Should we pursue inter-rater reliability or diversity? An empirical study of pilot performance assessment

Deveríamos buscar confiabilidade inter-avaliadores ou diversidade? Um estudo empírico sobre a avaliação do desempenho de pilotos

David E. WEBER[1]
Wolff-Michael ROTH[2]
Timothy J. MAVIN[3]
Sidney W. A. DEKKER[4]

ABSTRACT: *Reliably and equitably assessing the performance of commercial pilots has not always proven easy. It is thus necessary to take a closer look on how performance is assessed in practice. This study explores the reasoning behind this process as stated by experienced pilots who assess safety-critical pilot performance. Using a theoretical model of performance, three pairs of airline captains assessed a captain and a first officer in two video scenarios. The results show that assessors apply the same or similar reasons to arrive at different assessments or use different reasons to arrive at the same assessment. In addition, conclusions about inter-rater reliability and efforts intended to increase it are drawn.*

KEYWORDS: *performance assessment; pilots' performance; non-technical skills; nonlinearity; reliability.*

RESUMO: *Avaliar o desempenho dos pilotos comerciais com imparcialidade e confiabilidade nem sempre é tarefa fácil. Assim, é preciso examinar mais de perto como tal desempenho é avaliado na prática. O presente estudo explora o raciocínio por trás deste processo de acordo com pilotos experientes que avaliam questões fundamentais para a segurança no desempenho de pilotos. Utilizando-se de um modelo teórico de desempenho, três pares de comandantes de linhas aéreas avaliaram um comandante e um copiloto em dois vídeos. Os resultados demonstram que avaliadores fazem uso de razões iguais ou similares para chegarem a avaliações diferentes ou de razões diferentes para chega-*

[1] Safety Science Innovation Lab, Griffith University, Brisbane, Australia. E-mail: david.weber@griffithuni.edu.au
[2] Applied Cognitive Science, University of Victoria, Victoria, Canada. E-mail: wolffmichael.roth@gmail.com
[3] Griffith Institute for Educational Research, Griffith University, Brisbane, Australia. E-mail: t.mavin@griffith.edu.au
[4] Safety Science Innovation Lab, Griffith University, Brisbane, Australia. E-mail: s.dekker@griffith.edu.au

rem à mesma avaliação. Conclusões sobre a confiabilidade inter-avaliadores e os esforços intentos em aumentá-la são formuladas também neste estudo.
PALAVRAS-CHAVE: *avaliação de desempenho; desempenho de pilotos; habilidades não-técnicas; não-linearidade; confiabilidade.*

1 Introduction

In 2010, Air India operated flight IX-812 from Dubai to Mangalore. During landing, the Boeing 737-800 overshot the runway, hit airport infrastructure, and fell into a gorge. The aircraft was destroyed on impact and a subsequent fire killing 152 passengers and all 6 crewmembers. Only 8 individuals survived. The Ministry of Civil Aviation (2010) determined the following cause for the accident: "[the] Captain's failure to discontinue the 'unstabilised approach' and his persistence in continuing with the landing, despite three calls from the first officer to 'go around' and a number of warnings from EGPWS [Enhanced Ground Proximity Warning System] (MCA, 2010, p. 115). The report also listed several other contributing factors: not properly planning the descent profile, the copilot not taking over the controls, and fatigue. All of the reasons listed fall into an area of piloting performance generally referred to as non-technical skills, including decision-making (failure to discontinue, to plan), communication (failure to be receptive to input), or management (failure to mitigate threats or errors; inability to control self or crewmember's performance).

How the aircraft accident investigation board viewed and assessed the Air India disaster is one possible perspective, which benefitted from the knowledge that whatever happened in the cockpit led to an accident. However, one could approach the assessment of performance on the flight deck differently: from the perspective of the pilots by reconstructing the event and producing a videotaped scenario. The question then would be, "Under the given circumstances and without knowing the outcome, what are the most reasonable actions and decisions to be taken?" Experienced pilots and flight examiners could be asked to assess the performance without knowing the final outcome. Would these pilots have observed, and agreed upon, the same pilot-performance criteria? How would they have behaved if they could have seen themselves in the given situation with the first officer? The purpose of this study is to provide some answers to questions of this kind. We enquire how pilots view performance of their peers in difficult situations that compromise safety.

The Mangalore case once again showed that performance other than that relating to the technical skills (TS) of flying an aircraft—aviation knowledge and flying skills—may play a role in major disasters. Records suggest that there were a considerable number of accidents in the aviation industry, where the problems involved could not be related to TS or mechanical failure (ORLADY, ORLADY, 1999; HELMREICH, MUSSON, SEXTON, 2004). Examples include the collision of KLM 4805 with Pan Am 1736 in 1977, Air Florida Flight 90 in 1982, or the controlled flight into terrain (CFIT) of

Crossair 3597 in 2001. Until the late 1970s, pilot training primarily focused on TS. This changed dramatically in 1978 when NASA research concluded that the majority of aviation disasters were associated with so called non-technical skills (NTS), including leadership, communication, teamwork, and decision making (HELMREICH et al., 2004; WEICK, 1990; COOPER, WHITE, LAUBER, 1979). NTS became increasingly important with the introduction of multi-crew cockpits. When aircraft became larger and operation more complex, a first officer was added to the flight deck. For a considerable time, however, first officers were seen as redundant pilots, who could act as operational backups in the unlikely case that the captain became incapacitated (ORLADY, ORLADY, 1999). Besides providing assistance with flight planning and radio communication, first officers traditionally were there to reduce workload if required by their captains.

This perspective has largely changed. In today's multi-crew operation, *individual* prowess is no longer seen as a pilot's strength. Instead, safety is increasingly regarded to be the product of a team, which involves more practitioners than the cockpit crew (ORLADY, ORLADY, 1999). Large efforts have thus been made to increase pilots' NTS (FLIN et al., 2003). One of the first efforts to reduce problems related to NTS was the NASA-sponsored "Resource Management on the Flightdeck" workshop in 1979 (COOPER, WHITE, LAUBER, 1980; FLIN, O'CONNOR, MEARNS, 2002; HELMREICH, MERRITT, WILHELM, 1999). Based on NASA's findings that the majority of aviation accidents were attributed to NTS, the workshop aimed at promoting psychological research into aviation accidents (COOPER et al., 1980). During this meeting the concept "Cockpit Resource Management" (CRM) was created (HELMREICH et al., 1999). It was defined as the "process of training crews to reduce 'pilot error' by making better use of the human resources on the flightdeck" (p. 19).

After the NASA workshop there was a more widespread recognition of the importance to increase pilots' NTS (FLIN et al., 2002). Consequently, many airlines introduced CRM training programs to mainly improve crews' interpersonal (non-technical) skills (HELMREICH et al., 1999). CRM training has proliferated around the world and became mandatory under many civil aviation regulatory authorities. To make sure that the cabin crew is included, terminology was later adapted from "cockpit" to "crew" resource management (HELMREICH, FOUSHEE, 1993; ORLADY, ORLADY, 1999).

