Good reasons for high variability (low inter-rater reliability) in performance assessment: toward a fuzzy logic model

Wolff-Michael Roth\textsuperscript{a,c,\*}, Timothy J Mavin\textsuperscript{b,c}, Ian Munro\textsuperscript{d}

\textsuperscript{a}Applied Cognitive Science, Faculty of Education, MacLaurin Building A557, University of Victoria, Victoria, BC, V8W 3N4, Canada
\textsuperscript{b}School of Biomolecular and Physical Sciences, N44 3.24 Griffith University, Nathan QLD 4111, Australia
\textsuperscript{c}Griffith Institute of Educational Research, Griffith University
\textsuperscript{d}Mt. Cook Airlines, Christchurch, New Zealand

Abstract

Regular performance assessment is an integral part of (high-) risk industries. Past research shows, however, that in many fields, inter-rater reliabilities tend to be moderate to low. This study was designed to investigate the variability of performance assessment in a naturalistic setting in aviation. A modified think-aloud protocol was used as research design to investigate the reasoning pairs of pilots use to assess the performance of an airline captain in a high-risk situation. Standard protocol analysis and interaction analysis methods were employed in the analysis of transcribed verbal protocols. The analyses confirm high variability in performance assessment and reveal the good, albeit fuzzy, justifications that assessor pairs use to ground their assessments. A fuzzy logic model exhibits a good approximation between predicted and actual ratings. Implications for the practice of performance assessment are provided.

Relevance to industry

Given that a low performance assessment can lead to re-examination and change in employment status, many industries aim at achieving consistency in identifying true performance levels. In view of the complexity of flying a modern aircraft, variability in performance assessment may be the norm and high inter-rater reliability may never be achievable. However, if the variability in performance assessment is a real phenomenon, as reported here, then practitioners and research might have to test whether it can be used positively (e.g., as opportunity for fruitful discussions during training situations that improve the resilience of flight crews).

Keywords: Performance assessment; High-risk industry; Fuzzy logic model; Inter-rater reliability; Think-aloud protocol; Aviation

*Corresponding author. Tel.: +1 250 721 7764; fax: +1 250 721 7598.
E-mail address: mroth@uvic.ca (W.-M. Roth)
1. Introduction

In high-risk industries, such as in aviation, it is crucial that employees work continuously at maintaining required performance levels of staff to guarantee the safety of customers, bystanders, employees, and environment alike. As disasters in the maritime industry, aviation, and medicine show, human errors, often arising from lack of proper training and low competency levels, are at the source of many serious accidents (Emad & Roth, 2008; Helmreich et al., 2004). Continued examinations, tests, and performance assessments are integral elements to “help ensure that all requisite skills and knowledge are included, while reducing the possibility that an operator will be required to demonstrate skills or knowledge that are not necessary to perform the job” (Nuclear Energy Agency, 1988, ¶6). Importantly, as the International Civil Aviation Organization recently recognized, inter-rater reliability is an important component of achieving these requisites (ICAO, 2007). However, meaningful criteria for consistently assessing performance appear to be elusive (Rigner & Dekker, 2000). There is evidence of considerable variability in performance assessment when the same pilot performance segments are assessed by a large number of experts, even when these all derive from the same company and, therefore, have a common background in performance assessment (as assessing or assessed individual) (Mavin et al., 2013). On the one hand, such variability creates a problem because operators (here, pilots) are not consistently assessed. This means that lower-performing individuals could pass because of the specifics of the situation and generally higher-performing individuals might fail. On the other hand, as the title of this article suggests, there might be some good reasons for performance assessments to vary (widely). The present study was designed to investigate the justifications underlying performance assessment ratings for the purpose of better understanding the high variability in performance assessment conducted in naturalistic settings (e.g., Goevarts et al., 2011; Mavin et al., 2013). The anticipated outcomes of the study were possible applications of assessment variability in the training of pilots.

2. Background: performance assessment and inter-rater reliability

To the present day, performance assessment in some fields is made in global terms, such as when the suitability of a junior surgeon was decided, until recently, based on the recommendations of one or more senior surgeons (Schijven & Bemelman, 2011). As there tend to be legal ramifications in the case of accidents, especially in high-risk domains, proper training and assessment of competencies and performances are necessary to ascertain that the employer has done everything to guarantee the safety of clients specifically and the public and the non-human environment generally (e.g., MacDonald & Sulsky, 2009). Detailed performance criteria that are to be assessed using checklists, rating scales, and rubrics pertain to the widespread, standard practices in educational settings around the world. The measurement instruments themselves are not only to be used by assessors, but also should be intimately familiar to the assessed as well—(a) for them to understand
how they are assessed and (b) with the aim of developing self-assessment strategies (e.g., Government of Alberta, 2008). Yet all such instruments are based on judgments, a situation that leads to the fact that even simple observational checklists requiring the correct identification of types of tasks give rise to variability due to (a) different codes for the same actions and (b) the same codes for different actions (Horng et al., 2010). Furthermore, in an attempt to improve construct validity and inter-rater reliability, assessment is sometimes segmented into smaller focus areas with checklists used to direct the assessor into regions of importance, as is the case with some simulation exercises for surgical assessment. However, this approach is “divorcing technical and decision-making skills, compartmentalizing the various facets of a mature surgeon with no guarantee that the sum of the parts is equal to the whole” (Bodle et al., 2008, p. 212). At least one study suggests that aviation practitioners found the separation into technical and non-technical skills not only confusing but also deleterious, which had led to the development of an integrated model also used by the airline participating in the present study (Mavin et al., 2013). On the positive side, such techniques and instruments tend to have high external validity (i.e., are valid across situations) while being economical and practical. Such instruments, including the Objectively Structured Assessment of Technical Skills in the case of surgeons, therefore, make up an integral feature of assessment in the medical field and in medical training (e.g., Royal College of Ophthalmologists, 2013).

