend: More attention to short-term planning P3 worsened route management
C2 P5 Duration (corrected)	Route DFCP	0.451*	0.037 (>0.1)	Trend: More attention to appropriate actions P5 improved route management
C3 P3 Duration (%corrected)	Route DFCP	-0.626*	0.004 (>0.1)	Trend: Greater proportion of attention toward short-term planning P3 worsened route management
C3 P5 Duration (%corrected)	Route DFCP	0.426*	0.049 (>0.1)	Trend: Greater proportion of attention toward appropriate actions P5 improved route management
C4 SC1: Looping (desired)	Route DFCP	0.476*	0.028 (>0.1)	Trend: More looping of P4/5/6 improved route management
C5 SC1: Looping (%) (desired)	Route DFCP	0.551*	0.011 (>0.1)	Trend: More looping of P4/5/6 improved route management
C5 SC4: FFD 456 (%) (undesired)	Route DFCP	-0.476*	0.028 (>0.1)	Trend: Skipping ahead to P4/5/6 worsened route management
C6 TM Duration (observed)	Route DFCP	-0.426*	0.049 (>0.1)	Trend: More time spent on TM worsened route management
C7 TM Duration (%corrected)	Route DFCP	-0.501*	0.020 (>0.1)	Trend: Greater proportion of attention toward TM worsened route management
C7 CM Duration (%corrected)	Route DFCP	0.526*	0.015 (>0.1)	Trend: Greater proportion of attention toward CM improved route management
C10 PC3: Skip UM (%) (undesired)	Route DFCP	-0.426*	0.049 (>0.1)	Trend: Skipping UM worsened route management

* Uncorrected p is significant, trend ($p \leq 0.05$; $q > 0.1$) (two-tailed)

5 REFLECTION ON HYPOTHESES

The reflection is limited to the two hypotheses H_1 and H_2, at both the strategy phase and principle levels as these draw conclusions at different levels of abstraction. The implications of the findings for the operational and research domains are covered in Section 6.

5.1 Training effectiveness conclusions

Table 13 summarizes the response to hypothesis H_1, segregating key behavioural expectations.

Table 13. Overview of study conclusions for hypothesis H_1

Training in strategy/RIS will …	Conclusion
increase time & attention in Phases 4, 5 and 6.	**Reject.** The RIS group (also trained in the strategy) did show more activity in P5 and P6, but this was gone after correcting for multitasking. The RIS may have been causal to this shift. Furthermore, there are weak indications that the BSL may have spent more time in P4.
reduce time & attention in Phases 1, 2 and 3.	**Reject.** The training actually (slightly) increased time & attention in P2 for the STG group.
increase attention toward SC1 and SC2 sequences.	**Reject.** There are no indications that training increased SC1 and SC2.
reduce attention toward SC3 (mild effect), SC4 and SC5 sequences.	**Partial accept.** There are weak indications that untrained crews showed more SC3. Training effects for SC4 and SC5 cannot be confirmed.
increase time & attention in UM and CM.	**Reject.** There are no indications that training increased attention in UM and there are weak indications that the BSL may have spent more time in UM.
reduce time & attention in TM.	**Reject.** There are no indications that training reduced attention in TM.
increase attention towards PC1 sequences.	**Reject.** There are no indications that training increased PC1.
decrease attention towards PC2 and PC3 sequences.	**Reject.** There are no indications that training decreased PC2 and PC3.

The above findings indicate that the training was not effective to induce the behaviours that were hypothesized to have been more prevalent after training. There are many reasons for this, including the brevity of the training (only several hours of classroom and simulator familiarisation before the actual experiment flight). Another explanation may be that making crews consciously aware of their actions may result in them re-navigating certain phases and steps, which may be the reason for the STG group's increased time and attention for Phase 2. This was observed on some occasions, where crews would be taking actions and then, mid-way through their actions, remember they had a strategy to apply, which

caused them to retrace some of their steps. This would be attributable to insufficient time to truly internalize and familiarize with a new way of managing complex situations. Another reason for these results may lie in random effects due to the small sample sizes. If the STG and RIS groups were coincidentally provided crews that struggled more, and the BSL group featuring crews naturally proficient at managing complexity, then this would also limit the power of these between-group results. An experiment with the same group flying two flights – one before training and one after – may result in a better pairwise training effect study.

5.2 Behavior-performance conclusions

Table 14 summarizes the response to hypothesis H_2, segregating key expectations where behaviours improve performance.

Table 14. Overview of study conclusions for hypothesis H_2

Performance will increase with …	Conclusion
increased time & attention in Phases 4, 5 and 6.	**Partial accept.** There are trends that time and attention in P5 related positively to Route DFCP. Time in P6 was also related positively with Total and ENG1 DFCP scores.
reduced time & attention in Phases 1, 2 and 3.	**Mixed results.** Strong trends that time & attention in Phase 3 related negative with Route DFCP scores. At the same time, a trend indicates that attention in P1 relates positively with ENG2 DFCP scores.
increased attention toward SC1 and SC2 sequences.	**Partial accept.** Trend that SC1 related positively to Route DFCP scores.
reduced attention toward SC3 (mild effect), SC4 and SC5 sequences.	**Partial accept.** Trend that SC4 related negatively to Route DFCP scores.
increased time & attention in UM and CM.	**Partial accept.** Trends indicate relative attention to CM related positive to Route DFCP scores.
reduced time & attention in TM.	**Partial accept.** Strong trends indicate more time and attention in TM related negatively to Route DFCP scores.
increased attention towards PC1 sequences.	**Reject.** There are no indications that PC1 increases performance.
decreased attention towards PC2 and PC3 sequences.	**Partial accept.** There are indications that PC3 is negatively related to Route DFCP scores.