Airlines and civil aviation authorities have experienced difficulties developing reliable and valid measures to assess NTS (ORLADY, ORLADY, 1999). Initial advances occurred by developing so-called "behavioral markers" (HELMREICH, FOUSHEE, 1993) that can be used in "line oriented flight training" (LOFT) to assess the performance of a crew (see Line/LOS Checklist, LLC). In contrast to a binary system that distinguished between satisfactory and unsatisfactory performance, LLC provided the means to assess the crew's knowledge, skills, and abilities in a more descriptive way.

Another approach to assess pilots' NTS was developed in the NOn-TECHnical Skills (NOTECHS) project. While many larger airlines developed their own assessment systems, a number of smaller European airlines lacked resources or expertise to produce assessment tools (O'CONNOR et al., 2002). Consequently, the European Joint Aviation Authorities (JAA) called for a basic, generic method of evaluating NTS for all JAA countries and operators (FLIN et al., 2003; O'CONNOR et al., 2002). A research consortium—called the NOTECHS Project—was thus set up, which was required "to identify or to develop a feasible and efficient methodology for assessing pilots' non-technical (CRM) skills" (FLIN et al., 2003, p. 96). The authors suggested 4 main categories, divided into two subgroups: social skills (including co-operation, leadership, and management) and cognitive skills (situation awareness, decision making). Each category (i.e., situation awareness) consists of various elements (e.g., awareness of time) that summarize numerous behavioral markers (i.e., "Identifies possible future problems" (p. 106) refers to good practice).

NOTECHs and its derivatives continue to be used around the world. However, concern has emerged about its initial design. Separating TS from NTS is a common trend, yet has recently been critiqued (MAVIN, DALL'ALBA, 2010). These authors regard such separation as a distortion that does not reflect the judgments that are made about pilots' performances in the cockpits. Instead, they suggest the so-called "Model for Assessing a Pilot's Performance [MAPP]" (Figure 1), which combines TS (aircraft flown within tolerances [AC], aviation knowledge [KN]) with NTS (situational awareness [SA], decision-making considerate of risk [DM], management of crew [MN], and communication [CM]). The holistic character of complex performances is better represented in such an interconnected model than in technical rationalist models that break skills down into various smaller measurable parts (MAVIN, 2010). By making a distinction between essential and enabling skills, which both include a mix of TS and NTS, the MAPP takes into account that flight examiners draw their final decision about pilots' performance on these essential skills. Consequently, pilots who show difficulties within their essential skills may immediately fall below minimum standard.

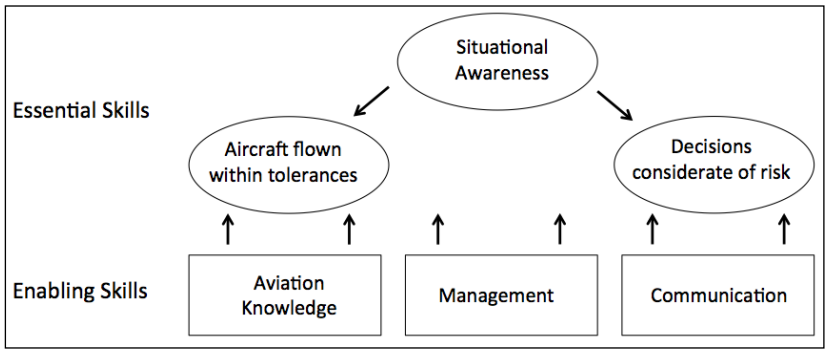

Figure 1: Model for Assessing a Pilot's Performance (MAVIN, 2010, used with permission of the author).

The MAPP was used to gain better insight into how airline professionals assess their colleagues' performance. The research showed large variations in the assessment scores of first officers (FO), captains (CAP), and flight examiners (FE) who assessed the performance of an airliner crew (MAVIN, ROTH, DEKKER, 2012). In one of the three scenarios, assessors' scores especially differed when judging the captain's performance: FO all passed the captain, while the FE all failed him. CAP assessor pairs, however, did not assess the performance in the same way: some regarded it as a pass while others as a fail. In certain assessment areas, the scores provided differed up to 2.5 marks (on a scale from 1 to 5). The differences found between assessors' scores when rating the performance of a pilot in the very same scenario necessitates further research. Whereas some theoretical progress has been made, better understanding is required of how pilot-performance is assessed in the practical field of airline aviation. The present study was designed to understand how pilots assess the performance of their peers in situations that compromise safety. By investigating assessment practices in depth, this study contributes to both our understanding of modeling the assessment of performance and improving pilot assessment. Of special interest are not the assessment scores themselves but the reasons that pilots use when assessing the performances of their peers.

2 Methods

2.1 Participants

Participants were six airline captains (CAP) that all work for the same airline. The selection was a function of the airline's roster: captains were randomly picked among those with free slots during the one-week data collection period. The mean age of the participating captains was 49.2 years (*SD* = 6.7) and they had a mean of 24.7 years (*SD* = 7.0) years of commercial flying experience with a mean of 15,420 flight hours (*SD* = 6,110). To encourage the production of verbalized assessment protocols without drawing on the *think-aloud protocol* (ERICSSON, SIMON, 1993), which practitioners frequently find unnatural, the pilots assessed performance in pairs.

2.2 Video scenarios

Each assessor pair saw two videotaped scenarios of a pilot-crew (captain and first officer) flying in an ATR 72 simulator. Table 1 outlines a summary of the two scenarios.

Video	Description
1	A captain (pilot-flying) and first officer (pilot-non-flying, pilot managing) conduct an instrument approach by day. They become visual close to the airport. During

	visual maneuvering to land, the aircraft encounters rain, which requires them to do a missed approach. The captain initially turns into the wrong direction, yet gets corrected by the first officer (duration: 6: 45).
2	The first officer (acting as pilot-flying) and captain (pilot managing) conduct an instrument approach. Poor weather forces them to fly a missed approach. Low fuel requires the crew to divert to the alternate airport. The captain attempts to convince the first officer to try a second approach before flying to the alternate. The first officer is reluctant (duration: 3:30).

Table 1: The video-scenarios used in the study.

The two pilots in the scenarios wore company uniform and took turns in acting as captain or first officer. Another employee acted as cabin crew. The videos were scripted in advance and recorded from the position of the flight examiner. The first camera captured both pilots from behind, whereas a second and third camera occasionally provided pertinent close-up pictures of the respective pilot's instrument displays. At several points during the flights, relevant airport landing charts (e.g., approach plate) were superimposed.

2.3 Procedure

The pilots watched the videos on a 52'' LCD TV monitor; they controlled the playback with a mouse on the table (Figure 2). The assessment sessions were video-recorded from three perspectives: CAM1 was positioned in front of the assessors, recording the assessors' gestures, faces, and interaction, together with the laptop that presented the scenario currently being watched. CAM2 provided a closer view of the assessors. In contrast, CAM3 recorded the area of work from above, capturing, for example, the notes taken, scores provided, and the pilots' pointing gestures toward certain word pictures on their assessment form.

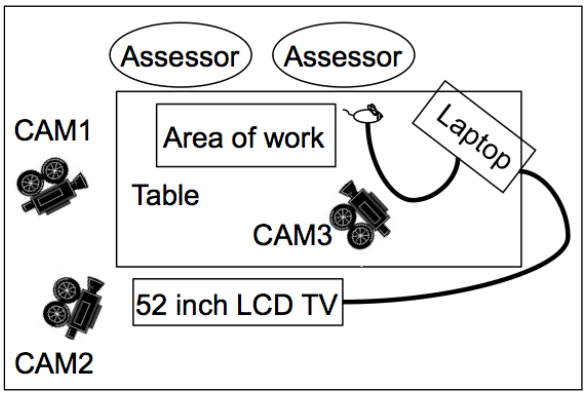

Figure 2: Study setting.