Rater training constitutes an integral part of efforts designed to ascertain inter-rater reliability. Such training includes, for example, coding sessions where two raters independently score the same situation with subsequent analysis of when and where the raters differed (e.g., Horng et al., 2010). Training sessions where pairs or groups code a number of samples jointly tend to increase inter-rater reliability, which improves further with total number of episodes of joint coding (e.g., Schoenfeld, 1992). Performance-dimension training—which involves using particular scenarios from an aircraft simulator session as an example of decision-making and training assessors to rate this dimension consistently—appears to be one of the most effective means for achieving reliable assessments (Baker et al., 1999). However, one notable study pertinent here shows that even after three years of training, the inter-rater reliability among instructor/evaluators assessing the performance of actual pilots in videotaped scenarios has not improved a lot (Holt et al., 2002).

In aviation, the context in which the present study was conducted, pilot training and assessment historically focused on technical (flying) skills and associated aircraft technical knowledge and procedures (Flin et al., 2009; Mavin & Murray, 2010). However, whereas there has been a decrease in technical skills- and knowledge-related accidents in the airline industry, non-technical skills in areas such as communication and decision making have been listed as the causes in airline disasters (e.g., Air India Flight IX-812) (Helmreich et al., 1999). Investigations revealed a mismatch between traditional training and assessment methods, and causes of accidents, highlighting an important need for changes in pilot training and assessment (Salas et al., 2004). As a result, the industry turned to crew resource
management training, which focused on non-technical skills including decision-making, situational awareness, management, and communication.

In Europe, the NOTECHS (non-technical skills) system was developed for assessing pilots’ crew resource management skills along the dimensions of cooperation, leadership and management, situational awareness, and decision making (Flin et al., 2003); in the southern hemisphere, the integrated MAPP system (Mavin & Dall’Alba, 2010)—which includes technical dimensions (aircraft flown within tolerances, knowledge/procedures) and non-technical dimensions (situational awareness, decision-making, management, communication)—is used by a number of airlines and the Australian national defence. However, a better understanding of the assessment process is required if it is to contribute to a decreased focus on threat and error management and an increase in crew resilience to surprises and anomalies in flight situations (Dekker & Lundström, 2007). This study was designed to better understand the sources of variability not only in pilots’ performance assessment but also in the reasons provided for a particular assessment score. The study was to account for variability—that is, its nature as grounded in everyday reasoning of pilots—rather than treat it as a source of measurement error.

3. Methods

The present study was designed to investigate the sources of variability underlying performance assessment ratings with a particular focus on understanding the reasons pilots differ in their assessment of the performance of their peers. Pairs of pilots of different rank (flight examiners, captains, first officers) were asked to assess the performance of a captain shown in a scenario filmed on a flight simulator. We looked for evidence to answer questions such as “Why do assessors differ in their assessments?” and “What are the reasons that assessors use to justify their assessment of the captain’s performance?”

3.1. Research design

A standard method used by cognitive scientists to investigate what and how experts and non-experts think is the think-aloud protocol (Ericsson & Simon, 1993). Given some task, (non-) experts in the field of interest are invited to talk aloud while solving it. Although it has been suggested that the think-aloud protocol does not interfere with thinking, practitioners tend to find this method of eliciting data unnatural, providing them with difficulties of continuously saying what they think (Roth, 2007). However, a modified think-aloud protocol task asks participants to collaboratively solve tasks to arrive at a common solution (Suchman, 2007). This requires participants to articulate for each other everything needed to arrive at a shared solution; solving a problem in pairs and talking about one’s reasons is experienced as more natural. This is particularly true in the airline industry where flight examiners and training captains externalize their assessments as part of their work; and the trainees and assessed pilots, as part of debriefing sessions, also
externalize the reasons for making decisions while flying. The modified (pair-wise) design was therefore considered to provide higher ecological validity, to be experienced more naturally by the participating pilots, and to produce natural protocols of performance assessment processes (Roth, 1996). This task design nevertheless offers challenges in the sense that participants tend to take for granted a lot of what they are conscious of and what enters their reasoning processes. We therefore included standardized prompts in this research protocol to ascertain that the relevant aspects of the assessment were articulated for the record (e.g., “What were you thinking [before]?” or “You mean [‘I circle poorly timed’] the fuel discussion?”); non-standard elicitations were also used (e.g., “Is that [rewinding the tape] the point where you said the check was missing?”). At the end of each session, participant pairs were asked whether they had said everything they thought relevant to an assessment or whether their company assessment was leaving out something that should have been included. All pairs ascertained that their assessment was complete and nothing further was to be said.