In contrast to the training effect results, there are more consistent relationships between behaviour and performance measures. It should be reiterated that all behaviour-performance correlations only provided trends (i.e., significant *p*-values but non-significant *q*-values). However, given relatively strong effect size (~ $\tau \pm 0.5$), confirmation of these relations may be possible with a stronger testing power (e.g., with more participants).

The observed trends do suggest that performance improves by not skipping phases (PC3) or going back (SC4), while also improving when focusing time and attention on later phases (Phases 4, 5 and 6) related to UM and CM, looping them (SC1) and spending less time in earlier phases related to TM. These results were particularly clear when considering performance measures related to managing future threats and contingencies (e.g., routing and energy management for this scenario). While not all behavioural measures were clearly correlated to performance, the strong correlations *between* the behavioural measures themselves may provide further indications about their effect on performance. For example, the strong *negative* correlation between PC1 and PC3 behavioural measures, and *negative* relation between PC3 and performance, may suggest a *positive* relation between PC1 and performance. Such proxy trends form specific hypotheses for future testing and require greater testing power to be confirmed.

In summary, the study was able to confirm that certain behaviours (brief TM, avoiding skipping/going back, UM and CM focus with looping) are possible performance drivers in the context of dealing with opacity and ambiguity on the flight deck. More powerful testing (possibly a meta-analysis) is required to confirm these trends beyond the False Discovery Rate (FDR). Furthermore, given the limited experimental data, the study concludes that the proposed treatment (strategy training approach, and RIS) cannot be proven effective in inducing desired behaviour. The indications that the RIS positively changed behaviour are stronger, as evident by more observed behaviour related contingency management in RIS crews (which was the purpose of the RIS). It may be that by illustrating more dimensions of risk, the RIS facilitates a broader threat assessment by the crew and identify better flight continuation options. Both improvements in the training, as well as improvements in the study design (increased power and group sizes) may yield more encouraging results in identifying what training or system developments may instil the desired behaviours.

6 DISCUSSION

Reflecting on the above study, there is some evidence that certain intended behaviours such as effective (i.e., brief) short-term stabilisation and planning to buy time, and subsequently using this time to managing uncertainties and contingencies, may increase crew performance in dealing with complex, ambiguous and opaque flight situations. This evidence supports the fundamental shift from an *executive* to a *learning* function or mindset that the crew must engage to maintain safety performance when faced with

opacity and ambiguity as in this study's scenario. This requires time and space for sensemaking, an explorative mindset ("managing uncertainties") and also a feedback process to reflect on the new knowledge gained through interacting with the situation and systems.

From an operational perspective, the results (pending further analysis) may drive a change in how flight crew training and cockpit JCSs can be reconfigured to be more effective in the context of a crew dealing with complexity, ambiguity and opacity which are becoming the Achilles' heel of modern airliners (as well as other complex systems in- and outside of aviation). While this study lacked the statistical power to confirm the maturity of the strategy and RIS in their current form, results are consistent and show positive effects which warrant revisiting them with a larger study population.

While the study's small population size limits its power to evaluate training and RIS effectiveness, qualitative observations indicate that the strategy, and how it was trained, may be subject to future improvements. While some STG/RIS crews easily integrated the new strategy into their workflow, most crews were observed recalling the strategy somewhere midway through the scenario, rather than engaging the strategy when the initial problems occurred (i.e., the ENG2 overheat or autothrust system failure). Longer training time (possibly months in advance) and more practice will likely improve strategy recall and application consistency. Improved training may also include more case studies to teach and explore a more dynamic response centred on the core principles of TM → UM → CM, rather than the rigidity that a quick reference card introduces. Despite the limited training, feedback from the STG group about the strategy does indicate that they deemed it as helpful, most notably by well-performing crews (based on Total DFCP scores).

Reflecting on the experimental setup and approach, this study could have benefited from a less complicated experimental logistics which undoubtedly contributed to data noise (i.e., two different locations, two different research simulators, two aircraft classes, two different airline participant pools, two research teams and two different experiment executions with and without RIS). Although great attention has been given to provide equivalent scenarios from a challenge-centric point of view (Niedermaier et al., 2018) and to align experiment execution, debriefing and scoring mechanisms, some noise in the data will be attributable to this somewhat divided setup. Reproducing the study with a more homogenous sample size and setup may reduce random effects and improve statistical power. This may

also be done by means of a longitudinal study design evaluating crew before and after introducing the strategy and RIS to reduce group effects and correct for differences in crew natural abilities. Finally, future (meta-studies) may also investigate improved behavioural measures (e.g., compound measures that cluster desired and undesired patterns) as well as new measures such as looking at the temporal distribution of phase activities and characterisations (not just the runtime sums). Despite these limitations, the analysis presented in this study aims to catalyse industry insight into the minutiae of flight crew sensemaking in the face of opacity and ambiguity on the flight deck, where both the results as well as innovative behavioural mapping methods invite new ways of looking—a new perspective—toward flight crew performance in modern transport aircraft.

DISCLAIMER *The data and results were produced as part of the FP7 2012 Aeronautics and Air Transport programme under EC contract ACP2-GA-2012-314765-Man4Gen. The NLR and DLR are owners of the data and results of this research and are published with permission from both NLR and DLR. The views and opinions expressed in this publication are those of the authors and are not intended to represent the position and opinions of the Man4Gen consortium and/or any of the individual partner organizations.*

References

Abbott, K. H. (2017). Human factors engineering and flight deck design. *Digital avionics handbook*, 302-328.

Banks, V. A., Plant, K. L., & Stanton, N. A. (2020). Leaps and shunts: designing pilot decision aids on the flight deck using Rasmussen's ladder. *Contemporary ergonomics and human factors*.

Boy, G. A. (2020). Aerospace human systems integration: Evolution over the last 40 years. *A framework of human systems engineering: Applications and case studies*, 113-128.

