Pilot Performance Assessment in Simulators: Exploring Alternative Assessment Methods

Pete McCarthy[1(✉)] and Arnar Agnarsson[2]

[1] Cranfield University, Bedford, UK
pete.mccarthy@cranfield.ac.uk
[2] Iceland Air, Reykjavik, Iceland
aja.crew@icelandair.is

Abstract. Flight crew performance and competency assessment are daunting tasks requiring expertise and training, and still will not be possible without a certain degree of subjectivity. On the other hand, collecting reliable data on flight crew competencies is at the core of Evidence-based Training, a major modernization of training methodology the industry as a whole has embarked on. Data from assessment informs training departments about where pilots seem to be lacking in proficiency, so those issues can be addressed in initial and recurrent training programmes of airlines. The effectiveness of the training hinges on the quality of the data. Accurate interpretation of the data is crucial for the decisions on training to respond to the true needs of commercial pilots. The industry has made a great effort to develop ways to measure crew performance. The role of Human Factors in incidents and accidents has been known for a long time, and the need to assess and train Human Factors has been identified. In recent years, with the introduction of Evidence Based Training (EBT) there has been a shift in focus from task-based assessment to competence based assessment. This study analysed crew performance in 25 videos from simulator sessions in a high fidelity full flight simulator. A checklist of Desired Flight Crew Performance (DFCP) was used to distinguish between high and low performing crews. Then the performance of selected crews was analysed in detail, using Performance Indicators (PI) as developed in EBT. The findings suggest that while the DFCP method was useful for the classification of high and low performing crews, the PI method provided detailed information for the understanding of underlying factors that affected the performance of the crews. The study also considers the value of using PI to understand and emulate well executed flying and problem solving, to change the focus of training from the study of error and accidents, to training best practices and safe operation.

Abbreviations

1. APK - Application of Procedure
2. CAA - Civil Aviation Authority
3. CAP - Civil Aviation Publication
4. CFIT - Controlled Flight Into Terrain
5. COM - Communication
6. CRM - Crew Resource Management
7. DFCP - Desired Flight Crew Performance
8. EASA - European Aviation Safety Agency

D. Harris (Ed.): EPCE 2018, LNAI 10906, pp. 615–636, 2018.
https://doi.org/10.1007/978-3-319-91122-9_50

9. EBT - Evidence Based Training
10. FPA - Flight Path Management Automation
11. FPM - Flight Path Management Manual
12. FSTD - Flight Simulator Training Device
13. G/A - Go-Around
14. ICAO - International Civil Aviation Organisation
15. IRR - Inter-Rater Reliability
16. JAA - Joint Aviation Authority
17. KNO - Knowledge
18. KSA - Knowledge, Skill & Attitude
19. LOFT - Line Orientated Flight Training
20. LTW - Leadership and Teamwork
21. NOTECHS - Non-technical skills
22. PI - Performance Indicator
23. PIC - Pilot in Command
24. PSD - Problem Solving and Decision Making
25. SAW - Situational Awareness
26. SOP - Standard Operating Procedures
27. WLM - Workload Management
28. NTSB - National Transportation Safety Board

1 Introduction

In general, commercial air transport has become one of the safest modes of transportation in the last few decades. There have been advances in technology, which have reduced the accident rates to historically low levels. With this new, advanced and reliable technology, there have been some challenging issues regarding pilot training, and especially how pilots respond to unexpected events. In this study, a critical look is taken at two different methods of assessing commercial pilots in aviation. While traditional methods, based on observations performed by highly experienced and well trained flight instructors or examiners, can provide reliable data, the expertise of evaluators might vary in practice, which can cause variability of the results. This may become an issue for a data driven concept such as Evidence-based Training. What pilots do on the job is commonly expressed in technical skills, and human factors, which are commonly named non-technical skills. It can be challenging to assess pilot performance as it is difficult to isolate one from another. There are different assessment methods in place which are looked at in this paper, and there is empirical research which has been conducted that are considered as well.

The aviation industry has understood the importance of simulating the real-world experience through virtual environment. Simulation is used in many professions, and extensively for pilot training. For the training to be effective there is a need for a methodology that provides systematic and structured learning experiences. The effectiveness of this training is dependent on the quality of performance measurement practices in place. Performance measurement during each session must be diagnosed;

that is, the causes of effective and ineffective performance must be determined. This diagnostic measurement drives the systematic decisions concerning corrective feedback and remediation (Rosen et al. 2008). There are always challenges when it comes to assessing pilots during recurrent training. In Europe, every pilot must be assessed every 6 months based on EASA regulation. The regulation states that every pilot must be assessed in technical and non-technical skills.

2 Background

2.1 Simulator

In fast moving and complicating domains such as the military, medicine, business, and aviation, the workplace is characterized by high degrees of complexity and competitiveness. The need to maximize human performance is essential for safety, effectiveness, efficiency, and even survival. Human performance can be resilient, adaptive, and flexible in ambiguous and information-intensive contexts, but it can also be plagued with error and inefficiencies. Preparing people for performance in complex environments requires a complex approach to training (Salas et al. 2008). The simulator training in aviation has evolved somewhat throughout the last decades but in general, regulations that control the training have been very tenacious. In the early 1950's with the introduction of the jet age, there was a training scheme that was largely based on the evidence of hull loss from early generation jets so in a simple way the training developed into a system where every new accident triggered the introduction of a new task into already saturated training programmes in response to the accident. This could easily lead to an attitude where training was all about *ticking the boxes*, instead of responding to the real needs of pilots. Aircraft design and reliability has improved substantially over time, and at the same time the industry experienced accidents with hull losses where there was no malfunction of the onboard equipment. This gradually led to a shift in focus towards the human factor, or human failure. The human factor has been researched very extensively in the past three decades. According to researches, human failure has contributed to more than 2/3 of all accidents (Helmreich et al. 1999). A good example of that is the CFIT (Controlled Flight into Terrain), resulting in hull loss where inadequate situational awareness is almost always a contributing factor (ICAO 2013). In the late 1970's Crew Resource Management (CRM) started to be developed. The beginning of CRM is normally traced back to a workshop by NASA in 1979 (Helmreich et al. 1999). The CRM was intended to address Human Error. CRM was mainly classroom based, and later implemented into simulator programs like LOFT (Line Oriented Flight Training) etc. In the late 1990's the JAA precursor of EASA came up with Non-technical skill evaluation or NOTECHS to evaluate non-technical skills in simulator.