3.2. Task

The task consisted of the analysis of a previously recorded flight scenario. It featured two company pilots flying in their organization’s simulator. In the scenario, the weather conditions are such that at the decision point, visual contact with the runway is not established as stated in the standard operating procedures (SOPs) leading to a “missed approach call,” that is, to an abandonment of the approach followed by an ascent toward the missed approach altitude. Already during the approach, the flying pilot (first officer) had commented on the low fuel situation. Following the missed approach but prior to reaching the missed approach altitude, the non-flying pilot (captain) and the flying pilot discuss what to do next. They do so within the 1,000-ft zone during which no such conversation ought to take place. The captain proposes to fly to the turning point (as specified in the standard operating procedures) and attempt another approach to the destination airport prior to departing, in the event of another “missed approach,” for the alternate airport. From a commercial standpoint, landing would have been preferable because it is the less costly option if the second try (from another runway direction) is successful. The first officer, however, has a strong preference for heading toward the alternate airport given that there is headwind, fuel is limited and the weather at the destination, while acceptable, is, in his words, marginal: “I hear what you’re saying but I’m I am quite worried about the weather back at [alternate airport] and the fact that we’ve got a headwind down there as well. So it’s not going to be a visual approach back there, it’s still going to be the ILS [instrument approach].” In this case, the first officer is outlining his concern about the aircraft being placed in a situation where the aircraft may go below normal fuel reserves and thereby compromise flight safety.

The scenario was filmed in a full-motion simulator (i.e., providing realistic sensations for the movement of the aircraft) but with motion turned off to avoid camera shake. Three cameras were used to record the scenario, two of which provide close-ups of the captain’s and first officer’s instruments, respectively,
whereas the third camera shows both pilots in the way a flight examiner seated directly behind (as during “line checks,” i.e., assessments during regular flights rather than on the simulator) would see them.

During the performance assessments, the participating pilot pairs had available two pages from their company-training manual. This allowed us to ascertain ecological validity and to increase likelihood that they had covered all assessment items with which they were familiar. The first page depicted the model of assessment of pilot performance (Mavin & Dall’Alba, 2010), which is used by the participating airline as well as several others and the Australian national defense. The model includes six performance components: situational awareness (SA), decision-making (DM), aircraft flown within tolerances (TOL), knowledge (KN), management (MAN), and communication (COM) (Fig. 1). The second sheet consists of a performance assessment grid that goes with the model. The grid has the six performance categories in vertical orientation and five performance levels (1 to 5) in horizontal orientation. In the airline of the participants, pilots receiving one “1” or three “2”s have to repeat the assessment exercise. A second failure would lead to the pilots being put on a strict training plan; and, if pilot continue to under-perform, they risk losing their jobs. In this assessment form, each performance component included sub-components (e.g., situational awareness included “perception, comprehension,” and “projection”). Each of the five performance levels for each of the sub-components has a corresponding performance description. Thus, for example, for the sub-component “time” under decision-making, the following five descriptors distinguished performance from “1” to “5”: “poor grasp of time available,” “limited time appreciation, led to rushed or delayed decision,” “decision made in time available,” “decision made within clearly established time constraints,” and “consistently identified problem, choosing best possible option in time available, appearing almost intuitive.”

The rater pairs were instructed to first make personal notes concerning their views on the performance observed and then to engage in a discussion for the purpose of arriving at a single score. Participants were told that the purpose of the research was to understand the reasons for giving one rather than another score. For each category, the participants first described their observations and then sought to match these with one of the performance scores. Each pair discussed the case until agreement was reached. In the few instances where the pairs did not appear to provide reasons, one of the two researchers present used standardized prompts as described above.

3.3. Participants

This study involved 18 pilots from the same company, 6 each from the ranks of flight examiner (FE, who are also captains doing routine flights when not assessing),
captain (CAP), and first officer (FO). \(^1\) (FE1 denotes the first pair of flight examiners, CAP3 denotes the third pair of captains, etc.) The participants assessed the performance of the pilots in the video (captain and first officer) in pairs of the same rank (to avoid power-over effects that sometimes are still noticeable in the industry). Table 1 provides background information on age, number of flight hours, and number of years as commercial pilots. Whereas the flight examiners and captains did not differ significantly in age, flying hours, and years as commercial pilots, there were significant differences on these parameters with first officers (\(t(7) = 11.06, p < .0001; t(7) = 6.87, p < .001; t(7) = 8.12, p < .0001\)). Two of the regular captain participants also had worked as training captains, and there were two ex-fighter pilots. Among the flight examiner participants were individuals who functioned in their airline as organizers of the simulator training and assessment profiles, crew resource management, and human factors coordination. All participants had prior training in assessing the performance of other pilots shown in recorded flight scenarios, using the company’s model and assessment form, as part of their crew resource management training. Each individual had assessed at least 3 videos using the same pages from the company-training manual that was also available here. The flight examiners had been using the model and assessment form for about 1 year.

3.4. Data collection

All assessor pairs viewed the laptop-based scenario on a 52” monitor in their company’s training facility, where they also had received training on using the company performance assessment model and assessment form. Three cameras were used to record each assessment session. The first recorded the assessor pairs face-on from slightly above eye level; this recording also featured, at its edge, the laptop on which a smaller version of the scenario was seen to allow precise identification of what the assessors were talking about. The second camera focused on the working area where the assessors made notes and marked the assessment form. This arrangement allowed precise identification of the topic talked about when assessors used deictic terms such as “this,” “that,” or “here” while pointing to the assessment form or model. The third laptop-based camera served as backup and recorded the same as the first camera but from the tabletop position. All tapes were transcribed word for word; the assessors’ individual notes as well as the researchers’ field notes were collected and included in the database.