Buch, J. P., Niedermaier, D., & Stepniczka, I. (2017). Managing the Unexpected-Human-in-the-Loop Simulation as Effective Tool for the Assessment of the Risk Information System in an Operationally Relevant Context. In *AIAA Modelling and Simulation Technologies Conference* (p. 4155).

Clark, A. N., & Wilson, A. L. (2024). From Levers to Glass: The Evolution of Aircraft Flight decks and the Future of Aviation Innovation.

DeSalvo, P., Fogarty, D. (2016). DOT/FAA/TC-16/39 Safety Issues and Shortcomings with Requirements Definition, Validation and Verification Processes – Final Report.

EASA (2018). Startle Effect Management Final Report.

Field, J., Fucke, L., Correia Grácio, B., & Mohrmann, J.F.W. (2016). Flight crew response to unexpected events: a simulator experiment. In *AIAA Modelling and Simulation Technologies Conference* (p. 3373).

Field, J., Rankin, A., Mohrmann, J.F.W., Boland, E., & Woltjer, R. (2017). Flexible procedures to deal with complex unexpected events in the cockpit. In *Resilience Engineering Association Symposium Liége, Belgum, 26-29 June, 2017*. Resilience Engineering Association.

Gago, C. P., Hansman, R. J., Edmondson, M. K., & Mosqueda, M. A. (2025, September). A Study of Pilot Response to System Failure in Transport Category Aircraft from 2000 to 2024. In *2025 AIAA DATC/IEEE 44th Digital Avionics Systems Conference (DASC)* (pp. 1-8). IEEE.

Helmreich, R. L., Klinect, J. R., & Wilhelm, J. A. (1999, May). Models of threat, error, and CRM in flight operations. In *Proceedings of the tenth international symposium on aviation psychology* (Vol. 10, pp. 677-682).

Hollnagel, E., & Woods, D. D. (2005). Joint cognitive systems: Foundations of cognitive systems engineering. CRC press.

Hosmer, D. W., & Lemeshow, S. (2000). Applied logistic regression (2nd ed.). John Wiley & Son

International Air Transport Association (IATA) (2014). Safety Report 2014.

Kelly, D., & Efthymiou, M. (2019). An analysis of human factors in fifty controlled flight into terrain aviation accidents from 2007 to 2017. *Journal of Safety Research*, *69*, 155–165.

Kharoufah, H., Murray, J., Baxter, G., & 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.

Landman, A., van Middelaar, S. H., Groen, E. L., van Paassen, M. M., Bronkhorst, A. W., & Mulder, M. (2020). The effectiveness of a mnemonic-type startle and surprise management procedure for pilots. The International Journal of Aerospace Psychology, 30(3-4), 104-118.

Landry, S.J. (2009). Flight Deck Automation. In: *Nof, S. (eds) Springer Handbook of Automation.* Springer Handbooks. Springer, Berlin, Heidelberg.

Loukopoulos, L.D., Dismukes, R.K., & Barshi, I. (2009). The Multitasking Myth: Handling Complexity in Real-World Operations (1st ed.). Routledge.

Man4Gen Consortium (2015). 6.7 - Final report of research methods and results.

Mohrmann, J.F.W., Lemmers, A., & Stoop, J. (2015). Investigating flight crew recovery capabilities regarding system failures in highly automated fourth generation aircraft. Aviation Psychology and Applied Human Factors.

Niedermaier, D., Buch, J. P., Mohrmann, J.F.W., & Durak, U. (2018). Simulating the Unexpected: Challenge-Centric Simulator Scenario Design for Advanced Flight Crew Training. In *2018 AIAA Modelling and Simulation Technologies Conference* (p. 1397).

Prinzel, L. , Krois, P., Ellis, K., Vincent, M., Stephens, C., Oza, N., Chancey, E., Davies, M., Mah, R., Ackerson, J., Infeld, S., Kiggins, D., & Matthews, B. (2024). *The Adaptable and Resilient Safety System: The Human Factor in Future In-Time Aviation Safety Management Systems*.

Proctor, R. W., & van Zandt, T. (2018). *Human Factors in Simple and Complex Systems, Third Edition*.

Rankin, A., Ekström, E., Sjölin, V., Woltjer, R., Stepniczka, I., Eder, M. (2016). Final report of research methods and results.

Sarter, N. B., & Woods, D. D. (1994). Pilot Interaction With Cockpit Automation 11: An Experimental Study of Pilots' Model and Awareness of the Flight Management System. In THE INTERNATIONAL JOURNAL OF AVIATION PSYCHOLOGY (Vol. 4, Issue 1).

Saurin, T. A., & Carim Junior, G. C. (2012). A framework for identifying and analyzing sources of resilience and brittleness: A case study of two air taxi carriers. *International Journal of Industrial Ergonomics*, *42*(3), 312–324.

Senate Committee (2020). US Senate Committee Investigation Report – Aviation Safety Oversight.

Stoop, J. A., & van Kleef, E. A. (2015). Reliable or Resilient: Recovery from the Unanticipated. In *International Journal of Performability Engineering* (Vol. 11, Issue 2).

Strauch, B. (2017). *Investigating Human Error*. CRC Press.

Woltjer, R., Field, J., & Rankin, A. (2015). Adapting to the unexpected in the flight deck. In *Proceedings of the 6th resilience engineering association symposium. Lisbon, Portugal: REA*.

Appendix A: Man4Gen Quick Reference Card

This Man4Gen quick reference card describes the six basic phases. It is reduced in font size to fit in this appendix but is usually printed larger for readability in the cockpit. Strategy-only crews receive a card without the display reference boxes in the bottom right of each phase.