2.2 Simulator Assessment

Under EASA authority there is a requirement for technical and non-technical skills assessment of flight crew. EASA regulation stipulates the requirement for non-technical

assessment in AMC1 ORO.FC.115 Crew resource management (CRM) training "Assessment of CRM skills is the process of observing, recording, interpreting and debriefing crews and crew member's performance using an accepted methodology in the context of the overall performance." (EU 2012). EASA stipulates also that the method must be accepted by the national authority. Most operators are using the NOTECHS system or a similar system to fulfil this requirement. The NOTECHS system was issued in late 90's and recommended by JAR.OPS (precursor of EASA). It was recognised that the task to assess non-technical skills was more subjective than assessing technical facts. The Notechs systems was intended to minimise that factor with using behavioural indicators to assist the raters (Flin et al. 2003).

There have been researches in the past regarding assessment in simulation based training. It is known in the military, medicine, business, and aviation world. In the commercial aviation world, there is an extra challenge, as the assessment is done both on the team and on individual performance. Aviation is a workplace characterized by a high degree of complexity and competitiveness. Therefore, maximizing human performance is essential for safety and effectiveness. The effectiveness is dependent on the quality of performance measurement practices in place. The effective and ineffective performance must be determined.

It is very important that performance is explicitly defined and measured. Otherwise it is impossible to change, or improve it systematically. This measurement gets more complicated with more complexity. Human performance is essentially behaviour in completing a task. The problem is to realise what behaviours are important components of performance (Salas et al. 2008).

The simulator assessment has evolved in the last two decades. In the beginning, NOTECHS was designed as a professional tool to be used by non-psychologists. It was not intended to judge flight crew personality or a toll for introducing psychological jargon (Flin et al. 2003). Examiners were not to fail pilots only on basis of non-technical skills, except associated with technical skills. This created some confusion and let the CRM be a kind of a stand-alone subject with its own requirements. The EBT competencies are a mixture of technical and non-technical skills and have their own performance indicators to assist examiners in assessing and debriefing with the intention of promoting learning. According to ICAO Doc 9995 there are 8 competencies and 59 performance indicators. One of the challenges is the methodology to assess pilots with 59 performance indicators. The data gathered from the system is the important bit as it gives a good picture for the operators to see where they should put in an effort, and how they should channel the resources in their training. But even the highly trained and most experienced examiners are limited in what they can reliably measure.

2.3 Evidence Based Training (EBT) Versus Traditional Training

There is a shift in training with the new method implemented by ICAO doc 9995 that introduces EBT training. This training is focusing on competency based training instead of task based training. It arose from an industry-wide consensus that reduction in aircraft hull loss and fatal accident is needed. It was necessary to review the existing recurrent and type rating training for commercial pilots. The EBT programme and

philosophy are intended as means of assessing and training key areas of flight crew performance in a recurrent training system (ICAO 2006).

Pilot core competencies were developed to support the Evidence-based Training (EBT) concept adopted by ICAO in 2013. An international industry working group was established in 2007 for the development of a competency-based approach to recurrent training and assessment. The first and critical step in the development of EBT was to identify a complete framework of performance indicators, in the form of observable actions or behaviours, usable and relevant across the complete spectrum of pilot training for commercial air transport operations. These competencies and performance indicators combine the technical and non-technical (CRM) knowledge, skills and attitudes that have been considered essential for pilots to operate aircraft safely, efficiently and effectively. A framework of behaviours was developed, divided into 8 core competencies, each with observable performance indicators. The competencies were published in the ICAO Doc 9995 Manual of Evidence-based Training. The core competencies are primarily an assessment tool, offering a different approach from the evaluation of outcomes and manoeuvres, the purpose being to understand and remediate root causes of performance difficulties, rather than addressing only the symptoms. The purpose of these performance indicators is to underpin the creation of performance expectations at all stages of training in a pilot's career. To complete the picture, a fair and usable system of grading performance is also required. The development of pilot core competencies was considered as the first important step towards the creation of the "total systems approach to training". By far the most significant challenges for operators using these competency frameworks is the creation of an effective performance assessment and grading system, and subsequently the need for instructor training and the assurance of inter-rater reliability (IATA 2014).

It is impossible to foresee all plausible accidents, especially in a complex and high reliability system, where the next accident may be something completely unexpected. Competency based training is trying to address this by shifting from pure scenario-based training, to prioritizing the development and assessment of key competencies. It uses the KSA (Knowledge, Skill, Attitude) to master the infinite number of competencies that allow a pilot to manage situations in flight that are unforeseen by the aviation industry. The EBT competencies encompass what has previously recognised as technical and non-technical knowledge, skills and attitude. The aim of EBT is to help pilots develop the identified competencies that are required to operate safely, effectively and efficiently in a commercial air transport environment (ICAO 2013).

2.4 Inter-Rater Reliability

A major challenge in conducting any kind of performance assessment is the development of strong inter-rater reliability, and consistency in the approach. It is of great importance for the system (ICAO 2013). Each scenario provides an opportunity for in-depth feedback for any of the 8 core competencies. In many such focused scenarios, a pilot is exposed to variations in his own working environment. Every pilot has opportunity to practise aspects of the same core competencies in different situations, which accelerates the acquisition of expertise in the complicated domain which aviation is. From an inter-rater reliability perspective, it is essential that what is measured is

based on reliable observation. Even highly trained and experienced observers are limited in what they can reliably rate (Rosen et al. 2008). It is therefore very important that the rater is trained in facilitation, debriefing, and in using the system to judge the technical and non-technical skills. It is important to understand what separates *exceptional* from *average* or *poor* performance among teams. So, it is important to understand how teams successfully interact to deliver superior performance. To evaluate superior performance, some evaluation tool needs to be in place. In the commercial aviation, there are normally two pilots who form the team. For the purpose of training the crew, it is necessary to diagnose differences between individual and team performance.

3 Research Question

It is critical for the aviation industry to have a valid assessment and measurement system in place for flight crew performance. For both individuals and the system, it is important to have valid and reliable assessment methods to reduce subjectivity to a minimum. In this paper, the purpose is to analyse flight crew performance based on two different methods.

Those methods are:

- Performance Indicators (PI) used in Evidence Based Training (EBT)
- Desired Flight Crew Performance (DFCP): a task based binary checklist

Is a binary Yes/No observation checklist usable to effectively reflect flight crew performance assessment?