3.5. Data analysis

Data were analysed both individually and in joint meetings that followed the precepts of protocol analysis (Ericsson & Simon, 1993) commonly used in cognitive science and studies of expertise, and interaction analysis (Jordan & Henderson, 1992). All pilots had signed ethics forms guaranteeing that their [non-] participation had no effect on their employment status and granting them confidentiality (unless they explicitly signed to have their photos released for publication purposes).
1995) typically used for understanding interaction in the workplace. The latter method was designed at XEROX PARC (Palo Alto Research Centre) for the purpose of understanding human-human and human-machine interaction. It essentially focuses on what participants in a particular setting make available to each other and does not allow claims that cannot be objectively verified in the data available. This is so because the researchers are forced to point to concrete evidence for each claim; researchers are not allowed to speculate but must provide each other with observable evidence. This allows researchers to quickly reach agreement about the fact that FE1 used the performance with respect to the 1,000-ft-limit and the ALT*-related standard operating procedures (SOP) in support of the rating of the captain’s knowledge of facts and procedures. Thus, a researcher may not state that a participant was thinking something unless there was clear, incontrovertible evidence that thinking had in fact occurred (e.g., “I was thinking”), that a participant meant a phrase in a particular way (e.g., “I meant to say” or “I can’t call that minor”), or that something was perceived (e.g., “Did you notice . . .?” or “We didn’t see that briefing part”). The method is based on the supposition that human participants make available to and for each other everything required in a situation to exhibit its rational, organized, patterned, and meaningful nature (Garfinkel & Sacks, 1986).

4. Results

This study was designed to better understand the reasons underlying the variability in the assessment of pilot performance completed by pairs of pilots of different rank. In this section, the results of the overall assessments are presented, followed by a detailed account for the reasoning underlying the assessments. The assessing pilot pairs recognized that the ultimate performance of the cockpit arises from the collaboration of the team members. As one recent study pointed out, team situational awareness is crucial, because the shortcoming of one team member can be compensated for by the attention of another (Sorensen & Stanton, 2011). The assessor pairs made their evaluations in view of the fact that the purpose of evaluation is to make judgements about individuals and their suitability for the job in the complex environment that the cockpit of a modern aircraft constitutes.

4.1. Overall assessments of the captain

Variability in assessment frequently is attributed to the subjective nature of the scoring process when conducted by individuals (e.g., Horng et al., 2010). The present study shows, however, that even when assessors work in pairs to arrive at a common score—forcing them to ground their judgment in observable facts—considerable variability exists (Fig. 2). Even more importantly, in this study variability existed even though the assessors could, and in fact did, repeatedly replay the scenarios. This is in contrast to those situations where scoring is done without opportunities to return and consider, such as when individuals score principals while following them around in their daily work, as Horng et al. have done.
Fig. 2 clearly exhibits considerable variation in differently ranked pilots’ assessment of the captain’s performance, reproducing and exemplifying the results of an earlier study involving 92 pilots (Mavin et al., 2013). The variability is especially notable in the knowledge (KN) component, where the performance ratings ranged from a low of “1” (CAP1) to a high of “5” (FO2). Fig. 2 shows that of the nine assessments of the decision-making component, four received a rating of “2.” The knowledge (KN), management (MAN), and communication (COM) components also had 2 (of 9) assessments scored “1” or “2.” The inter-rater reliability scores between any two assessor pairs of the 6 performance categories are below $r = .80$ in all but 2 of the 15 comparisons, and are below $r = .70$ in all but 4 instances. Although these results are to be taken with caution given the small number of rater pairs and low statistical power, the results suggest low inter-rater reliability in a case where reasons for a score were explicitly discussed and where pairs only continued rating another category when both members were satisfied that the evidence supported the score ultimately marked.

In the assessing pilots’ airline, one “1” rating or three “2” ratings lead to a required repeat assessment. Fig. 2 shows that two of the captain pairs (CAP1, CAP2) considered the captain’s performance as a fail and two further groups (FE3, FO1), with two “2” ratings each, came close to recommending a repeat assessment. Thus, for example, in the flight examiner group FE3, a participant suggested: “I mean I’m happy that both of them have passed. But there is deficiencies, especially with the captain, that need to be dealt with. So no, I’d be happy with that [assessment].” All pairs indicated at the end of the rating that they felt comfortable with the performance assessment and, therefore, with the implied recommendation of pass/fail that they had made.