COMPLEX SITUATION MANAGEMENT GUIDE

STABILIZE FLIGHT PATH

- **FLY THE AIRCRAFT**
- **CONFIRM FLIGHT PATH CONTROL (OR IF SEMI-STABLE)**
- **CONSIDER USE OF AUTOFLIGHT**
- **ASSIGN PF/PM**

NO PAGES

IMMEDIATE THREATS

- **IDENTIFY IMMEDIATE THREATS**
- **PRIORITIZE THREATS**
- **THREAT MEMORY ITEMS**
- **CONFIRM THREAT STATUS**

NO PAGES

SHORT TERM PLAN

- **IF POSSIBLE, MAINTAIN FLIGHT AND BUY TIME**
- **CONSIDER FLIGHT PLAN OPTIONS**
  Original destination - alternate destination – holding – land ASAP
- **CONFIRM SHORT TERM FLIGHT PLAN**
- **CONSIDER NOTIFYING ATC/CABIN/COMPANY**

RISK LEVEL PAGE (0)
FLIGHT PHASE RISK PAGE

IDENTIFY SITUATION

- **ACKNOWLEDGE CERTAINTIES, UNCERTAINTIES AND CONCERNS**
  Issues to consider: fuel or time limit? Structure integrity? Controllability/performance? Information reliability? Secondary failures? External complication factors such as weather, traffic and routing/destination?
- **CROSS-CHECK SUSPECTED SITUATION**
- **IF SEVERAL POSSIBILITIES: PREPARE FOR WORST CASE**

CATEGORY RISK PAGES (1-4)

PERFORM APPROPRIATE ACTIONS

- **ASSURE THAT ACTIONS WILL BE SAFE, EFFECTIVE AND INFORMATIVE**
- **PERFORM PROCEDURES & OTHER ACTIONS**
- **VERIFY EFFECT OR KNOWLEDGE GAINED FROM ACTIONS**

CATEGORY RISK PAGES (1-4)

LONG TERM PLAN

- **IDENTIFY EFFECTS OF THE SITUATION**
  Issues to consider: Time/endurance limits? Struct./system functionality? Controllability? Performance? Info. reliability?
- **ADAPT MONITORING TO DETECT IMPORTANT CHANGES**
- **CONSIDER REGULAR FLIGHT PLANNING ASPECTS (Wx, COMPANY, PAX, DEFERRED TASKS)**
- **CONFIRM LONG TERM PLAN**
- **CONSIDER NOTIFYING ATC/CABIN/COMPANY**

RISK LEVEL PAGE (0)
FLIGHT PHASE RISK PAGE

Appendix B: Man4Gen Quick Reference Card (Adapted from Buch et al., 2017)

All risk information pages have been designed to mimic the visual layouts and screen sizes of the Airbus A320 ECAM system pages. This particular page shows a risk calculation for various flight phases. Calculations should take into account many different aspects of the aircraft state including current altitude/airspeed, location, malfunctions, endurance and other factors such as weather and geographic risks. For the sake of an experiment, the pages used in the study presented a select number of "drawn" pages that would change based on the scenario time and if particular (failure) events occurred. In this particular page, it is clear that the Landing/Go-Around flight phase presents the most risk and includes two pieces of information that have the greatest contribution to this risk ("Reduced Climb Rate" and "Yaw Oscillations"). This should drive crews to mitigate with appropriate threat and error management briefings and strategies.

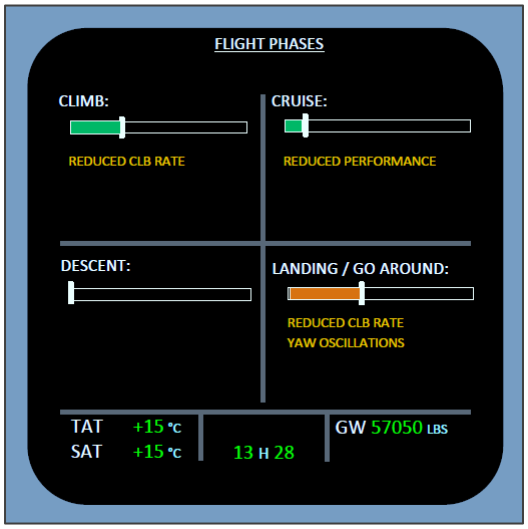


Figure B.1. RIS flight phase page

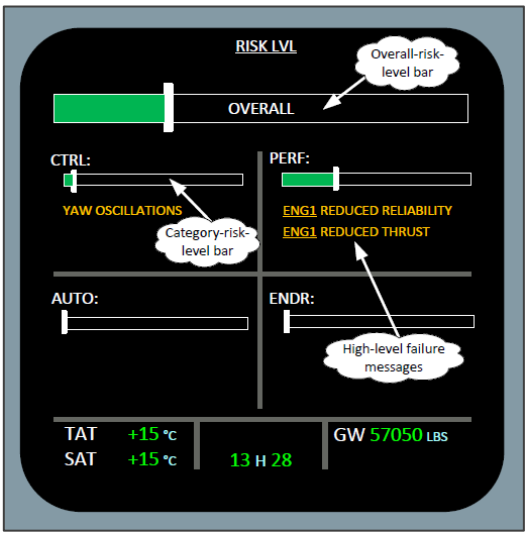


Figure B.2. RIS overview page

In this page, the RIS provides crews with another "cross-section" of risks, namely those associated to different aircraft state parameters: controllability, performance, automation and endurance. In this example, *aircraft performance* is most affected, with two failures driving this risk assessment ("ENG1 reduced reliability" and "reduced thrust"). This page also provides an "overall" score, which supports the other categories in driving crew awareness to mitigate and brief accordingly.