Do the PI's reflect that specific scenario the same way as the DFCP method does, does it give more insight into what is the core of the problem versus indicating that task was done or absent?

What competencies helps high performing crews and hinder low performing crews in a challenging scenario which includes responding to the unexpected?

As the binary checklist if more straight forward and simpler to use, which recommendations can be made for improvement of the binary checklist approach with regards to use in training?

4 Methods

4.1 Aim

The aim of the experiment was to create a scenario which includes elements which are normally not covered in routine training programs and would enable the researcher to monitor pilot performance in unexpected events. The experiment took place in a full flight, approved level D simulator, which enabled the crew to react exactly as they are used to in traditional training.

The researcher was provided with 25 exclusive access research videos of flight simulator events, by a large airline with a multicultural pilot workforce. Each crew was

exposed to the same scenario. The scenarios offered considerable challenges with elements of surprise and operational complexity. The presented videos were extracted from a real simulator environment. The pilots showed authentic nonscripted behaviours as they were not able to prepare themselves beforehand, but had to make fast decisions. Each video is about 30 min long for the whole scenario. Five of the videos were discarded as the content was obscured or sound corrupted which hindered the researcher to get a clear understanding of what was going on and made thorough analysis impossible.

4.2 Scenario

The scenario started at descent into an international airport in the Middle East. Three key events in the scenario formed the unexpected elements which were studied. In the first event, the crew (Captain and First Officer) were given clearance for an instrument approach. During the approach, the tailwind exceeded the limit for the aircraft, and this should instigate the crew to do a Go-Around; or later, for the crews that proceed with the approach, a loss of visibility below decision height forcing them to Go-around. During the Go-Around, the second event occurred, which was a subtle heading control failure in the automation, forcing manual reversion to regain heading control. During turn the final event took place: a bird strike hitting both engines, causing internal damage. The damaged engines would surge and stall until thrust was reduced on both engines. The crew were free to select any appropriate response to the failures. The pilots were faced with automation limitations and partial engine power on both engines. The six flight phases were the following:

- Instrument Approach (first approach)
- Go-Around
- Heading failure in automated system (subtle)
- Bird strike affecting both engines
- Planning the subsequent landing
- Approach and landing.

4.3 Analysis

The study was carried out in two steps: First an assessment of the performance of all crews based on a DFCP checklist; and then in step 2, the scores of the DFCP assessment were used to select the three highest, and three lowest performing crews, which were analysed further with Performance Indicators.

Step 1: Initially an objective method was selected to reduce or eliminate subjectivity. It is a known procedure to objectively measure the performance of pilots while undertaking certain tasks, like flight path management, or procedure adherence. It is more challenging to objectively measure the non-technical skill parts, like problem solving or situational awareness. To assist in that a known method was used: Desirable Flight Crew Performance (DFCP). This method is used to determine the action of the pilots against a safe outcome in the scenario. The DFCP list uses ranking of the flight crew against expected behaviour in the scenario. The industry is known for thorough

operating procedures or guidelines and training which were considered (Field et al. 2016). A list of 25 items was generated for this specific scenario. To increase validity in the choice of items on the list of tasks, two experienced Airbus simulator instructors were consulted in addition to the author who is experienced simulator instructor as well. The list was made up independently at first and then compared and finalised. As the videos were exclusively for the author, because of confidentially issues, the other experts could only theoretically assume the tasks needed for a safe outcome. They were given details of the scenario and used their experience to suggest the tasks that were considered important. There was agreement about all the items in the list. The DFCP list is divided into six phases see Fig. 1.

Phase	No.	DFCP:
First approach	1	Brief low visibility approach
First approach	2	Approach checklist
First approach	3	Arm Approach
First approach	4	Landing checklist before 1000´.
First approach	5	Clearly verbalize tailwind
Go-Around	6	GA due to tailwind (Critical)
Go-Around	7	GA due to loss of visual contact
Go-Around	8	Execute G/A actions - G/A Procedure (Flaps&Gear)
Heading failure	9	Verbalize heading failure
Heading failure	10	Execute immediate manual flying and turn right
Birdstrike	11	Verbalize birdstrike
Birdstrike	12	Verbalize surge/stall or ECAM actions
Birdstrike	13	Reduce power to try and stop surge
Birdstrike	14	Run correct Abnormal checklist (ECAM), ENG STALL
Birdstrike	15	Mayday call
Planning	16	Inform cabin crew of birdstrike/immediate landing
Planning	17	Review of aircraft/system status, options
Planning	18	Get info from ATC on WX, RWY avail
Planning	19	Assessment of RWY for second landing based on WX
Planning	20	No unnecessary delay for second approach
Planning	21	No landing with tailwind
Approach and landing	22	Verbalize energy management considerations
Approach and landing	23	Briefing, G/A no option
Approach and landing	24	Execute landing checklist
Approach and landing	25	Landing with flaps 3

Fig. 1. A list of 25 items divided into six phases in the scenario.

Each item on the list is considered essential for a safe outcome. The researcher watched all the videos and kept a score about items that were either completed or not completed. It was a score against the DFCP list, which included a number of safety critical items already identified. The list was used to collect each item and thereby draw a comparison between the crews. The rating was simply the sum of the DFCP items that were performed by relevant crew. The result was used to identify the high performing crew and low performing crew. Out of the twenty crews, one crew opted to land after the first approach despite having no visibility. For that reason, the crew was not included in the final comparison.

Each crew was scored according to the desired parameters, and total score calculated to select the three best performing crew and three least performing crew. One point was given for completing the task, and no point if the task was absent. For two of the items in the DFCP list "Go-Around due to tailwind" and "Reduce power to try and stop the surge" were considered critical and were given the weight of three points due to their role in the safe outcome of the exercise. Maximum score for each crew was therefore 28 (the crew got 3 points for *Go-around because of tailwind* and in case of *Go-around because of no visibility at decision height* one point was given).

The DFCP analysis only shows if certain task was completed, or absent. It does not explain why certain crews performed better than others according to the DFCP list.