Assessors were given a booklet to take notes during the assessment process. Furthermore, each pair received two sheets from their company-training manual: the MAPP model (one for assessing the captain, the other for the first officer) and two assessment forms (word pictures) derived from the MAPP (Table 2). A "word-picture" is a performance marker within the MAPP assessment form (Table 2), made up by the MAPP category (SA, DM, AC, KN, MN, CM) and the assessment scale (1 to 5, where 1 denotes "unsatisfactory" and 5 "very good standard"). Examples for word pictures are "Correct procedure performed, with minor errors" and "Predicted future events and impact on flight safety." The former was one of the two word pictures of the category KN related to a score 3 ("satisfactory") whereas the latter represented one of the three word pictures in SA (score 4, "good standard"). A score 2 stands for "minimum standard or repeats required." To refer to these word pictures, the following conventions are used. "SA3," for instance, refers to a score 3 for the pilot's SA and "SA3.3" denotes the third word picture "Some difficulty predicting future events" in SA3. Within the participating airline, one 1 on any of the six criterion (SA, DM, AC, KN, MN, CM) means a "fail", making a repetition of the flight exam unavoidable. The same is the case when given three 2s (as i.e. for the captain assessed by CAP1 in scenario 1).

As part of their CRM training courses, all pilots had been introduced to the MAPP and had used the assessment form to rate the performances of other pilots. Each pilot had rated the performances of crews featured in at least 3 videos other than those they were rating for the present study. Those pilots working as training captains would also use the MAPP and related assessment form in the preparation of new employees.

	1	2	3	4	5
SA • Perception • Comprehension • Projection	• Lacked awareness of clearly obvious systems or environmental factors. • Misinterpreted or did not comprehend factors effecting flight safety. • Did not predict future events, even those obvious to flight safety.	• Missed some minor systems or environmental factors not critical to flight safety. • Comprehended some factors and implications on flight safety. • Difficulty predicting future events.	• Perceived significant systems or environmental factors affecting flight. • Comprehended significant factors and implication on flight safety with few errors. • Some difficulty predicting future events.	• Perceived all systems or environmental factors affecting flight. • Comprehended the implication of all factors. • Predicted future events and impact on flight safety.	• Perceived all systems or environmental factors, with an active approach to seeking further information. • Clearly comprehended the meaning of all factors. • Actively considered future events and impact on flight safety.
Knowledge • Facts • Procedures	• Unable to recall facts or made fundamental errors in their recall. • Correct procedure not identified, major deficiencies or excessive time taken in execution.	• Able to recall facts with some difficulty, some errors. • Displayed lack of familiarity with procedure; correct procedure selected but executed slowly or with errors.	• Recalled facts with only isolated errors. • Correct procedure performed, with minor errors.	• Easily recalled facts. • Correct procedure performed.	• Demonstrated an error free understanding of all facts. • Correct procedure selected and executed with certainty.

Note. SA=Situational Awareness, 1-5 = assessment scores (1 very poor performance, 5 very good performance).

Table 2: Excerpt of the assessment form, including the MAPP word-pictures (SA and KN)

2.4 Task

Participants were instructed to watch, discuss, and assess the performance of each pilot shown in the video. Assessor pairs received only one assessment form for each pilot. They were asked to come to an agreement about the ratings for the captain and first officer. Participants were instructed that the study aimed at understanding the reasoning behind their assessments of pilot performance. The researchers emphasized the importance of explicitly articulating the reasons and thoughts for a specific assessment. During the discussion, assessors were free to start, stop, pause, replay or go back to the video at any time they wanted. It was left to the assessors to decide where to start the discussion and how to assess performance. Assessors were asked to circle word pictures on the assessment form they deemed applicable to each pilot's performance. If they did not find evidence in the video for a certain aspect they were free not to circle the word picture at the corresponding level of the assessment form.

Two researchers were present during the assessment sessions. They used a fixed protocol to encourage assessors to speak up, clarify a comment when it was unclear (e.g., the pilot the assessors were talking about [such as, "Which pilot are you talking about? The captain or the first officer?" "Why did you circle this word picture? What evidence did you see?"]; why they circled a specific

word picture; to ask for the reasons for a specific rating), or provide reasons for a decision when these were not made explicit while assigning a score. After each rating exercise, assessor pairs were asked whether they had assessed all aspects of the performance as they perceived it, whether the instrument was missing a performance component, and whether they had felt constrained by the rating instrument.

2.5 Transcription

After data collection, the videos from the three cameras were combined into a single picture-in-picture video. The videos were transcribed word for word by a commercial transcribing service. Two of the authors checked and corrected all transcriptions in their entirety. Subsequently, several supplements were added to the transcripts: (a) the MAPP categories assessors were talking about (i.e. SA, KN); (b) which person (captain or first officer) the pairs were assessing; and (c) which "word picture" was currently talked about or pointed to by each of the assessors.

2.6 Coding and Data Analysis

In this study, we followed the precepts of the Grounded Theory Methodology (CHARMAZ, 2008; CORBIN, STRAUSS, 2008). Grounded theory is a way of creating a system of categories ("theory") that is directly "grounded" in the entire data set and provides a complete, objective (verifiable) description of the data. We applied Grounded Theory in order to arrive at emerging topics (see tables 4 and 5) stated by the assessor pairs.

Grounded Theory was implemented in the following way. To analyze the transcribed protocols, we employed the following procedure. We began by watching each assessment video, looking for, and highlighting, the assessors' key statements, subsequently called "criteria". For example, we coded the following speaking turn by the participant on the right as means of the codes 116 and 117:

> 01 R: So, if we get back to basic facts, I would have actually said, the Captain [Captain in the video] hadn't actually been able to actually recall what's written in the SOPs [Standard Operating Procedures] – which is quite clearly, you're not allowed to go below your 600 [kilogram fuel] unless you've got a payload issue before departure. (CAP1, p. 32)

This statement was summarized in two criteria:

- 116, The CP was not able to recall what is written in the SOPs (KN, CAP1, neg)
- 117, The SOPs are quite clear: you're not allowed to go below 600 kg fuel, unless there is a payload issue before departure (KN, CAP1, neg).

Each criterion was identified by a running number (e.g., 116), topic (e.g., KN), assessment pair (e.g., CAP1), and influence on the decision (e.g., neg). The criterion is the summarized statement made by an assessor pair (e.g. *The CP was not able to recall what is written in the SOPs*). The *topic* reflected whether the criterion was stated as such by the assessors in the discussion of SA or KN; the *assessment pair* (e.g., CAP1) pointed out which assessor-pair stated a certain criterion (CAP1, CAP2, or CAP3). *Influence* (i.e. *neg*) captured the assessment whether a criterion influenced the score provided in a positive (pos), negative (neg), or neutral (neu) way. Two researchers independently coded the transcriptions. Any disagreements were discussed until agreement was achieved.