The analysis reveals that the assessors isolated (identified) specific aspects in the video but these were not consistent across the pairs. For example, FE1 noted that the first officer, who flew the aircraft, did not confirm the setting of the missed approach altitude as required by the SOPs, and the captain, the non-flying pilot, failed to seek this confirmation as required by the SOPs. In a similar vein, this pair noted that the missed altitude capture mode was not mentioned in the cockpit; it should have been if the pilots were following company standard operating procedures. Flight examiner group FE1 also noted that the captain had pulled out the checklist but then put it back and did the check from memory. These facts did not become a topic in other groups, where, however, different (sets of) facts were perceived and articulated. For example, only group FE3 noted that the first officer did not verbalize the fact that a warning signal showed up on the ADU (advisory display unit), but that he actually addressed the issue by changing the setting on the rudder trim. Only first officer group FO7 noted that there was a discrepancy between the flight level requested by the first officer who was flying the aircraft (1-7-0) and the one that the captain (monitoring, non-flying pilot) requested from the tower (1-5-0).
4.2. Some good reasons for the observed variability in performance assessment

It might be suggested that non-technical performance aspects are more difficult to assess because these are more judgment-based than technical performance aspects (objective knowledge and [company] standard operating procedures). Even more difficult to assess might be considered those aspects that arise from power-over issues, including the level to which rank difference is brought to bear on crew resource management. In the following, pilot assessor reasons in each of these three dimensions are presented.

4.2.1. Assessment of (objective) technical performance dimensions: knowledge

It might appear that knowledge is easily assessed, because, as the subcategories facts and procedures show, assessors are asked to identify the presence or absence of facts and whether correct (standard operating) procedures have been followed. However, in this study there was considerable variability in how assessors rated the performance on this dimension and in the reasons for doing so (Table 2). The ratings ranged from the lowest possible (CAP2) to the highest possible (FO2).

For CAP2, the reasons for rating the performance as “1” included that the captain’s proposal compromised actual fuel reserves (Table 2). The only time this is possible, according to standard operating procedures, is before take off. This pair considered the standard operating procedures to be “quite categorical”; the pair found major deficiencies in understanding the procedures and considered as excessive the time it took to execute the procedure. On the other end, FO7 suggested that the captain knew the procedure, and knew where to cut a required 80 kg of fuel to mitigate another attempt at landing, which the pair considered to be consistent with the standard operating procedure (Table 2). In between, the pairs suggested that the standard operating procedures could be interpreted differently, and that what the captain proposed was “legal,” though perhaps risky. There were also suggestions that the captain (a) was “creative” in his interpretation (FE2), (b) was “reading between the lines” and “blurring the boundaries” (FE3), and (c) was “applying ‘poetic’ licences to the intent of the SOP” and “aware that there was a way to fudge [the standard operating procedures]” (CAP1). Thus, 7 of the 9 pairs found evidence that the captain clearly knew the standard operating procedures (Table 2). Other facts mentioned as violating the procedures were that the future plans were discussed within the 1,000-feet limit of the missed approach altitude and that a decision was made prior to alerting the tower that a missed approach had been flown (CAP1).

4.2.2. Assessment of non-technical performance dimensions: decision-making
There was similar variability in the dimension of decision-making, a dimension that belongs to the non-technical performance area. The ratings varied from 2 to 4 and there was considerable variability and fuzziness in the reasons provided (Table 3). Among the main reasons cited for the “2” assessment level were the untimeliness of the decision-making with respect to retrying the landing versus choosing the alternate approach, the high risk involved in depleting fuel reserves, lack of consideration of the weather situation, and insufficient option generation (Table 3). Again, all assessor pairs did not evenly generate these facts. On the other end of the rating scale, the decision-making was considered to be “4” (FE1).

An important consideration was the timing of the discussion about whether to attempt another approach to land or to divert. Table 4 shows that the assessments ranged from “too late” (FE2, FE3, CAP2, FO2) to “too early” (FE1); two groups (CAP1, FO1) considered the timing both as too late, because the discussion should have been had prior to the approach, and as too early, because it happened within 1,000-feet of the missed approach altitude, when such a discussion should not take place because of the risks to the flight brought about when the altitude is not captured (by the automatic pilot). One pair considered the timing completely appropriate given that the missed approach procedure had been completed perfectly and that now was the time to consider alternatives. FO3 did not explicitly comment on the timing of the decision-making process but noted that the process was rushed.

One chief reason given for the assessment that the discussion was timely was that the fuel consumption would be reduced if the aircraft would climb straight away to the return altitude rather than levelling off at missed approach altitude and then continuing (FO1, CAP3). FE3 suggested that they could have “drawn the line in the sand” by establishing clear criteria prior to engaging in the descent for taking one or another option. Above a certain fuel level, the future decision would be retrying; below the level, the decision would be to divert: “So the decision point’s been delayed. But there’s sound planning involved. So they wouldn’t have had to have the discussion.” A simple look at the fuel gauge would have been a satisfactory means of selecting which option to fly.

4.2.3. Assessment of “elusive” performance dimensions: rank-related pressure

An important point that appeared to mitigate the assessment discussions—especially when communication was the topic—existed in the question about the degree to which the captain tried to influence the decision-making process. In the cockpit, there still exist power-over issues even though less than this has been the case historically. This was quite apparent in the 2010 crash of Flight IX-812 of Air
India upon trying to land at Bangalore airport. The aircraft came down too late and too hard, and, as a consequence, overshot the runway. A total of 158 people, including 6 crewmembers, died. During the landing, despite repeated warnings and go-around options from the first officer, the captain (flying pilot) had continued. Among other reasons, the investigators stated that “the captain 'forcefully prompted' the first officer to give an affirmative call” (Aviation Medicine, 2013). Although the first officer made another, third, go-around call, the captain persisted with an approach that ended in tragedy. In the present study, as our analyses reveal, whether a power-over situation is at play may not be clearly established (Table 5).