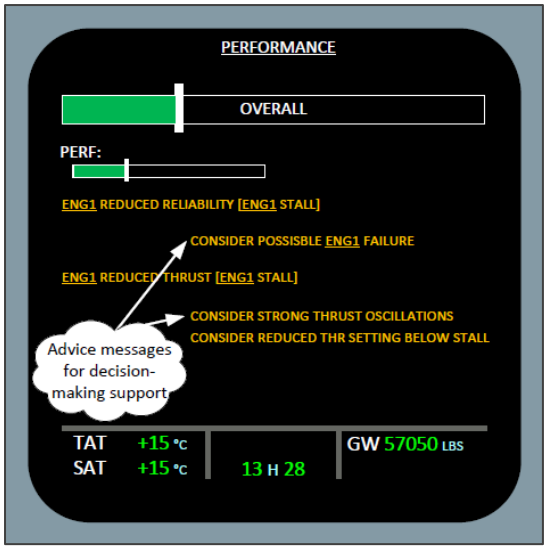
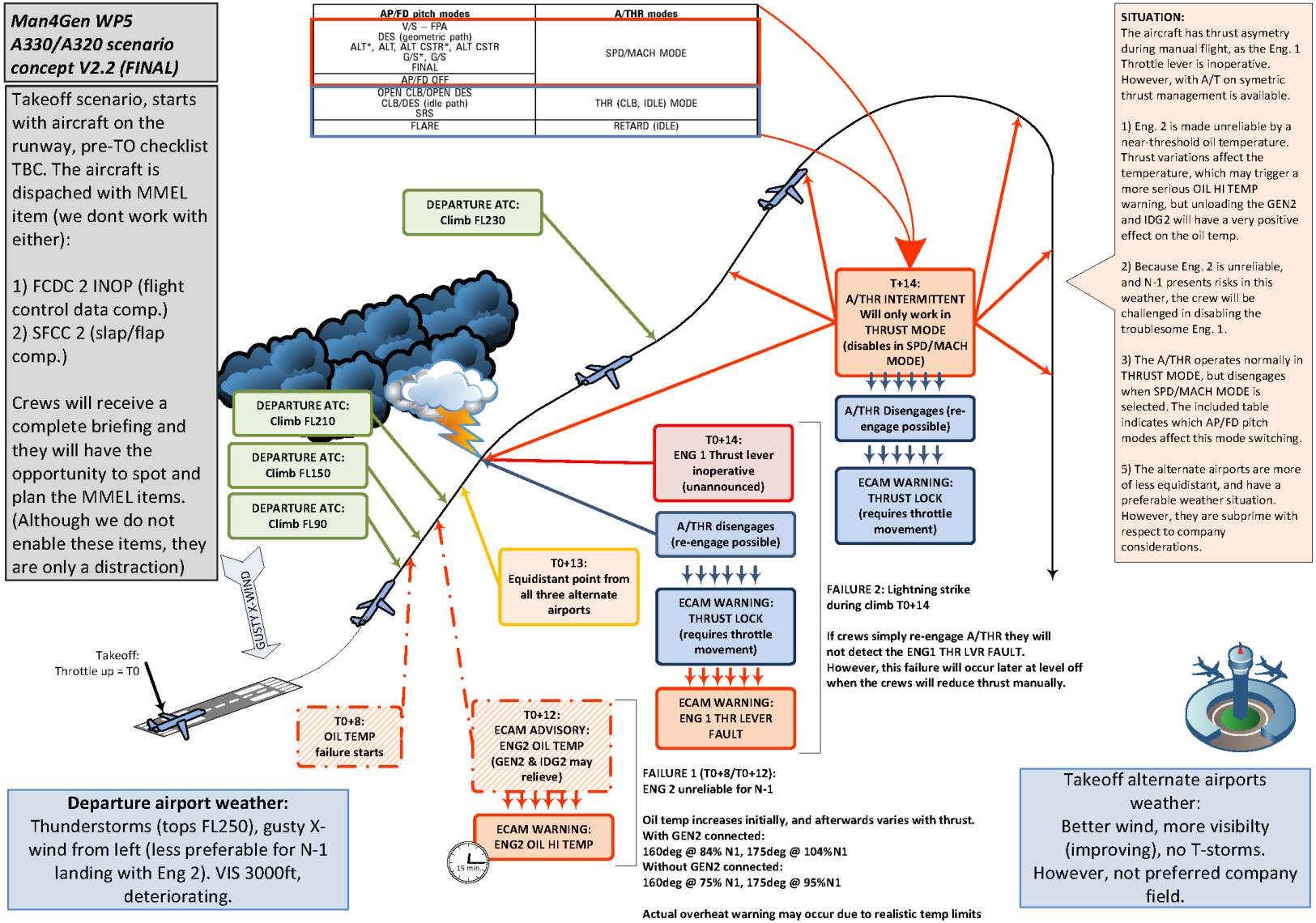


Figure B.3. RIS detail page

Figure B.4. RIS in DLR AVES A320 simulator

Each risk (in this case *aircraft performance*) can be selected to inspect more details about the systems considerations and how particular failures are driving its risk evaluation. These detailed pages also include *suggested* actions or considerations to support the crew's threat and error management. However, the verbatim is carefully selected as to not stimulate automation bias and prescription-based resolutions. Figure 4 illustrates how the RIS is embedded in the natural cockpit environment, in order to minimize the impact on crew training and make use of human-machine interactions already familiar to the crews (e.g., ECAM system page selection).