The analysis of the criteria was conducted as follows. Initially, the list of criteria was printed out and a card made for each criterion. Then, criteria were grouped together that related to the same topic (a topic is made up by a range of criteria). For example, the criterion "135 The captain did make errors (KN, CAP3, neg)" and "88 I'd say it is a fundamental error. If there no payload issues, it was a fundamental error (KN, CAP1, neg)" were grouped together under the topic "Errors made." Each criterion could be related to more than one topic within the same scenario. This grouping process resulted in two schemes, one for each scenario. Both schemes included all criteria stated by the assessors, as well as the topics. With the help of these schemes, it was checked whether each criterion related to any other criteria or topics. By doing so, many additional connections between criteria and topics emerged. In a final step, the topics (that were made up by the criteria) were listed in two tables, one for each scenario (Tables 4 and 5) and sorted according to the assessor-pairs who stated them. Topics were either addressed by single pairs (CAP1, CAP2, CAP3), by two pairs (CAP1 & CAP2, CAP1 & CAP3, CAP2 & CAP3), or by all of the three pairs. These two lists were then used to conduct the analysis. Two authors independently coded the data. In the few instances of initial differences, the cases were discussed until agreement was reached.

3 Results

This study was designed to investigate the reasons (criteria) underlying pilots' assessments of the performance of other pilots that range from a scale of "unsatisfactory" to "very good standard." In the following, we present our findings under four aspects: (a) the overall assessment pass versus fail; (b) analysis of situations where the same reasons were used as evidence for different scores; (c) analysis of situations where different reasons led to the same score; and (d) comparisons of the assessment processes by assessment pair and by scenario.

3.1 Overall assessment pass versus fail

In multi-crew-operations, the pilot flying the aircraft (PF) makes certain manipulations and decisions that have to be cross-checked and acknowledged by the second, non-flying pilot (PNF). Whereas the captain retains full authority to the safety of the flight, PF and PNF duties are generally alternated between the captain and first officer. Whatever happens when there are no (strong) disagreements in the cockpit can be taken as evidence that both pilots are aligned and confident that the aircraft is not subject to any risk. However, this does not mean that peers evaluating the current performance would rate it as satisfactory with respect to the level of risk that the aircraft is exposed. In fact, our data suggest that there is considerable variance of how performances were rated. In both scenarios, one assessor pair gave the respective performances of the captain a passing grade, whereas two pairs assigned a failing grade (Table 3).

	Assessor pair	SA	DM	AC	KN	MN	CM	Pass/Fail
Scenario 1	CAP1	2	3	3	2	2	3	Fail
	CAP2	1	2	3	1	2	3	Fail
	CAP3	3	2	4	2	3	3	Pass
Scenario 2	CAP1	3	2	5	1.5	2	3	Fail
	CAP2	1	2	4	3	2	2	Fail
	CAP3	3	3	5	4	4	5	Pass

Note. CAP=Captain assessor pair, SA=Situational Awareness, DM=Decision Making, AC=Aircraft maintained within tolerances, KN=Knowledge, MN=Management, CM=Communication.

Table 3: The CAP-pairs' scores provided to assess the performance of the captain in scenario 1 and 2

In this situation, a "pass" reflected the fact that a performance maintained the aircraft in a safe state; a "fail" meant that the assessed pilot exhibited one or more behaviors that put the airplane at risk (unsafe performance). In the assessing pilots' company, a "fail" meant that the pilot had to repeat the assessment exercise. The data indicate that the three assessor pairs rated the same performance differently. The question this study was designed to address concerned the reasons for these differing assessments of the same performance. To gain a deeper, more detailed understanding of the variance, we focus on SA and KN in Scenarios 1 and 2, the performance aspects where lowest ratings were attributed (Table 4).

Topics stated by CAP 1	Topics stated by CAP 1&2	Topics stated by all pairs
- The missed approach as a risky area of flight	*- Asking about MDA*	*- Problems with the turn direction*
- Teamwork issues	*- Not predicting the go around*	*- Problems with knowledge*
		- Asking for information
	Topics stated by CAP 2&3	*- High workload*
Topics stated by CAP 2	*- Callouts not made*	*- Problems with the situational awareness*
- Deficient knowledge of the risks of a circling approach	*- Terrain*	*- Good initial approach and circling*
Topics stated by CAP 3	Topics stated by CAP 1&3	
- Low capacity to remember the key points due to high workload	*- Confusion*	
	- The missed approach and its procedure was messy	

Note. This table only provides an incomplete overview of the topics found in the analysis.

Table 4: Excerpt of the topics resulting from the assessor-pairs' statements in scenario 1

Out of a total of *N* = 41 topics addressed in Scenario 1 by all three pairs, only *n* = 6 (~15%) topics were shared. In Scenario 2, the total number of topics was N = 21 (~50% lower), whereas assessors concurred with each other on *n* = 7 topics (33%). In both scenarios, pairs addressed similar numbers of topics (Scenario 1: *n* = 23, 22, and 21 for CAP1, CAP2, and CAP3, respectively [*SD* = 1.0]; Scenario 2: *n* = 13, 15, and 12 for the three pairs, respectively [*SD* = 1.5]). For example, CAP2 and CAP3 stated in the discussions of Scenario 1 the following criteria (Table 4): "the *terrain* surrounding the destination airport" and "not making the correct *calls* during the missed approach." The latter topic "Callouts not made," for instance, is made up by the following two criteria: "No go around/flap15-call was made" (CAP2), and "A power up/flap15-call was never made" (CAP3). In total, pairs stated 79 criteria in Scenario 1, whereas 59 criteria in Scenario 2.

Do assessor pairs identify the same criteria? To answer this question, those aspects that some pairs rated as leading to unacceptable risk were investigated. Comparing the scores (Table 3) with the lists of topics (Tables 4 and 5), it is found that in some cases pairs state the same reasons (criteria/topics) but arrive at a different score. On the other hand, pairs also note different reasons but assessing the performance by the same scores. In other words, even if pairs come to the same conclusion that the captain's performance is a fail, they do so for different reasons. These findings are examined in the following two subsections.

3.2 Same reasons, different score

In regards to SA and KN, the same scores were attributed by CAP1 and CAP3 in the discussion of the KN dimension concerning Scenario 1 (both pairs attributed a score of 2 [Table 3]) and SA in Scenario 2 (score 3). In most instances, however, the assessor scores varied. Yet despite disagreement in terms of the quality of the performance, the pairs often noted the same or similar reasons. For example, the topic "problems with the turn direction" (Table 4), which was addressed in the SA discussion of Scenario 1, the pairs made the same observations but disagreed about the quality of the performance. In Scenario 1, the crew had to engage a missed approach procedure, because the aircraft entered a cloud during the base turn while flying a circling approach at the destination airport. Initially, the captain initiated a right turn. The first officer immediately picked up on the captain's mistake and instructed him turn to the left. The issue of turning the wrong way entered the discussions of all pairs. Yet the quality of the performance was rated differently. CAP1 noted (leading to SA score = 2):

> 01 R: And obviously, when it got to the missed approach, they both got a wee bit overloaded there. There was confusion with regards to the turn direction and the procedure went out of the window largely.[2]

In this situation, the assessor highlighted for his partner the apparent confusion that was expressed in the captain's actions. The pair ended up qualifying the performance as less than satisfactory. In contrast, CAP3 addressed the difficulties with the turn direction as follows (SA score = 3 [satisfactory]):

> 01 R: And then when they got onto that left base, it was a bit of a surprise; well, they both seemed a little bit surprised when they went IMC [instrument meteorological conditions].
> 02 L: Yes.
> 03 R: A bit taken aback by it. So that led to some confusion because they had perhaps already got this, the captain didn't have a plan of what to do, but the [FO] did.
> 04 L: Yeah, I think they both had an idea of what to do as far as a missed approach is concerned, *but it was just the execution, turning left or right*.
> 05 R: [The FO] seemed a little bit more in tune with the situational awareness perhaps than the captain. Because the captain was flying the airplane—busy. Whereas the [FO], he's monitoring role. So he's got a bit more capacity there to remember key points.