In the course of the conversation about whether to retry or divert, the captain in the scenario presented says: “But I am not influencing you (two-handed gesture, ‘I am keeping my hands off’),” “Can’t do anything to change your mind?” and “You’re a hundred per cent?” The pairs made good albeit fuzzy distinctions in what exactly was happening between the crew, leading to scores between “2” and “5” for communication. Table 5 shows that some assessor pairs heard the communication as being non-threatening, without pressure upon the first officer, and without any undue effort to push a decision (FE1, FO2, FO3). Others heard and saw evidence in the videotaped scenario that the captain applied subtle pressure to make the first officer repeal his view that going straight for the alternate airport would be the safest option. The three captain groups (CAP1, CAP2, CAP3) assessed the pressure as being more than just subtle. Interestingly, two of the lower-ranked first officer groups (FO2, FO3) did not hear/see evidence that the captain attempted to unduly pressure the first officer, whereas one group felt that there was at least some subtle pressure, though it was applied in a non-threatening way. That is, those pilots whom one may have anticipated to be most sensitive to possible pressure actually assessed the situation as exhibiting no pressure and as being non-threatening. Although there was no specific item on the assessment form that specifically asked participants to rate this aspect in the relation between the two pilots, what the assessors felt the relation to be appeared to mitigate how they evaluated the situation, especially in terms of the communication category. The difficulty to assess this aspect was explicitly articulated by one participant in FE2, who emphasized the fact that there are different ways that such situations can be perceived and described. Where would an examiner put the observation that “body language” and the tone and nature of the questioning appear to belie the verbally articulated “I don’t want to put pressure on you”? One of the flight examiners in FE2 said,

Trying to fit that in to one of these boxes, it’s quite a hard, one because it’s very, very subtle. . . . I mean what can you say to someone afterwards, you say “You were trying to put pressure on me,” you say “No I wasn’t, I was not trying to put pressure on you.” But you can say, “Well, why would you say that if you’re not trying to put pressure on someone?”

Those assessor pairs who perceived the captain as “leaning” on the first officer tended to remark positively on the latter’s standing his ground and convincing the
captain with the suggestion of diverting (e.g., “The first officer stuck to his guns” [FE3] or “[the co-pilot] definitely resisted the first call, we’ve only got 1,200 of fuel” [CAP1]). One group (CAP2) pointed out that with another co-pilot, the same situation might have ended in trying the landing again, “which would have seriously compromised the safety of the flight” (Table 5), rather than in the diversion.

5. A fuzzy logic model of performance assessment

The results of this study provide extensive evidence (a) that pilots, in assessing the performance of peers, have good reasons for the scores they attribute and (b) that there is great variability in the scores and the reasons. The data provided in Tables 2–5 exhibit that the evidence used in making a performance assessment tends to be fuzzy, such as when the assessors talk about communication. Here, they saw varying levels of pressure. These results, therefore, strongly suggest the need for an approach to modeling assessment that differs from the inter-rater reliability concept. In the literature, there have been suggestions that (a) performance assessment in the field is a categorization and judgment rather than measurement issue (Govaerts et al., 2007); and (b) categorization and judgments generally—as in medical diagnosis (Esogbue & Elder, 1980)—and performance assessment specifically is well modeled by fuzzy logic reasoning (e.g. Ciavolino et al., 2013; Özdaban & Özkam, 2010; Roth & Mavin, 2013). In the following, we articulate how the results presented in Table 2 are modeled using a fuzzy logic approach. In this approach, intrinsically fuzzy aspects of assessment are modeled mathematically. This approach has been shown not only to be practical but to result “in models of greater validity than those based on classical and set theoretic approaches” (Esogbue & Elder, 1983, p. 223).

In the fuzzy logic model proposed for performance assessment (Ciavolino et al., 2013; Özdaban & Özkam, 2010; Roth & Mavin, 2013), two fuzzy sets $B_L$ and $B_U$ ($0 \leq B_{i,j} \leq 1$) define the lower and upper boundary of the five performance levels associated with each category used in the assessment. Because performance raters have been reported to tend to avoid extremes and exhibit a preference for the middle range, we defined the boundaries for unsatisfactory (u), minimal (m), satisfactory (s), good (g), and very good (vg) as $\{0 \leq u < .15 \leq m < .35 \leq s < .65 \leq g .85 \leq vg \leq 1\}$ (i.e., a satisfactory rating has $B_{l,s} = .35$ and $B_{u,s} = .65$ as lower and upper boundary). In the case of the assessment of knowledge of facts and procedures, a total of 5 facts were used as observational evidence: (a) confirmation of the ALT* setting, (b) speaking during the 1,000-ft range prior to altitude capture, (c) SOPs concerning fuel supplies, (d) SOPs concerning flying a missed approach, and (e) SOPs concerning weather at alternate airport. A fuzzy relation $W$ specifies the weight of a specific assessor observation on the performance level. For example, in the specific scenario, the central aspect concerned the discussion of how much the captain knew the standard operating procedures concerning fuel. This dimension, therefore, was associated with a weight of $W_{F,i} = 1$ for all performance levels $i \in \{u, m, s, g, vg\}$. Every observation used in support of making a judgment is coded in a fuzzy set $O$. For example, CAP2’s reason that the rated pilot “fundamentally
compromised fuel” and thereby clearly acted against the standard operating procedure constituted a fuzzy rating of $O_{F,CAP2} = 0$. A fuzzy code of .5 was used to code (a) the observation that the pilot “was graying the area, reading between the lines” $O_{F,FE3} = .5$ and (b) that the pilot “was fudging the procedures, but ended up okay” $O_{F,CAP1} = .5$. Thus, for example, the pair FE2 used 2 observations as evidence (Table 2): knowledge of the missed approach procedures (“good knowledge all around” $O_{MA} = .8$) and the fuel issue (“being creative,” “legal but not allowable in terms of company procedures,” “knew the procedure” $O_F = .5$). (See Table 2 for all fuzzy codes used for modeling the knowledge component; the fuzzy codes for other dimensions are provided in Tables 4 and 5.)