Step 2: The second phase of the analysis tries to determine the differences in behaviour using the performance indicators (PI's) that are used in EBT for behavioural measurement. As previously mentioned, the PI's form observable behaviours and have been established by ICAO. Some modification is recommended for operators to adjust the PI to their specific needs in evaluation or assessments (Iata 2013). The PI's used in this experiment were adjusted by EBT Foundation and LOSA collaborative in a joint venture, and one competency added to the 8 established competencies. The added competency is *Knowledge* and it has its PI's as well. List of all competencies and their PI's can be found in appendix. The PI's are intended as a guiding tool to look beyond the outcome and subsequently look at the process that either hindered or helped the applicable crew in that specific scenario.

The researcher took the three-highest performing crew, and three lowest performing crew according to the DFCP and compared them. They were analysed by using observation and tangle them with applicable PI. Each PI was either positive (P) or negative (N) for each segment of the scenario. As the PI's are globally designed to capture as much as possible in the real world, some of the PI were not relevant to this specific scenario and therefore not observed. There are total 66 performance indicators that were consulted for each six phases of each scenario. The observation was made against the individual and not as the crew as a collective unit. This method is considerably different from real environment as the researcher was able to use pauses, rewind etc. extensively. This allows the analysis to be very detailed. The PI's were used as a checklist in effort to reduce the subjectivity because of one rater.

Each phase was then compared between the crews to try to establish what competencies were hindering, or helping the crew to deal with the situation.

5 Results

During this challenging scenario, all crew except one managed to land the aircraft on the runway. It was very different between crews how they handled the situation and situational awareness at first glance seemed to be lacking for low performing crews. Each crew had to spend some time on analysing what was actually going on, and this provided an opportunity to notice some differences. As all the crew were doing this on voluntary basis, no crew actually had engine failure on both engines, which would have been a more dramatic ending. That could however be expected in an actual scenario if

the crew does not reduce the engine power below the engine EGT (exhaust gas temperature) limit. After analysing the three-high performing crew and three low performing crew, a clearer better picture was starting to form. The intension was to compare the DFCP list with the Performance Indicators and see if there would be the same outcome.

A descriptive analysis was performed. The results of the two methods indicated a similar outcome. Both indicated that low-performing crew and high-performing crew scores were concurrent. The DFCP method is a checklist composed of items that were considered necessary for a safe outcome. There were considerable differences in the performance of the crews. After all the videos had been watched, a score based on the DFCP list was generated. See Fig. 2.

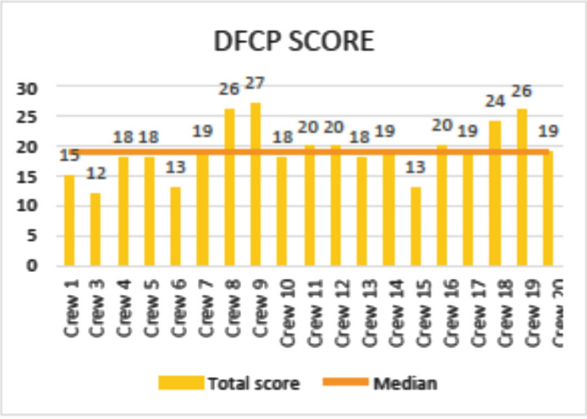

Fig. 2. DFCP total score for all 19 crews analysed. Median line for reference

The analysis gave a strong indication of differences in performance, see Fig. 3. The range of scores were from 12 as the lowest, to 27 the highest. Average is 19.15 and median value was 19. This indicates that there were considerable differences between high and low performing crews, and fairly high spectrum between high and low. High performing crews scored on many of the tasks required, or almost all of them. Low performance crews scored much less where tasks were commonly absent. It turned out that the decision to give two tasks in the DFCP list more weight than others, based on their criticality, did not affect the selection of crews. DFCP could distinguish between high and low performance crews but could not show reasons for crew actions. This will be further discussed in discussion and conclusion.

Figure 4 Drafts up simplified but typical track of the scenario. The green dotted line demonstrates track chosen for typically higher performance crews while the lower performance crews typically followed the red dotted line after the go-around. The pattern indicated by the red dotted line meant that the crew was landing with more tailwind then the aircraft is certified to do, while the green dotted line demonstrates the preferred track. This figure is just for demonstration it does not depict the exact track nor all courses of action decided by each crew.

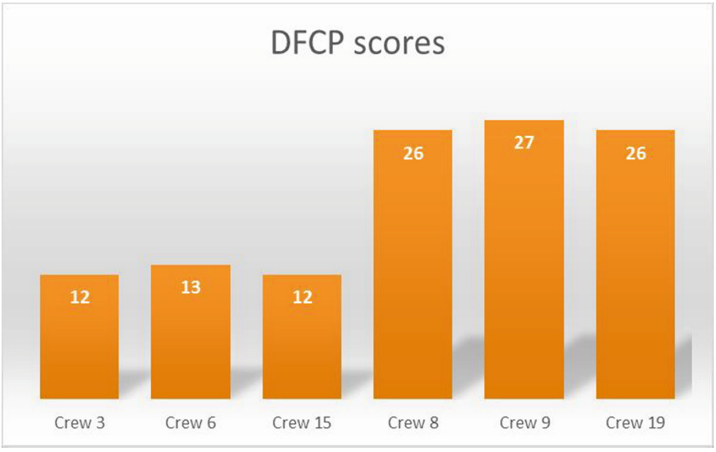

Fig. 3. Score for three highest performance crews and three lowest performance crews.

The result of the detailed PI analyses is presented, based on the numbers of (P) positive and (N) negative scores given in the evaluation. The score is made against observation that tangles with PI's and is either positive or negative. This allowed to obtain an overall score for each flight segment. Some of the PI are not observed. Each flight phase is scored separately in all competencies to see the difference when crew is faced with unexpected scenario, like bird strike and planning. Figures 5, 6, 7, 8, 9 and 10 show the difference in observed PI's in each flight phase.

The results show that in the first two phases there were not much differences between observed crew PI's. There are negative indications which are related to not detecting the wind change on the approach, which affected the situational awareness and decision making. When crews are faced with standard operation or threat they can expect (typical training scenario which pilots are frequently exposed to in simulator training e.g. Approach and Go-arounds etc.) they seem to handle that within certain criteria. Not much difference is detected between high and low performing crew. On the other hand, when unexpected events are introduced, there is more difference detected. Where the heading failure is introduced, which is subtle failure in the automation, a greater difference is detected. One crew was hesitant to revert to manual flying and therefore scored negatively. Some difference is detected there but not as much as would be expected based on the DFCP. During bird strike and subsequent planning phase, considerable difference is detected in PI's. High performing crew demonstrates more positive PI's but there is some evidence of negative PI's. That would be expected as the scenario is challenging and not typical for pilot routine training. The low performing crew is scoring high on negative PI's during these phases and reduction in positive scores are detected.