Appendix C: Scenario schematic



Appendix D: DFCP items

DFCP Item	ENG2 Management	ENG1 Management	Route/situation management	Communication/Interaction	A/THR Management	Other flight tasks
Avoiding storm cells in path			X			
OT Advisory						
Verbalize oil temp rising	X			X		
Refer to QRH ADV page	X					
Reduce Thrust	X					
GEN2 OFF	X					
IDG 2 OFF	X					
Collect information of weather at nearby fields			X			
Discuss weather situation before diversion decision. FRA/AMS deteriorating (& enroute to Saarbrucken/Zweibrucken).			X	X		
Do not land at AMS/FRA			X			
OT ECAM Warning						
ENG2 IDLE before shutting down	X					
APU GEN ON	X	X				
APU ON Before ENG2 shutdown	X	X				
Use ENG 2 within oil temp limits	X					
N-1 descent and approach planned			X			
Identify ENG1 throttle issue & effects		X				
THR LVR FAULT & A/THR FAILURE						
Own initiative to find FCOM		X				
Call maintenance/FT/Dispatch own initiative				X		
Declare MAYDAY/PANPAN to ATC				X		
Identify THRUST LOCK		X				
Keep/restart ENG2	X					
No unexpected ENG1 IDLE (gear/slats)		X				
Respect gear/slats extension speeds			X			
Control ENG1 using gear/slats extensions		X				
Continue using ENG1 for approach/landing	X					
Identify A/THR as a problem/unreliable					X	
A/P ON after LS						X
Inform cabin crew of situation				X		
Landing prep						
Check impact of failures on landing performance (LPC-NG / QRH)			X			
Discuss GA performance risk: ENG2 overheat, clean up for ENG1, Flap 3 landing			X	X		
request equipment on runway/engines off				X		

Appendix E: DFCP and strategy phase behavioural measures data

Table 15. DFCP, strategy phase duration (observed) and strategy phase duration (attention corrected) scores

Group	Crew ID	Total DFCP	ENG2 Management	ENG1 Management	Route Management	Communication	P1 Duration (observed)	P2 Duration (observed)	P3 Duration (observed)	P4 Duration (observed)	P5 Duration (observed)	P6 Duration (observed)	P1 Duration (attention corrected)	P2 Duration (attention corrected)	P3 Duration (attention corrected)	P4 Duration (attention corrected)	P5 Duration (attention corrected)	P6 Duration (attention corrected)
BSL	102	67%	45%	71%	88%	83%	00:03:07	00:03:00	00:15:43	00:33:09	00:40:20	00:33:55	00:00:59	00:01:09	00:05:27	00:15:58	00:20:23	00:15:54
	103	57%	36%	57%	50%	67%	00:06:00	00:00:00	00:19:50	00:24:37	00:29:00	00:22:15	00:02:05	00:00:00	00:08:39	00:14:53	00:16:46	00:15:23
	202	70%	55%	71%	88%	67%	00:08:59	00:03:20	00:15:13	00:24:47	00:33:00	00:13:25	00:02:41	00:01:10	00:10:10	00:11:44	00:17:00	00:05:05
STG	101	73%	73%	57%	100%	67%	00:04:17	00:03:54	00:10:33	00:19:11	00:30:10	00:15:31	00:01:23	00:01:13	00:03:50	00:09:01	00:17:45	00:09:17
	104	70%	64%	86%	63%	67%	00:23:47	00:06:00	00:23:12	00:15:01	00:24:20	00:29:08	00:09:02	00:01:58	00:12:37	00:06:13	00:11:16	00:16:11
	106	77%	55%	86%	100%	33%	00:10:00	00:04:10	00:06:06	00:12:45	00:26:34	00:15:40	00:04:09	00:01:27	00:02:22	00:08:44	00:15:45	00:11:07
	107	63%	55%	86%	38%	67%	00:12:06	00:06:07	00:26:23	00:16:01	00:13:37	00:09:04	00:05:48	00:01:57	00:13:31	00:09:43	00:08:34	00:04:24
	108	67%	36%	71%	88%	67%	00:03:33	00:06:11	00:12:10	00:29:53	00:43:56	00:23:19	00:00:47	00:01:49	00:06:59	00:18:19	00:28:05	00:14:48
	109	43%	45%	14%	50%	67%	00:10:54	00:06:25	00:19:44	00:14:59	00:26:03	00:09:00	00:03:19	00:01:50	00:07:34	00:05:15	00:13:46	00:03:09
	110	60%	45%	71%	50%	33%	00:03:00	00:06:00	00:25:46	00:20:54	00:21:38	00:14:45	00:00:50	00:02:01	00:11:47	00:07:54	00:08:25	00:06:45
RIS	206	67%	36%	86%	88%	67%	00:03:00	00:03:00	00:10:08	00:18:43	00:34:35	00:34:15	00:01:09	00:00:48	00:03:22	00:06:28	00:16:31	00:16:09
	207	67%	45%	86%	88%	50%	00:13:49	00:03:00	00:13:56	00:18:09	00:49:45	00:31:08	00:04:31	00:00:38	00:04:48	00:05:45	00:24:40	00:13:31
	208	80%	64%	100%	88%	50%	00:11:22	00:03:20	00:19:26	00:32:27	00:46:12	00:34:46	00:03:32	00:01:04	00:09:12	00:13:30	00:20:18	00:15:06
	209	83%	73%	86%	100%	67%	00:13:35	00:03:21	00:19:46	00:32:45	00:46:28	00:36:01	00:04:29	00:00:50	00:07:59	00:13:56	00:21:43	00:15:23