To model the assessment of a particular category, such as knowledge of facts and procedures, a clustering technique is used based on Euclidean distance between the fuzzy set of $n$ observations the assessors make and the performance levels.

$$D = \sum_{j=1}^{n} W_j (B_{L,j} - O_j)^2 + \sum_{j=1}^{n} W_j (B_{U,j} - O_j)^2$$

Table 6 presents the distances for each of the 5 rating categories for all nine assessor pairs. The smallest distance (bold-faced box) corresponds to the performance level that the fuzzy logic model predicts; grey shading is used to indicate agreement with the actual rating. The table suggests good agreement between model and actual performance ratings, thereby confirming the results of another study (Roth & Mavin, 2013). There is one point of divergence. In the case of FE2, the mismatch occurs where the difference between neighboring distances is actually very small. This may have been the result of too little weight being assigned to flying the correct missed approach procedure. Taken together, these results suggest that variability can be modeled and therefore can be explained rather than being an effect that leads to random error, as assumed in inter-rater reliability approaches (e.g., Roth & Mavin, 2013).

6. Discussion

This study was designed to investigate the variability that arises in the performance assessment of pilots. It shows that the assessors have good, albeit fuzzy, reasons for their assessments. This suggests that the variability of the performance scores is not random and, therefore, needs to be treated as explained (explainable) variability. This study therefore allows understanding of why training studies that attempt to improve inter-rater reliability, even after three years, have achieved only moderate improvements (Holt et al., 2002). The fact that a mathematical model approximates well the wide variability further underscores the hypothesis that variability is a real phenomenon. If this were to be the case, to be affirmed by further research, then researchers and practitioners may embrace the phenomenon and consider ways in which it can be used to enhance training rather
than spending financial and human resources on training that only marginally improves upon variability.

The investigation shows that there were a considerable number of facts noted only by one or two groups. For example, the unacknowledged advisory display unit (ADU) signal or the unacknowledged missed approach altitude call and failure to elicit the acknowledgment, and the wrong flight level height call each was noticed by one pair only. Because all assessor pairs were routinely flying pilots, we can anticipate that they would have acted in the way they saw the pilots in the video act. Thus, if the first officer did not call the captain on the wrong flight level call (15[0] instead of 17[0], Table 1), then this might have just as well escaped the participant had he been in the cockpit with the captain. This would have considerable consequences, for example, for the assessment of situational awareness, an integral dimension of which is perception (e.g., Taber et al., 2013). The fact that the captain in the scenario continued to question the first officer regarding whether he would consider trying again, despite the information made available—the latter’s articulation of the weather situation in the alternate airport and headwind on the way there—may suggest that he was apparently not comprehending the signification of this information and projecting to the future, indications of lower situational awareness (e.g., Kaber et al., 2006). However, all but one of the assessor pairs found the captain to perform at the “3” or “4” levels in this dimension. Inconsistent observation of facts mitigates the way in which assessors rate the situational awareness of the examined pilot. This, in turn, leads to variability in assessing situational awareness. Given that all participants were knowledgeable in the domain assessed, this raises questions about the possibility to reliably measure situational awareness—in the way research such as that by Taber et al. (2013) is designed to achieve—in an ecologically valid (rather than workplace-independent and contrived) way.

This study shows that there is great variability in the performance assessment even though each pair spent considerable time in elaborating evidence to support any given score. The variability has as its effect low pair-wise inter-rater reliabilities. However, although there is great variability in the assessments within assessment categories (e.g., all levels between 1 and 5 for the knowledge of facts and procedures category), the fuzzy logic model provides a good approximation to the actual ratings provided. Where there are differences between actual and predicted scores, these are associated with small distance differences between pairs of categories in the model. The results are consistent with those reported in another aviation-related study (Roth & Mavin, 2013) and those of fuzzy logic models of categorization and judgment more generally (e.g., Esogbue & Elder, 1980). Such models are in support of suggestions of treating performance assessment as a categorization and judgment issue rather than as a measurement issue (e.g., Goevarts et al., 2007).