To explore further and see which competencies were helping or hindering the crew, a comparison was also made regarding all the 9 competencies. In Figs. 11, 12, 13 and 14 which is presented as a heat map to visually demonstrate the mostly affected competencies.

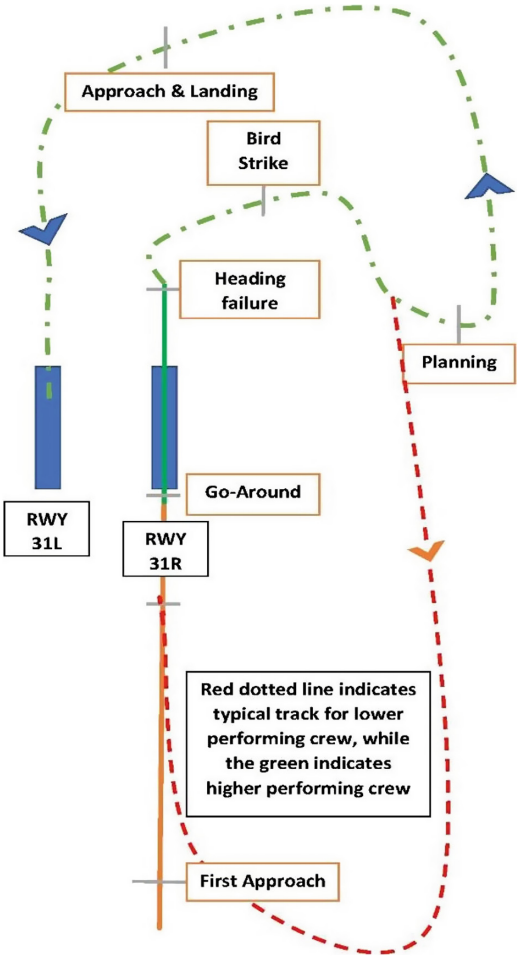

Fig. 4. Two typical tracks decided by crews (Color figure online)

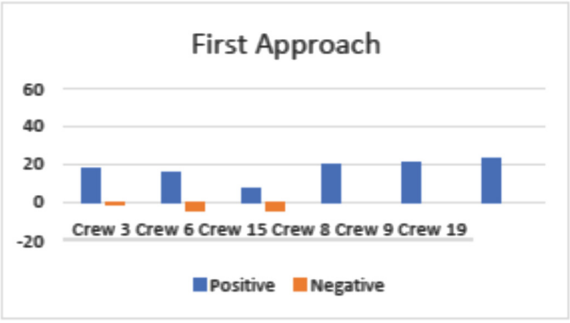

Fig. 5. Performance indicators identified for First Approach

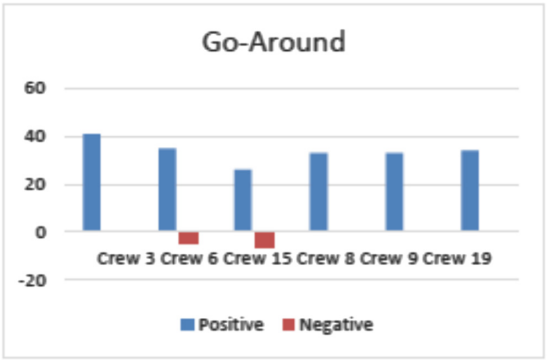

Fig. 6. Performance indicators identified for Go-Around

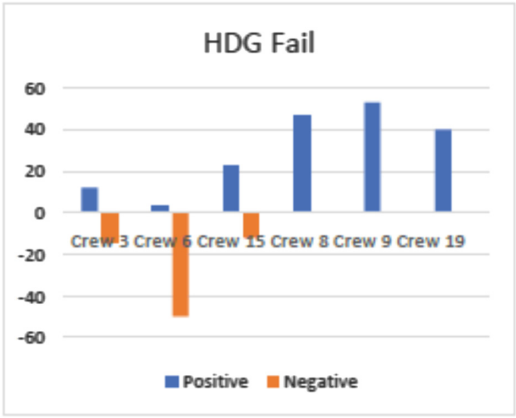

Fig. 7. Performance indicators identified for Heading Failure

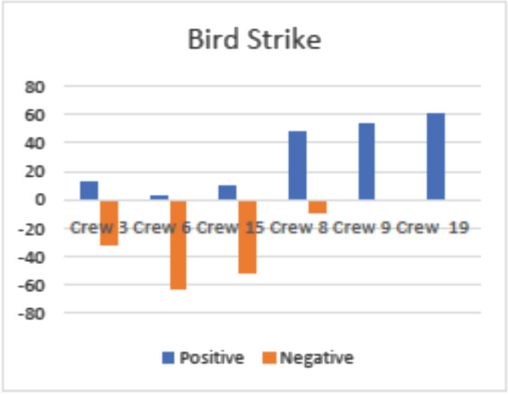

Fig. 8. Performance indicators identified for Bird Strike

Fig. 9. Performance indicators identified for Planning

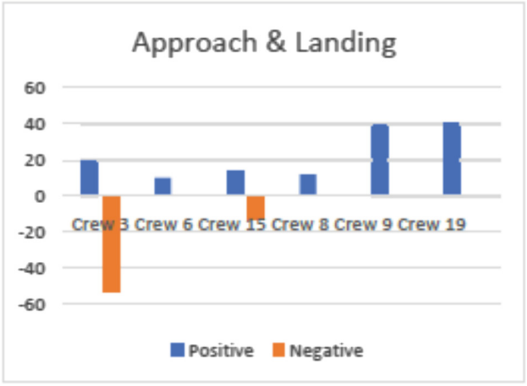

Fig. 10. Performance indicators identified for Approach & Landing

Positive - High Performance Crew						
Phase / Competence	First Approach	Go-Around	Landing Inflare	Touchdown	Planning	Approach & Landing
APK	18	18	16	18	14	17
COM	30	31	28	27	26	16
FPA	0	10	3	2	0	0
FPM	0	1	13	11	10	6
KNO	4	13	14	13	16	8
LTW	0	8	15	28	20	15
PSD	0	1	19	23	14	8
SAW	12	12	12	15	12	11
WLM	0	6	20	25	20	12

Fig. 11. Demonstrates the positive performance indicators that were analysed with the three high performing crews

Positive - Low Performing Crew						
Phase / Competence	Post Approach	Go-Around	Heading Failure	Bad strike	Planning	Approach & Landing
APK	14	16	8	2	3	7
COM	21	24	7	3	10	8
FPA	0	15	6	0	0	2
FPM	0	2	5	6	9	9
KNO	0	10	2	1	4	2
LTW	0	10	6	3	9	8
PSD	0	3	0	6	4	0
SAW	6	12	3	1	5	6
WLM	0	10	1	2	4	2

Fig. 12. Demonstrates the positive performance indicators that were analysed with the three low performing crews.