Table 16. SC (total, normalized) and SC (count, normalized) scores

Group	Crew ID	SC1: Looping (total, normalized)	SC2: Sequence (total, normalized)	SC3: Skip 3 or 5 (total, normalized)	SC4: Back to 123 (total, normalized)	SC5: FFD 456 (total, normalized)	SC1: Looping (count, normalized)	SC2: Sequence (count, normalized)	SC3: Skip 3 or 5 (count, normalized)	SC4: Back to 123 (conut, normalized)	SC5: FFD 456 (count, normalized)
BSL	102	300%	123%	129%	21%	36%	8.67	4.60	4.50	1.00	2.00
	103	203%	125%	167%	34%	38%	5.00	2.80	5.50	1.42	1.38
	202	100%	97%	117%	26%	26%	3.33	3.00	4.00	1.25	0.75
STG	101	128%	73%	100%	26%	31%	3.67	2.40	3.00	1.17	1.25
	104	76%	74%	74%	41%	64%	4.33	3.40	4.50	2.33	2.88
	106	172%	65%	129%	9%	22%	4.00	1.60	4.00	0.83	0.63
	107	39%	118%	150%	63%	31%	1.00	3.00	2.00	2.25	0.88
	108	244%	250%	29%	23%	27%	5.67	5.60	1.00	1.58	1.13
	109	67%	49%	84%	23%	47%	2.33	4.40	4.50	1.83	3.00
	110	58%	82%	121%	27%	24%	4.00	4.40	4.50	1.67	2.13
RIS	206	72%	79%	23%	19%	36%	3.67	3.60	2.00	1.08	2.00
	207	94%	79%	69%	17%	39%	3.67	4.00	2.50	1.33	2.88
	208	160%	94%	79%	23%	45%	8.00	4.40	2.50	1.33	2.00
	209	175%	102%	117%	28%	41%	6.33	4.80	6.00	1.42	2.50

Appendix F: Philosophy principles behavioural measures data

Table 17. Principle duration (observed), principle duration (attention corrected), PC (total, normalized) and PC (count, normalized) scores

Group	Crew ID	TM Duration (observed)	UM Duration (observed)	CM Duration (observed)	TM Duration (attention corrected)	UM Duration (attention corrected)	CM Duration (attention corrected)	PC1: Sequence and loop (total, normalized)	PC2: Return to TM (total, normalized)	PC3: Skip UM (total, normalized)	PC1: Sequence and loop (count, normalized)	PC2: Return to TM (count, normalized)	PC3: Skip UM (count, normalized)
BSL	102	00:19:00	00:33:09	00:53:21	00:07:43	00:17:17	00:34:51	196%	39%	50%	6.50	2.67	2.50
	103	00:20:15	00:24:37	00:46:40	00:10:05	00:15:10	00:32:32	183%	94%	92%	5.25	3.33	2.00
	202	00:25:00	00:24:47	00:15:31	00:14:28	00:12:00	00:23:24	123%	67%	113%	4.75	3.33	3.50
STG	101	00:12:29	00:19:11	00:39:47	00:05:37	00:09:44	00:27:07	98%	50%	96%	3.75	1.67	2.50
	104	00:37:04	00:15:01	00:42:16	00:22:57	00:07:25	00:26:55	71%	117%	133%	3.25	3.67	3.50
	106	00:11:45	00:12:45	00:37:25	00:07:10	00:08:58	00:29:03	100%	39%	42%	3.00	1.00	1.50
	107	00:30:40	00:16:01	00:22:37	00:20:11	00:09:50	00:13:55	104%	111%	67%	3.00	3.67	2.00
	108	00:13:28	00:29:53	00:55:28	00:08:30	00:19:00	00:43:16	283%	56%	25%	6.25	2.67	1.50
	109	00:24:11	00:14:59	00:28:25	00:11:21	00:06:18	00:17:14	42%	61%	117%	3.25	3.33	3.50
	110	00:26:27	00:20:54	00:29:10	00:13:37	00:08:51	00:15:14	102%	50%	38%	5.00	2.67	3.00
RIS	206	00:14:16	00:18:43	00:00:00	00:05:35	00:07:49	00:15:56	58%	28%	67%	2.00	2.33	2.00
	207	00:23:53	00:18:09	00:14:48	00:10:13	00:06:57	00:30:06	69%	44%	113%	3.75	3.33	4.50
	208	00:29:17	00:32:27	00:14:48	00:15:10	00:15:06	00:23:41	113%	78%	100%	5.25	2.67	4.00
	209	00:27:00	00:32:45	00:38:43	00:12:46	00:16:27	00:32:50	165%	72%	79%	6.75	2.00	3.00

Appendix G: Characterisation modalities analysis results

Characteristics measures modality analyses for both strategy phase level and principle level measures.