As this transcript shows, the confusion was noted (turn 03), but the assessment of it appeared mediated by the fact that the pilots appeared to have "had an idea of what to do as far as a missed approach is concerned," but the execution was problematic. That is, whereas in the discussion of CAP1 the same video material provided evidence that there was confusion, CAP3 suggests that the pilots

[2] In the transcriptions, "R" and "L" refer to the assessor captain seen on the right and left, respectively, from the perspective of the viewer.

knew what they did but merely performed the procedure poorly. As a result, CAP3 immediately concluded that the captain's performance was satisfactory and rated it a 3 for SA. CAP2R spoke about the captain's SA right after having watched scenario 1 (SA score = 1).

> 01 R: It seemed to me that if I had to critique the flying pilot, there was a couple of areas there in his situational awareness that I think were some pretty big gaping holes. And that is that he—the two things on this approach as the pilot flying that you should know, what altitude is our MDA [minimum descent altitude], and what way are we going to turn in the missed approach. He was unaware of the altitude they were descending to, and had to ask what minimum are we descending to? And then the confusion there when he carried out, he had to enquire as to the correct way to turn on the missed approach. Yeah, I felt they were two sort of gaping holes in his situational awareness.

The assessor pair CAP2 discussed the fact that the captain was not performing standard operating procedures. He was obviously surprised by having to fly a missed approach, which is a consequence of inappropriately assessing the observed poor weather conditions. They concluded that the pilot had difficulties "predicting future events and implications on flight safety." CAP2 immediately agreed that the performance was such that the pilot should be failed (SA score = 1).

When the issue of the turn direction is investigated more deeply, one can find additional differences between the reasons. Related to the importance of turning the correct way during the missed approach was the consideration of the high terrain that surrounded the destination airport. All pilots in the company are aware of the fact that there had been a major accident at this airport where a wrong missed approach procedure had led to a crash in the mountains to the right-hand side of the approach.[3] The danger of the terrain did not show up in the assessment of CAP1 (Table 4), but was explicitly articulated as a reason by CAP2 (SA score = 1) and CAP3 (SA score = 3). CAP2 noted that the captain had "missed what way to turn on the missed approach," which is critical "With high terrain. On that particular approach there is high terrain." The assessor concluded that therefore there is "a necessity to turn the correct way".

Right from the beginning, CAP2 tended towards a very low SA score, yet it followed an extensive discussion before they arrived at a decision: "On a difficult, bad weather, circling approach into an area where there's high terrain, the fact that you can't recall what you were going to descend to, what way you were going to turn on the missed approach, definitely puts [the captain] into that one [SA score = 1] category." The second assessor agreed. CAP2's final statement shows the importance the topics "turn direction" and "high terrain" had for this assessor pair. CAP2 spoke about, and thus

[3] To maintain confidentiality, no additional identifying information is provided about the airport.

knew, that a score of 1 in any of the performance components would immediately lead to a failing rating, which in turn makes a repeat of the whole simulator session unavoidable. Here, "not turning the correct way in an area with high terrain" is treated as a non-compensatory criterion (e.g., EINHORN, 1972): no matter how good the captain performed previously or following this situation, he cannot compensate for his behavior with excellence in another area of his performance.

Whereas the assessment criterion "not turning the correct way in an area with high terrain" led to a score of 1 for CAP2, it was not regarded as severe by CAP3. Despite making the same observations (high terrain; turn direction), CAP3 rated the SA performance component as "satisfactory" (SA score = 3). This exemplifies our observation that for any pair of assessors, a specific criterion does not have to be non-compensatory. The same observation is weighted differently. This is in accordance with the view of the third step in the process of making performance judgments: having (a) *observed* relevant work-related behavior and (b) *evaluated* each behavior in terms of its effectiveness, assessors (c) *weight* the observed behavior (BORMAN, 1978). Differences may result from unequal weights attributed. Whereas a rater might assess specific behavior to be critical, another may deem it less important in the overall scheme and even irrelevant. In our example, CAP3 gave less weight to the "turn direction in an area with high terrain" than CAP2 and, therefore, rated it less severe in terms of having an impact on the safety of the aircraft.

3.3 Same scores, different reasons

In the preceding section, descriptions are provided of how the assessors come to different evaluations of the flying pilot's performance despite stating the same reasons. The data also reveal the opposite: assessments result in the same scores although the assessors state quite different reasons. In the discussion of SA in scenario 2, for example, CAP1 and CAP3 both rate the performance as "satisfactory" in the dimension of SA (score = 3, Table 3); CAP2 deemed the quality of the performance "unsatisfactory" (score = 1). Related to a topic that we outline in the following paragraphs, CAP1 and CAP2 addressed the same topic but came to opposing conclusions. In contrast, this topic did not appear in the deliberations of CAP3. Table 5 gives an overview of several topics stated by the pairs in regards to Scenario 2. In this scenario, the crew conducted an instrument approach at the destination airport. Not becoming visual with the runway at a low altitude forced them to conduct a missed approach. Related to SA in this situation was the discussion if the captain was aware of, and expected, the missed approach. Whereas this point was not addressed by CAP3, it was exclusively discussed by CAP1: "Well, [the crew] was pretty aware of the fact that there was a possible missed approach. [The captain] didn't really have difficulty predicting future events, because they knew that the missed approach was possible at [the destination airport]." Following their deliberation, CAP1 came to the conclusion that the captain's performance was satisfactory (SA score = 3).

Topics stated by CAP 1	Topics stated by CAP 1&2	Topics stated by all pairs
- KN in general was good	*- Inappropriate position of discussing fuel issues*	*- The captain's lack of knowledge*
- Being aware of, and expecting a missed approach	*- Payload issues*	*- Lack of information available to the assessors*
Topics stated by CAP 2	Topics stated by CAP 2&3	*- Headwind on the way to the alternate*
- The missed approach as a surprise	*- Questioning whether the captain had insufficient knowledge or knew what was allowed but tried to "screw the scrum"*	*- Weather at the alternate airport*
- The planning was inadequate		*- Questioning the captain's dealing with risk and uncertainty*
- The gravity of the low fuel state		*- Issues related to Standard Operation Procedures (SOPs)*
Topics stated by CAP 3	Topics stated by CAP 1&3	*- The captain's handling of the (low) fuel*
- The captain's offering of an alternative	*- The captain made errors*	
- Doubt about the appropriateness of suggesting a 2nd approach		

Note. This table only provides an incomplete overview of the topics found in the analysis.