Negative - High Performance Crew						
Phase / Competence	Post Approach	Go-Around	Heading Failure	Bad strike	Planning	Approach & Landing
APK	0	0	0	0	0	0
COM	0	0	0	0	2	2
FPA	0	0	0	1	0	0
FPM	0	0	0	0	0	0
KNO	0	0	0	4	2	0
LTW	0	0	0	0	0	0
PSD	0	0	0	4	2	0
SAW	0	0	0	0	0	0
WLM	0	0	0	0	0	0

Fig. 13. Demonstrates the negative performance indicators that were analysed with the three high performing crews

Negative - Low Performing Crew						
Phase / Competence	Post Approach	Go-Around	Heading Failure	Bad strike	Planning	Approach & Landing
APK	4	2	6	11	11	5
COM	0	2	20	29	23	15
FPA	0	3	9	0	0	2
FPM	0	0	5	4	9	2
KNO	0	0	4	16	14	7
LTW	0	2	7	19	11	10
PSD	0	2	5	24	24	11
SAW	6	0	9	13	12	8
WLM	0	0	11	28	22	13

Fig. 14. Demonstrates the negative performance indicators that were analysed with the three low performing crews

The heat map indicates that high-performing crew were scored positively in Communication (COM), Leadership and Teamwork (LTW), Problem Solving and Decision Making (PSD) and Workload Management (WLM).

Indication for the low performing crew is similar for positive score in the first two phases or slightly lower. Negative indication is very low for the first two phases which indicates that the low performing crews are performing normally under familiar, or expected events. This indicates that it is hard to detect the difference between the high and low performing crew under expected or familiar events. When unexpected events come into play, the positive scoring drops rapidly for the low performing crew. The outcome of the negative scores supports the picture as well: Low performing crews score negatively especially with unexpected events, which is reflected in Communication (COM), Leadership and Teamwork (LTW), Problem Solving and Decision Making (PSD), and Workload Management (WLM). This is quite opposite of the results of the high performing crews. The four identified competencies are helping the high performing crews, while at the same time hindering the low performing crews, indicating that certain non-technical skills are required to deal with this unexpected event. Although the high performing crews are scoring much more favourably, there is an indication of minor rise in negative PI score during a challenging situation, which would have been expected.

6 Discussion

The DFCP method was used to rate the performance of the crew in safety related decisions and actions. There was consensus from three experts regarding each task on the list. The list of tasks was successfully used to identify the desirable and less desirable actions taken by flight crew in response to expected and unexpected events they encountered in the simulator. DFCP could distinguish well between high-, and low performing crews that were subsequently analysed further with the Performance Indicators. The DFCP identified the flight crews that performed well but it did not explain potential reasons for the decisions, actions or differences in performance. To investigate these descriptions of performance and to investigate the potential reasons behind it an additional analysis was performed.

Reflecting on the results from the performance indicator analysis, low performing crew struggled in the competencies where high performing crews were strong. This was mostly evident in high workload situations. It was evident in the planning phase that low performing crew were suffering from inadequate communication which resulted in poor planning and lack of situational awareness. From the heat map (Figs. 11–14) where each competency is presented, Leadership and Teamwork, Problem Solving and Decision Making, Workload Management, and Communication are the ones that are most prominent. The competencies cannot be viewed as separate, independent entities. They are interconnected, and some of them are a consequence of good or bad in others. Communication is a good example: Communication is not an indicative of good or bad performance in itself, rather it propagates positive or negative behaviour in other competencies.

Communication was integrated into the four categories of non-technical skills in the NOTECHS, and not used as a separate competency (Flin et al. 2003), which, based on the findings of this study, might have left out some valuable data. Studies have found that high performing crews discussed in-flight problems more thoroughly than low performing crews. The also used low workload phases to plan ahead and anticipate by using strategy and planning. They were found to use fewer commands during a high workload situations (Grossman and Salas 2011).

As all the crews in this study are working for the same operator, and are current and active pilots on the type. This means that they are all trained and used to the same operating procedures. There was no clear difference between the crews when dealing with expected events. Contradicting to the other study, the low performance crews were not obviously detected in the first two phases. This concludes a slightly different picture as that had been detected before from other study on similar subjects, that the competencies did not show that much of a difference during routine operation between high and low performance crews (Field et al. 2016).

Events and tasks in the first two phases of the scenario are what pilots are exposed to in their regular recurrent training. This supports that training should focus on unexpected events, frequently exposing pilots to demanding situations, requiring them to utilise all the spectrum of the core competencies. The collection of observed competencies makes it possible to draw a clear picture of the difference between high and low performing crews.

The DFCP is essentially a checklist of tasks performed or not performed. Once the list of tasks has been agreed on, the marking is relatively straightforward, and different observers are likely to check the same boxes so there is a high likelihood of consistency in marking. It was more relevant to use that method as the use of rewind and pauses was not used extensively, which lowered the work load on the researcher. In testing terminology: the method is reliable. Performance indicators are observable behaviour which reflect the competencies as described in EBT. As the goal of the PI's is to reveal root causes to performance below, at, or above the expected level, this method probes deeper than DFCP. It seeks to give the observer an understanding of why a task was completed or omitted, not just if. However, as there is substantial judgement involved on behalf of the instructor, and a high probability of rater bias, the instructors need to be trained and standardised. Again, in testing terminology: this method presents challenges in reliability, but it scores high in validity because it attempts to capture things that are important to assess, not just those that are easy to assess. However, during this study the author could use the checklist in a different way than would be possible in simulator training, as the use of pauses, rewind etc. were useful. Even if the validity composes some real challenges, the positive thing is getting to the root cause. It would have been likely that at first glance the Situation Awareness would have been the cause of some of the failures. Situation Awareness is one of the labels in CRM which is commonly referred to. This label is commonly referred to in accident investigations as well, where the probable cause was loss of situational awareness. The performance indicators in this scenario, although detecting some negative performance in situation awareness, indicate that the crew was lacking in other domains more dominantly. These evidences indicate that it is important to retrain and debrief the crew to help with transfer of training. Behind the terms, human error and situational awareness is another