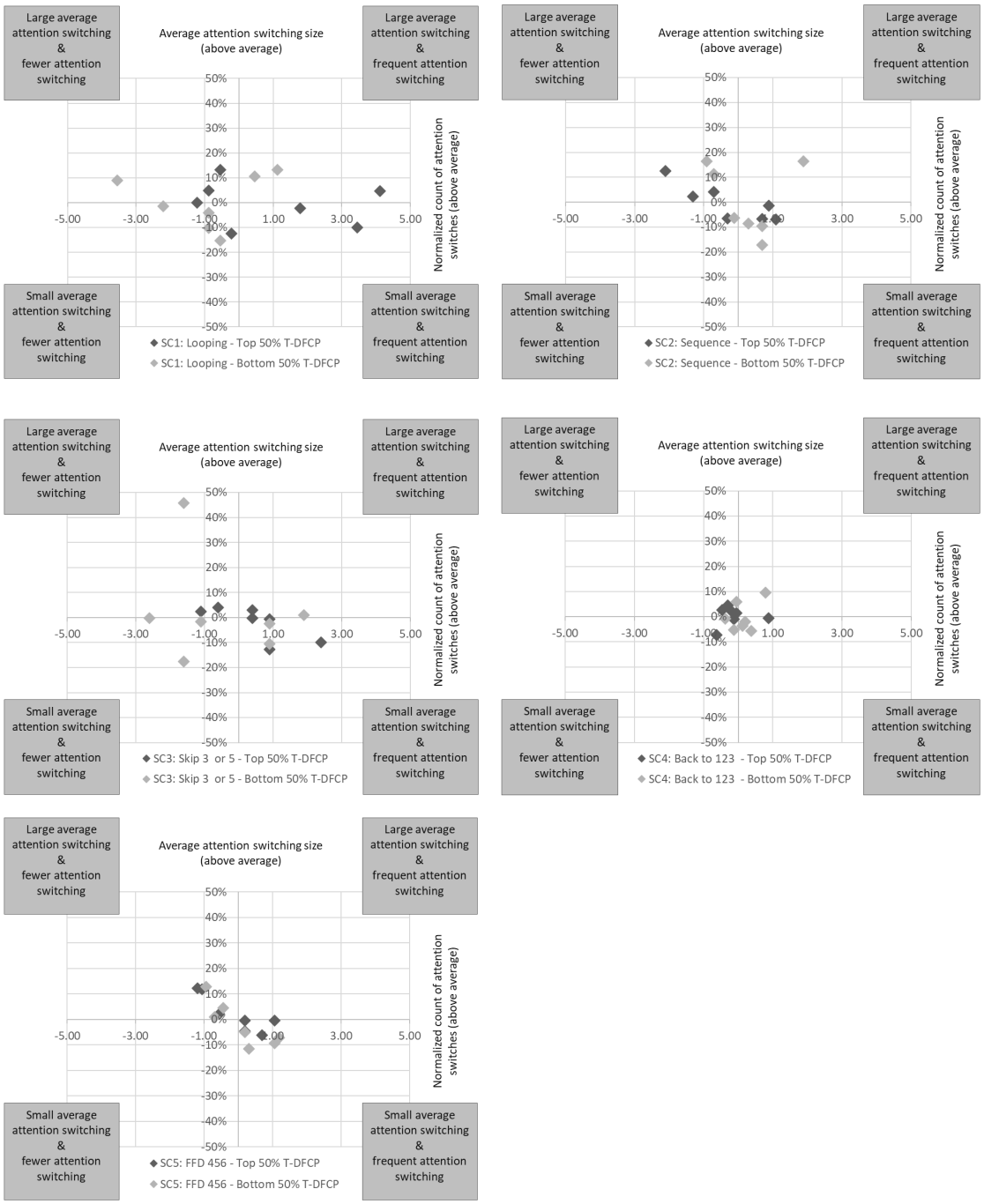


Figure 19. Comparison of modalities for all five strategy phase characterisations indicating top and bottom half of performers (according to Total DFCP scores)

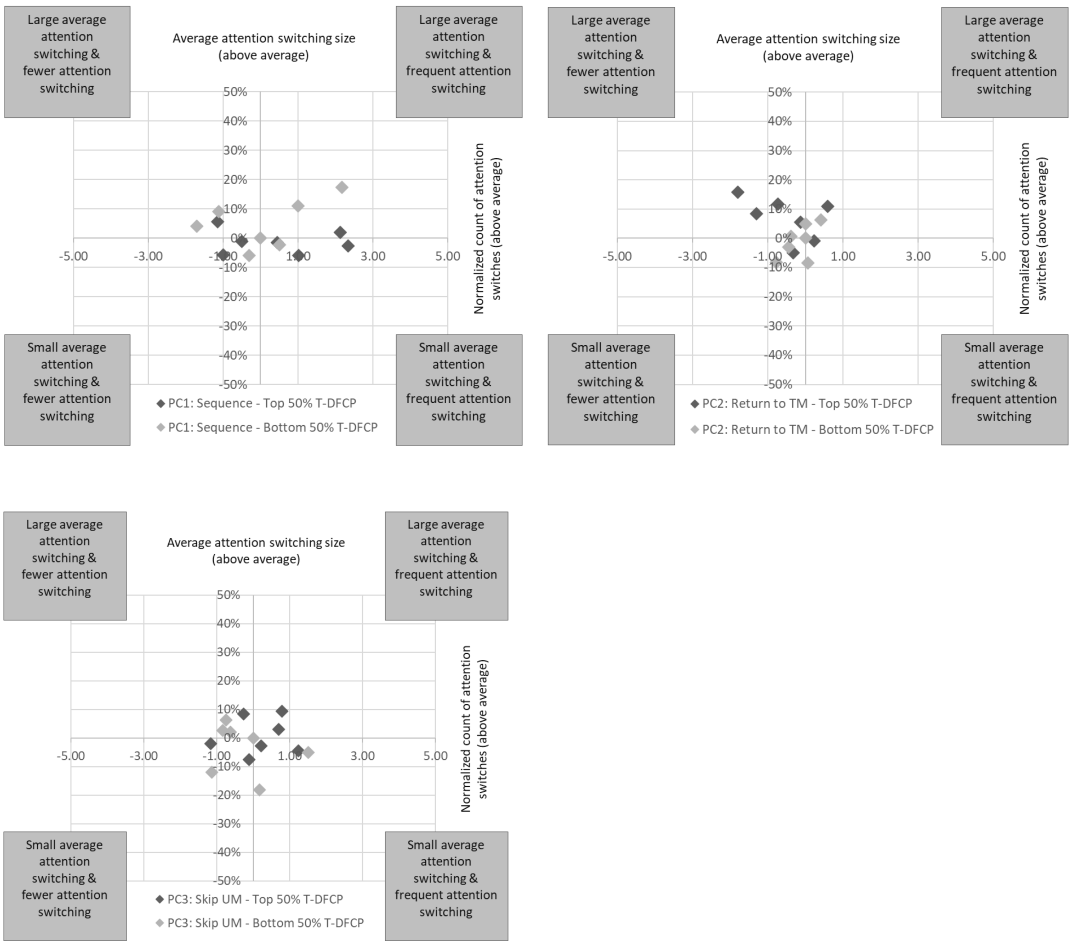


Figure 20. Comparison of modalities for all three principle characterisations indicating top and bottom half of performers (according to Total DFCP scores)