Table 5: Excerpt of the topics resulting from the assessor-pairs' statements in scenario 2

CAP2 came to a different conclusion: the missed approach must have been a surprise to the crew, which the assessor pair saw as an indicator that the captain had problems with SA:

> 01 R: If what the first officer said was correct and the weather at [the alternate] wasn't flash [good], and they had a head wind to go home, then [the captain] wasn't perceiving.
> 02 L: And the fact that they left it so late to discuss their plan: they can't have perceived, they can't have predicted that they were going to have to go around at [the destination airport]. So, it wasn't very good.

Having watched the video again, CAP2 further affirmed that the crew did not mention possible fuel problems: "If that's their first comment, about the fuel, 'We haven't got much fuel,' it implies that they haven't discussed it." But, so the assessor pair, they ought to have discussed the fuel issue. One of the two then summarizes:

> 05 R: So clearly, the miss[ed approach] has been a surprise to them. And therefore there are some issues with their SA in terms of the big picture. They haven't predicted the future event, even those obvious to flight safety. They didn't predict they weren't going to get in [at the destination airport], which you can't always do. But the fact that on min[imum] gas, with an alternate nominated that wasn't flash, they don't seem too worried about not getting in at [the destination airport].

CAP2 subsequently critiqued the captain for not being sufficiently worried about the weather and for not having developed a plan. This was deemed inacceptable on bad weather days: The assessed pilot "hadn't worked out a divert plan, when they were clearly on min[imum] fuel. And they mentioned that their alternate wasn't flash either, and they had a headwind to get there." The assessors then came to the conclusion that the performance was unsatisfactory with respect to situation awareness (SA score = 1).

CAP3 did not speak about whether the crew predicted the missed approach or if it was a surprise to them. Instead, they discussed whether the captain was offering an alternative (Table 5) by suggesting going around and trying a second approach at the destination airport. One of the assessors suggested that the captain was not really surprised about the weather situation but that he was possibly offering an alternative way of looking at the situation that they were in. The assessed pilot was merely "pushing the risk envelope a bit too much." The assessor then elaborated that they needed to err on the side of caution, in favor rather than against the pilot to be assessed.

> 01 L: Well you see, you don't really know. But I would err on the side of caution. It's difficult to tell whether he's just throwing—admittedly [the captain] did mention it a couple of times. As if he was trying to, not manipulate, but just push his suggestions in a little bit further as "don't quite give up yet. We'll have another go." But the more time you spend [around the destination airport], the less you've got back [at the alternate]. And nothing's ever given. There's nothing there to say that the weather's going to be perfect back at your alternate airfield. It could go either way, as we've found out in the past. It's just erring on the side of caution.

The second rater subsequently noted his reasons for tending towards a satisfactory rating rather than for a "good" (score = 4) by suggesting that the captain clearly should have known about the possible risks that arose from the weather situation. The two assessors then concluded that the performance was satisfactory.

These exemplary extracts from the deliberations illustrate that assessors make different interpretations of the same pilot performance. Despite assessing the same scenario, they note different facts. Some find evidence that the crew in Scenario 2 expected the missed approach, whereas other assessors conclude that the weather was a surprise or that the captain was simply suggesting an alternative of looking at the situation. That is, the same assessment concerning the quality of the performance—i.e., the same score—does not imply the congruence of the assessors' reasons: the same scores are sometimes based on entirely different reasons and aspects of the scenario.

3.4 Comparison of the assessment processes between the pairs and between the scenarios

The process of assessing pilots' performance was not only very similar in relation to the two scenarios, but also between the assessor pairs. Having watched the video in its entity, all pairs had a more or less extended introductory discussion without using any of the categories of the MAPP or the associated assessment form. The groups discussed their observations, drawing on the notes they had taken during the first viewing of each scenario. The pairs then started to use the assessment form. Assessors read certain word pictures, discussed each, and then sought to fit their observations to one of the descriptions of the assessment form. By doing so, they circled the specific word pictures belonging to a category (e.g., SA or CM), which were used later on to decide the overall-score. During their discussion, assessors referred to observations they made and provided examples. In case they disagreed, the pair had a discussion until agreement was achieved or replayed parts of the scenario.

Not all of the pairs used the assessment form in the same way. Both CAP2 and CAP3 first marked the essential skills (SA, DM, AC), followed by the enabling skills (KN, MN, CM). CAP1, on the other hand, initially did not provide a SA score for Scenario 1. They moved to other performance components before returning to discuss SA. In the end, CAP1 had first rated those performance components related to the enabling skills before rating those pertaining to the essential skills.

Great similarity was found in the way assessors approached the assessment. All pairs first assessed the performance of the captain before turning to the first officer. They all spent most of the time to finish the assessment of the captain in Scenario 1. Subsequently, less time was required to assess both the first officer and Scenario 2.

Not all of the captains collaborated with each other in the same way. CAP1 and CAP2 both assessed each pilot together, one after the other. This was different for CAP3 (e.g., SA in Scenario 2): One assessor circled the word pictures related to the captain, whereas the other did the same for the first officer. This technique seemed to negatively influence the discussion in terms of coming to a conclusion about the scores, because each assessor was focused on the pilot he assessed. In Scenario 2, for example, one of the CAP3 assessors continually made suggestions about a SA score for the captain. But the other, in his response, did not take up the offer for assigning a specific performance rating.

When starting to use the assessment form, pairs read the word-pictures (Table 2) in a horizontal way: in terms of the KN element "procedures," for instance, they went through the various word pictures, such as "Recalled facts with only isolated errors" (KN3.1) or "Easily recalled facts" (KN4.1). However, assessors not necessarily considered all the word-pictures related to a certain MAPP element, especially when they became more familiar with the assessment form. Instead, when there was evidence that the performance was poor, they immediately addressed lower scores, without consider-

ing higher ones. These findings are mirrored in Table 6, which summarizes the word pictures addressed by each pair.

		CAP1	CAP2	CAP3
Scenario 1	SA	1, **2**, 3, 4	**1**, 2	2, **3**
	KN	1, **2**, 3	**1**, 2	**2**
Scenario 2	SA	2, **3**, 4	**1**, 2, 3	**3**, 4
	KN	1, **1.5**, 2	2, **3**, 4	**4**

Note. The highlighted scores represent the overall marks provided to the captain being assessed

Table 6: MAPP word-pictures addressed by the CAP-pairs in scenario 1 and 2

Table 6 shows that pairs addressed unequal numbers of word pictures. On average, CAP1 debated about 3 word pictures, CAP2 about 2.5, and CAP3 about 1.5. The pairs did not discuss all of the five word pictures in relation to a MAPP element or take a score 5 into consideration for the captain. Whereas CAP1 addressed up to four word pictures (SA, Scenario 1), CAP3 twice spoke about a single word picture (KN, Scenarios 1 and 2). CAP3 never took more than two word pictures into account. When the scores provided (Table 3) are contrasted with the word pictures addressed (Table 6), it can be found that the assessor pairs applied different techniques to come to a conclusion about a score: CAP1 and CAP2 always "backed up" the scores they provided by considering word pictures above and below a certain score. For example, when CAP2 concluded that the situation awareness was unsatisfactory in Scenario 2 (score = 1), they checked the two word pictures to the right in the matrix; when they decided to rate KN as satisfactory (KN score = 3), they questioned their decision by considering the word pictures on the left (KN2) and right (KN4) in the assessment form. This was different for CAP3, where the assessors did not consider the performance in light of the word pictures to the left and right of the one they discussed.