psychological world to do with attention, perception, decision making, and so forth. Human factors have produced or borrowed terms that try to capture these phenomena. Complacency, situation awareness, crew resource management, shared mental models and workload are common currency today that are deep rooted in science, but at the blunt end, people have difficulties putting finger on, and don't dare to ask what they actually mean (Dekker 2006). The importance of the performance indicators is training the instructors. They are not intended to apply a psychological jargon; the performance indicators are simply a behavioural marker that are intended to tangle with observations that are made by the instructor or examiner. From this study, there are detailed data which would not be expected in normal recurrent simulator, where the simulator instructor is not only observing the crew. The simulator instructor is also occupied running the simulator, acting as Air Traffic Controller (ATC), playing cabin crew, and so forth. For the instructor, it is important that he acts like he normally does, observes and records what he sees from the crew, and after the simulator, he takes the recording and observation and applies it to the performance indicators. This assists the instructor by not overloading him with tasks and the performance indicators are something which he should be able to observe. As mentioned before, the instructor training is very important as the data gathered in the simulator is only as good as the validity of the data. For training transfer of the pilots, which might add is the ultimate goal, facilitation debriefing technique is considered essential for training cognitive ability and training transfer (Grossman and Salas 2011). It is important for the student or pilot to understand his weaknesses or the root cause instead of simply stating the label of non-technical skill that needs improvement. Same apply to high performing crew, realising why things went well is critical to motivation and transfer of training (McDonnell et al. 1997).

As explained above, two methods were used to analyse crew performance in this study: DFCP (Desired Flight Crew Performance) checklist; and Performance Indicators. One of the videos revealed an interesting aspect of the different things that the two methods capture.

The above-mentioned video was very interesting with respect to the two methods. The crew performed nearly all the DFCP tasks – but crashed the aircraft. The list did not specify that the aircraft had to land on the runway. This crew scored above average, but the outcome was unsafe, to say the least. So, the conclusion in this scenario, the method is lacking in validity. The PI was much more effective in capturing deficiencies in the performance of this crew. Although the crew did many things well, they would have gotten a negative indication for Application of Procedure, Problem Solving, Situation Awareness, and Communication.

It is outside the scope of this paper, but it would be very interesting to try to improve the DFCP list and link the items to performance indicators. The DFCP list served its purpose to distinguish between high and low performance crews but to combine these two methods would assist the instructor. If successful, that would strengthen the overall validity and reliability of the combined method and enable the instructor to capture more information. This was given some consideration but there seem to be some difficulties in doing that, and it might be impracticable. For example, the PI's are based on individual performance, but the DFCP assesses tasks performed by the crew. This would need to be solved.

One of the competencies that is identified strongly in this experiment was Leadership and Teamwork (LTW). The Teamwork competency explains the effective teamwork which has lately attracted more attention and interest in researches. Empirical evidence suggests, that there is an increase in requirement for teamwork in complex domains. Teamwork studies have indicated that individuals become less willing to accept input or feedback from their team members when in high workload situations. Teams that are oriented have shown to perform better under high pressure or high workload. Team orientation is based on effective communication, decision making and being able to manage workload (Salas et al. 2008). Although there is not full consensus on Teamwork researches, this coincides with the data that were received in this study on identified competencies where the high performing crew scored positively in these domains and where the low performing crew scored negatively. We are seeing this study supporting this empirical finding. With closer look, there is also another empirical finding where the study can be connected: Safety II is a recent term which describes safety from another point of view. Safety II is defined as the ability to succeed under both expected and unexpected conditions. It becomes a characteristic of how systems function and is based on that the human is clearly an asset rather than a liability (Hollnagel 2014). With the Performance Indicators, more accurate data will be received to study both negative and positive performance. With the positive data, there is a possibility to study expertise, taking into account which competencies are dominant in different scenarios, and use that data to look more closely at safety II. This way it could help to achieve more total system approach to training. It has the potential to be beneficial for operators that are willing to adopt this way of thinking as is defined in Safety II. Current safety systems for a typical airline today are relying on flight data that it receives from their own aircraft (Flight Data Monitoring). These data are in fact quite similar to the DFCP. They are real data and can tell you what went on, but they are limited as a tool to explain the reason. If it is possible to study why things go right, the data given from the performance indicators in EBT will be very valuable. In this study, the analysis done with the performance indicators give a detailed picture of the performance, hence they can be used to create a feedback loop into the training and safety system. They are currently being used for investigating human (BEA 2013). What is further possible is to use the data to study expertise per se, so we know why things are going right and continue from there.

7 Limitations

The author is experienced simulator examiner and instructor. Nevertheless, is the research analysis very different from real simulator environment, the ability to use rewind and pauses to thoroughly look at whatever is said and tone of voice gives more detail in analyses. It was new for the author to be able to look at body language as the details are very subtle. For example, hand gesture, etc. It also possible that confirmation bias took place after the selection of low and high performing crew had taken place. Knowing that might have influenced the result of performance indicators. In order to reduce or hopefully remove the confirmation bias, the discussions among the crew were mostly written up to get better understanding and being aware that things might be biased.

8 Recommendations

Further studies on training transfer is needed. Further knowledge is needed to study if there is a transfer in KSA (Knowledge, Skill, Attitude) between different scenarios or domain. There are positive indication that training transfer is likely specially in cognitive skills, so the cognitive ability is transferring to other domains (Grossman and Salas 2011).

9 Conclusions

Systematic gathering of valid data about crew performance is essential for effective training and safe flight operations. A binary yes/no checklist is usable to assess flight crew performance, but it has serious limitations as, while it gives useful data about the completion of tasks, it is not a suitable tool to analyse why a task is not completed. As the checklist is used in this study it may fail to capture important information so it needs to be developed further.

The two methods that were used in this study were similar in their ability to detect high-, and low performing crews. However, the information they provide the researcher with are very different. The Performance Indicators were very effective in identifying the underlying competencies that helped, or hindered the crews. This can help training departments to make more informed decisions about training needs.

The competencies that help high performing crews are, LTW, COM, PSD, and WLM. Conversely, lack of those competencies was a barrier to low performing crews.

The use of Performance Indicators poses challenges in rater reliability because instructor bias is likely to cause subjectivity. These challenges can be mitigated with effective training of instructors. The performance indicators are very useful to capture the competencies, which are important for pilots to have to effectively master complicated, unexpected situations. The study detected more negative performance indicators in competencies that are more related to non-technical skills rather than technical skills for the low performing crews. The difference between high and low performing crews in technical skills was not as evident, which indicates that training needs to be more effective in non-technical skills, despite the effort that the industry has made e.g. with the emphasis on NOTECHS and CRM training in the last few decades. Conclusions should state concisely the most important propositions of the paper as well as the author's views of the practical implications of the results.

References

Barry Issenberg, S., Mcgaghie, W.C., Petrusa, E.R., Lee Gordon, D., Scalese, R.J.: Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med. Teach. **27**(1), 10–28 (2005). https://doi.org/10.1080/01421590500046924

BEA: Sx-Bhs, August 2013

Cannon-Bowers, J.A., Salas, E.: Team performance and training in complex environments: recent findings from applied research. Curr. Dir. Psychol. Sci. **7**(3), 83–87 (1998). https://doi.org/10.1111/1467-8721.ep10773005

Dekker, S.: The Field Guide to Understanding Human Error. Ergonomics, vol. 51 (2006). https://doi.org/10.1080/00140130701680544

Ericsson, K.A., Ward, P.: Capturing the naturally occurring superior performance of experts in the laboratory: toward a science of expert and exceptional performance. Curr. Dir. Psychol. Sci. **16**(6), 346–350 (2007). https://doi.org/10.1111/j.1467-8721.2007.00533.x

EU: Commission Regulation (EU) No. 965/2012. Official Journal of the European Union, 5 October 2012

Field, J.N., Mohrmann, F., Fucke, L., Grácio, B.C.: Flight crew response to unexpected events: a simulator experiment. In: AIAA Modeling and Simulation Technologies Conference (2016). https://doi.org/10.2514/6.2016-3373

Flin, R., Martin, L., Goeters, K.-M., Hörmann, H.-J., Amalberti, R., Valot, C., Nijhuis, H.: Development of the NOTECHS (non-technical skills) system for assessing pilots' CRM skills. Hum. Factors Aerosp. Saf. **3**(2), 95–117 (2003). http://www.safetylit.org/citations/index.php?fuseaction=citations.viewdetails&citationIds%5B%5D=citjournalarticle_37801_6, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.472.6866&rep=rep1&type=pdf

Fowlkes, J.E., Dwyer, D.J., Oser, R.L., Salas, E.: Event-Based Approach to Training (EBAT). Int. J. Aviat. Psychol. **8**(3), 209–221 (1998). https://doi.org/10.1207/s15327108ijap0803

Grossman, R., Salas, E.: The transfer of training: what really matters. Int. J. Train. Dev. **15**(2), 103–120 (2011). https://doi.org/10.1111/j.1468-2419.2011.00373.x

Harris, D.: Human Performance on the Flight Deck. CRC Press, Boca Raton (2012)

Helmreich, R.L., Klinect, J.R., Wilhelm, J.A.: Models of threat, error, and CRM in flight operations. In: Proceedings of the Tenth International Symposium on Aviation Psychology, pp. 677–682 (1999)

Helmreich, R.L., Merritt, A.C., Wilhelm, J.A.: The evolution of crew resource management training in commercial aviation. Int. J. Aviat. Psychol. **9**(1), 19–32 (1999). https://doi.org/10.1207/s15327108ijap0901_2

Hollnagel, E.: Is safety a subject for science? Saf. Sci. **67**, 21–24 (2014). https://doi.org/10.1016/j.ssci.2013.07.025

IATA. Evidence-Based Training Implementation Guide (2013)

IATA. Data Report for Evidence-Based Training (2014). http://www.iata.org/whatwedo/ops-infra/itqi/Documents/data-report-for-evidence-basted-training-aug2014.pdf

ICAO. Procedures for Air Navigation Services - Training (Doc 9868). Planta (2006)

ICAO. Manual of Evidence-based Training (2013). http://www.icao.int/SAM/Documents/2014-AQP/EBTICAOManualDoc9995.en.pdf

Kanki, B., Helmreich, R., Anca, J.: Crew Resource Management. Crew Resource Management, pp. 1–5. (2010). http://www.scopus.com/inward/record.url?eid=2-s2.0-84882499276&partnerID=40&md5=4aec85e9c8960fc3d3629013edeb580b

Kanki, B.G., Greaud, V.A., Irwin, C.M.: Communication variations and aircrew performance. Int. J. Aviat. Psychol. **1**(2), 149–162 (1991). https://doi.org/10.1207/s15327108ijap0102

McDonnell, L.K., Jobe, K.K., Dismukes, R.K.: Facilitating LOS Debriefings: A Training Manual, March 1997

Orlady, H.W., Orlady, L.M.: Human factors in multi-crew flight operations. Aeronaut. J. **106** (1060), 321–324 (2002)

Osgood, C.E.: The similarity paradox in human learning: a resolution. Psychol. Rev. **56**(3), 132–143 (1949). https://doi.org/10.1037/h0057488

Rankin, A., Woltjer, R., Field, J., Woods, D.: "Staying ahead of the aircraft" and Managing Surprise in Modern Airliners. In: Proceedings of the 5th Resilience Engineering Association Symposium, pp. 209–214 (2013). http://www.resilience-engineeringassociation.org/download/re-sources/symposium/symposium-2013/

Rosen, M.A., Salas, E., Wu, T.S., Silvestri, S., Lazzara, E.H., Lyons, R., Weaver, S.J., King, H. B.: Promoting teamwork: an event-based approach to simulation-based teamwork training for emergency medicine residents. In: Academic Emergency Medicine, vol. 15, pp. 1190–1198 (2008). https://doi.org/10.1111/j.1553-2712.2008.00180.x

Salas, E., Rosen, M.A., Held, J.D., Weissmuller, J.J.: Performance measurement in simulation-based training: a review and best practices. Simul. Gaming **40**(3), 328–376 (2008). https://doi.org/10.1177/1046878108326734