There were a number of topics that bore on the decisions in both scenarios. There were 7 themes that were stated in relation to both scenarios. All assessor pairs extensively addressed *situation awareness* and aircraft *knowledge*. Furthermore, they all spoke about *threats and risks*. In Scenario 1, for example, CAP1 described the missed approach as a risky area of flight, and CAP2 critiqued the captain's deficient knowledge of the risk associated with a circling approach. In Scenario 2, all pairs noted how the crew dealt with risks and uncertainty, and CAP2 emphasized the gravity of the low fuel state. Another topic that was shared between the scenarios was how the crew dealt with *anticipating future events*. CAP2 underlined the importance of looking out to see whether they would go IMC

(Scenario 1), and CAP1 questioned if the crew expected a missed approach (Scenario 2). In both scenarios it was mentioned that *errors were made* during the missed approach. Another topic was the crew's *surprise* when going IMC and its consequences (CAP3, Scenario 1). This theme also emerged in CAP2's discussion of the evident surprise observable in Scenario 2. Finally, the CAP pairs also noted difficulties related to *planning*.

The majority of topics applied to the scenarios were largely different. When the number of topics shared between the two scenarios (7 topics) is compared with the total number of topics addressed in each scenario (S1: 41 topics, ~17%; S2: 21 topics, ~33%) it is found that assessors largely apply a different set of topics to each scenario. Whereas a smaller number of topics were shared in both scenarios, the majority of topics are dissimilar.

4 Discussion

This study was designed to better understand how airline captains assess the performances of peers in situations that compromise the safety of the aircraft. The aim of the study was to find the criteria (or combination of criteria) that assessors identify as affecting flight safety and to reveal whether all pairs identified the same criteria that for some assessors lead to unacceptable risk. Given the variations in the nature of the assessment processes and assessment outcomes, it may be hypothesized that in the Air India example, too, assessors—without knowledge of the final outcome—might have arrived at different ratings of crew performance. Thus, for example, even if some level of risk had been identified, a qualification of the PF as being a "risk taker" (made by CAP3) might lead to rating the quality of a performance as "satisfactory." On the other hand, other pilots might have judged the performance as unacceptable and, therefore, as unsatisfactory and considered it an automatic "fail." That is, a criterion deemed non-compensatory and critical to overall safety by one assessor (pair) may have been deemed compensatory and less or not critical to safety by another. Despite addressing the same criteria, they might have been weighted in different ways. In addition, the data presented show that raters do not make the same observations; and the observations they make enter and weigh on the final assessment in varying ways.

4.1 Variability in assessment

The present data show considerable differences between CAP pairs: not only in their assessment of specific skillsets (SA, KN, etc.) but also in their overall ratings (pass/fail). Some regarded performance to be safe (pass) while others saw the aircraft at risk (fail). Possible reasons for such rater differences are addressed in the literature. Evaluating performance is assumed to include three steps (BORMAN, 1978): observation, evaluation, and weighting. In the first step, assessors have to observe

work-related behavior. Disagreement between raters is said to result from differences in the observations made. This is in accordance with the present findings: not all the pair made the same observations, yet some were similar. Some pairs stop after having found evidence, while others continue to look out for more critical behavior. In a second step, assessors have to evaluate the observed behavior in terms of its effectiveness (BORMAN, 1978). Because this step is rather made implicitly and not very systematic, it often is a source of disagreement. We observed different evaluations of observations in our data: some pairs regarded the captain to be totally uninformed about the SOPs, while others reached the conclusion that he exactly knew what was going on, but just attempted to "screw the scrum a little bit" (meaning: not exactly stick to the procedures, CAP2). The same observations were thus evaluated in different ways. The third step is to weight criteria (i.e. BORMAN, 1978; EINHORN, 1974; SLOVIC & LICHTENSTEIN, 1971). This weighting can result in the observed differences between assessors (BORMAN, 1978). An incident deemed critical by one assessor (pair) might not be assessed in the same or similar way by another. This, in turn, is consistent with our findings: being unaware of the turn direction, for instance, in an area where there is high terrain was weighted differently.

The data demonstrate that there is no consistent correlation between topics and assessment. Addressing a specific criterion did not necessarily lead to a certain score. Based on the present data set, it may be concluded that the performance assessment is not a linear, additive process (e.g., DAWES, CORRIGAN, 1974). Assessors do not compile a finite number of criteria to arrive at an overall decision (EINHORN, 1971; DIECKMANN, DIPPOLD, DIETRICH, 2009). In the present database, assessors do not bring to bear the same observations. Instead, the interpretation of the same criteria may vary considerably. Our data support the contention that assessors combine criteria in a nonlinear, nonadditive way (e.g., EINHORN, 1970, 1972; BRANNICK, BRANNICK, 1989). They make certain observations, combine information in different ways, apply compensational and noncompensational criteria (even to the same criteria), search for further information to back up their opinion, or stop after having found some critical evidence.

4.2 Inter-rater reliability

In this study, considerable differences are observed between assessments of the same performance. Such differences have already been reported in the literature (e.g., SMITH, NIEMCZYK, MCCURRY, 2008) and have considerable impact when they occur in flight examiner-examinee debriefing sessions, where assessments bear on the lives of the assessed pilots. To reduce disagreement between assessors and increase agreement within assessment scores, the literature provides a series of approaches (e.g., BAKER, MALQUEEN, DISMUKES, 1999; WOEHR, HUFFCUTT, 1994), such as rater-error training (RET), performance-dimension training (PDT), frame-of-reference (FOR) training,

and behavioral-observation training (BOT). All of these techniques aim at increasing agreement between assessors, and with this, inter-rater reliability (IRR). IRR measures (such as the Pearson product-moment correlation, e.g., GOLDSMITH, JOHNSON, 2002) are widely used to assess the health of the aviation assessment-system and are usually based on correlations of assessment-scores. A high IRR-score indicates agreement between raters. However, this study shows that the reasoning behind the same scores can largely vary between assessors. The same scores are given for different reasons, and vice versa. Consequently, high IRR-scores do not imply that assessors made the same observations and assessed the same phenomena. A score of 3 for one assessor pair might mean something different for another pair. Despite scores being same, the underlying reasons may be different. This study thus suggests that great care is to be taken when considering IRR as a means of measuring agreement between raters.

4.3 Strive against, or value diversity between assessments?

The literature regards FOR training to be the most effective strategy to improve IRR (BAKER et al., 1999; see WOEHR, HUFFCUTT, 1994 for closer explanation of FOR-training). The results of this study suggest that there is an increased value in BOT because assessors make different observations and isolate different facts. Training assessors on how to improve their observational skills might thus be of value to achieve increased agreement between raters.

It is certainly important to achieve agreement among assessors about whether the performance of a pilot was safe. Yet any scenario to be assessed might be perceived in different ways, not unlike the Mangalore accident mentioned in the introduction. An important question is whether the reduction of variance does justice to the various ways in which a scenario can be interpreted and in which learning may be derived from it for the pilots concerned. The results of this study show that in two scenarios only 7 topics were shared (out of 41 topics in Scenario 1 and 21 topics in Scenario 2). This, in turn, illustrates the extent to which scenarios are unique, diverse, and lend themselves to different observations. It may thus remain difficult, or even impossible, to eliminate variance in observations. Future research is required to study whether diversity in observations is a source for engaging the pilots involved in reflection on action and, in the process, improve performance.

5 Conclusions and future research

The present study focused on captain assessor-pairs. It was noticed that CAP pairs watched the scenario, had a general discussion, and then tried to fit their observations in the assessment form. There is a need to investigate how airline professionals of other experience and seniority levels (such as flight examiners and first officers) assess performance. By doing so, future studies need to investi-

gate whether more or less experienced pilots apply a similar technique, make the same observations, and if the assessment-process follows a similar scheme in all of the three groups.

This study used an assessment form for evaluating performance. Yet providing assessors with a range of word pictures might already have influenced the assessment process. Future studies should investigate how assessments come about when assessors are not given any model or assessment form. This may require assessors to fall back on their own knowledge and assumptions. Possible research questions might pertain to the sets of criteria addressed and to the level of agreement achieved.

The data analyzed is qualitative in nature. An approach inspired by grounded theory was applied to extract the criteria from the discourse of assessor pairs and to arrive at the topics addressed. It goes without saying that the small sample size of three assessor pairs prevents a generalization from the sample to the population. A closer look is needed on how a larger number of airline professionals, namely flight examiners, assess performance. The analysis of the captain pairs, however, provided valuable insight into their reasoning, pointing towards large variance in both the scores and reasons.

Reliably and equitably assessing pilot performance in modern multi-crew cockpits has proven difficult. The present study traces how assessor pairs come to different assessments about a performance component, such as situational awareness or knowledge. Different assessments of the same performance were found even when assessors used the same reasons and same assessments may be reached for very different reasons. This questions the value of IRR measures currently used in the industry.

Acknowledgments

This research was supported by a grant from Griffith University with additional funding coming from the partner airline. The views expressed here are those of the authors. We thank both the participating airline and the participants for their time and support.

References

Baker, D., Malqueen, C., & Dismukes, R. (1999). Training pilot instructors to assess CRM: The utility of frame-of-reference (FOR) training. In R. Jensen (Ed.), *Proceedings of the 10th International Symposium on Aviation Psychology* (pp. 291-300). Columbus, OH: Ohio State University.

Borman, W. C. (1978). Exploring the upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology, 63*(2), 135-144.

Brannick, M. T., & Brannick, J. P. (1989). Nonlinear and noncompensatory processes in performance evaluation. *Organizational Behavior and Human Decision Processes, 44*(1), 97-122.

Charmaz, K. (2008). Grounded Theory. In J. A. Smith (Ed.), *Qualitative psychology: A practical guide to research methods* (pp. 81-110). London, UK: SAGE.

Cooper, G. E., White, M. D., & Lauber, J. K. (1979). *Resource management on the flight deck*. (NASA Conference Publication 2120). Moffet Field, CA: NASA - Ames Research Center.

Cooper, G. E., White, M. D., & Lauber, J. K. (1980). *Resource managment on the flight deck: Proceedings of a NASA/Industry Workshop*. (NASA CP-2120). Moffet Field, CA: NASA Ames Research Center.

Corbin, J., & Strauss, A. L. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (3rd ed.). Los Angeles, LA: SAGE.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*(2), 95-106.

Dieckmann, A., Dippold, K., & Dietrich, H. (2009). Compensatory versus noncompensatory models for predicting consumer preferences. *Judgment and Decision Making, 4*(3), 200-213.

Einhorn, H. J. (1970). The use of nonlinear, noncompensatory models in decision making. *Psychological Bulletin, 73*(3), 221-230.

Einhorn, H. J. (1971). Use of nonlinear, noncompensatory models as a function of task and amount of information. *Organizational Behavior and Human Performance, 6*(1), 1-27.

Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance, 7*(1), 86-106.

Einhorn, H. J. (1974). Expert judgment: Some necessary conditions and an example. *Journal of Applied Psychology, 59*(5), 562-571.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (revised ed.). Cambridge, Mass: MIT Press.

Flin, R., Martin, L., Goeters, K. M., Hörmann, H. J., Amalberti, R., Valot, C., & Nijhuis, H. (2003). Development of the NOTECHS (non-technical skills) system for assessing pilots' CRM skills. *Human Factors and Aerospace Safety, 3*(2), 95-117.

Flin, R., O'Connor, P., & Mearns, K. (2002). Crew resource management: Improving team work in high reliability industries. *Team Performance Management, 8*(3/4), 68-78.

Goldsmith, T. E., & Johnson, P. J. (2002). Assessing and Improving Evaluation of Aircrew Performance. *The International Journal of Aviation Psychology, 12*(3), 223-240.

Helmreich, R. L., & Foushee, H. C. (1993). Why crew resource management? Empirical and theoretical bases of human factors training in aviation. In E. Wiener, B. Kanki & R. Helmreich (Eds.), *Cockpit Resource Management* (pp. 3-45). San Diego, CA: Academic Press.

Helmreich, R. L., Merritt, A. C., & Wilhelm, J. A. (1999). The evolution of crew resource management training in commercial aviation. *International Journal of Aviation Psychology, 9*(1), 19-32.

Helmreich, R. L., Musson, D. M., & Sexton, J. B. (2004). Human factors and safety in surgery. In B. M. Manuel & P. F. Nora (Eds.), *Surgical patient safety: Essential information for surgeons in today's environment* (pp. 5-18). Chicago, IL: American College of Surgeons.

Mavin, T. (2010). *Assessing pilots' performance for promotion to airline captain.* Unpublished doctoral dissertation, University of Queensland, Brisbane, Australia.

Mavin, T., & Dall'Alba, G. (2010, April). *A model for integrating technical skills and NTS in assessing pilots' performance.* Paper presented at the 9th International Symposium of the Australian Aviation Psychology Association, Sydney, Australia.

Mavin, T., Roth, W.-M., & Dekker, S. W. A. (2012). *Should we turn all airline pilots into examiners? The potential that evaluating other pilots' performance has for improving practice.* Paper presented at the 30th EAAP Conference, Villasimius, Sardinia, Italy.

Ministry of Civil Aviation (2010). *Factual Report: Air India Express, Boeing 737-800, VT-AXV, Mangalore, 22 May 2010. New Delhi: MCA.*

O'Connor, P., Hörmann, H. J., Flin, R., Lodge, M., Goeters, K. M., & JARTEL Group. (2002). Developing a method for evaluating crew resource management skills: A European perspective. *International Journal of Aviation Psychology, 12*(3), 263-285.

Orlady, H. W., & Orlady, L. M. (1999). *Human factors in multi-crew flight operations*. Burlington, VT: Ashgate.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6*(6), 649-744.

Smith, M. V., Niemczyk, M. C., & McCurry, W. K. (2008). Improving scoring consistency of flight performance through inter-rater reliability analyses. *Collegiate Aviation Review,* Fall 2008.

Weick, K. E. (1990). The vulnerable system: An analysis of the Tenerife air disaster. *Journal of Management, 16*(3), 571-593.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*(3), 189-205.