In the scenario used, the ultimate outcome of the flight remains unknown. There could have been a tragedy at the new target airport, in which case an after-the-fact coroner’s analysis might have suggested that the pilots should have tried landing at the airport that they did not land at on request of the FO. In such a scenario, very different forms of reasoning could be anticipated to have occurred in the assessment of the conversation in the cockpit. The title of our paper suggests that there are good
reasons for the considerable variability in assessment of piloting performance. Good reasons, from an assessment perspective, are those that are defensible (Kubiszin & Borich, 2010). Here, the assessors also are experienced pilots, giving each other, within the pairs, reasons for scoring a performance one rather than another way. Our modified think-aloud sessions, therefore, constitute natural protocols of the sense pilots make of their practice, as seen in their assessment of the performance of peers. The participants can be thought of as team members in the sense of Langan-Fox et al. (2009)—here virtually participating in such team processes as task preparation, post-task reflection, analysis, strategy formulation, planning—and their understanding of the events reflects what they might have perceived with some likelihood had they themselves been in the situation depicted in the scenario. The way in which they hear the pilots in the scenario and perceive their respective actions is likely to be the way in which they would hear their partner in the cockpit speak and perceive his/her actions. If there are pilots who perceive the captain to have subtly pressured the first officer, we can expect that they would have heard him in this way had they been in the first officer’s seat; if they felt there was no pressure or a non-threatening offer to consider another option, then this is a likely hearing we can expect from them.

This study throws light on the attribution of causes after accidents have occurred because there is insufficient appreciation for how pilots generally might perceive some critical situation. Thus, after an accident—such as an airline crash, cruise ship crash, or radiation accident in a nuclear power station—will be the attribution of fault and what the agents did or did not do to prevent it. This is known as “the coroner’s problem,” which involves a particular form of reasoning practice that—though an everyday form of reasoning in its own right—differs from the way in which the people involved actually reasoned (Garfinkel, 1967). The coroner, having a corpse, constructs a story that led to it. His/her account has to be able “to withstand claims that that account was incomplete or that the death happened differently than—or in contrast to or in contradiction of—what” (p. 15–16) the coroner and his/her team has stated. But rather than taking the coroner’s perspective, one could also take a different point of view. For example, pilots are asked to assess performance without knowing the end result, as in this study, or are asked to assess performance in situations that potentially place an aircraft in a situation that could compromise the safety of the aircraft. In these cases, the discussion and final assessments would not be biased by the fact that there is a cause to be found. Rather, all aspects that ought to be critical to flight safety would be discussed without the teleological reasoning that one or the other was the cause of the crash that followed. In the present situation, the participants were provided with an excerpt (scenario) from a flight where a decision had to be made about trying to do another approach or immediately asking for permission to divert. How would assessments change if the assessors knew the ultimate outcome of the flight, for example, if there had been a major incident with the pilots having to declare a fuel emergency at the alternate airport? We do have evidence in our database that the entire assessment changes if assessors subsequently identify a serious safety issue (e.g., an engine left running on the side that an emergency evacuation was taking place). Future studies might be designed to better understand the way in
which assessments systematically change when outcomes are known. This might also assist in better understanding how senior management pilots within airlines might investigate issues that arise in their company.

7. Implications for research and practice

The results of this study suggest that there are many different, albeit fuzzy, reasons assessors of pilot performance provide in support of the scores they attribute. A mathematical model of the scoring based on fuzzy facts and fuzzy reasoning provides a good approximation of the actual assessments. Both of these aspects of the study therefore support the hypothesis that variability is a real phenomenon. This hypothesis is consistent with the outcomes of an extensive European study, which focused on increasing the validity and reliability of the scoring instruments and provided rater training over three years (Holt et al., 2002). If variability is a real phenomenon, then there are considerable implications, especially for practice, because companies such as the one involved in the present study have to decide whether they want to spend financial and human resources on efforts that might end with marginal effects, as the Holt et al. study did. If variability turns out to be a real phenomenon that no training or work on the rating instrument could remove, then one way of dealing with it would be to embrace variability and use it to enhance team resilience (Saurin et al., 2013), as we suggest below.

One possible explanation for the considerable variability in the assessment scores and assessment reasons may be the method of eliciting assessment used in the study. Thus, the scenarios employed here as in studies by other researchers (e.g. Holt et al., 2002) present only a brief episode (5–10 minutes) from a longer flight. A way to mitigate the effect of analyzing a decontextualized scenario would be to use longer flight situations to be assessed (e.g., an entire flight from beginning to the end). However, only smaller aircraft fly distances short enough to make it feasible within a regular airline to have their pilots analyze entire flights during a crew resource management training session. A more feasible option would be to show sequences of three or four clips from different stages of the same flight. Research is required to show whether this decreases variability. Whereas the preceding solution might be thought to lead to decreased variability given that the total incidences, number of facts, and complexity of events is more limited, the fact that other aspects of the same flight are not available may lead to an increase of variability. For example, the participants often made assumptions about what the assessed crew would likely have done or said, for example, during the briefing of an approach. That is, their knowledge of the company’s standard operating procedures would have mitigated their assessment of the instant. To tease out the role of familiarity with the aircraft and standard operating procedures, a study is under way to investigate assessment results when the crew in the scenario operates another type of aircraft used by a different airline in the same region. These assessment protocols will be compared to assessment protocols of the same scenario from pilots familiar with the aircraft in two conditions: (a) familiar with the particular route and (b) unfamiliar with the route.
Checklists are sometimes recommended as a means to improve the identification of factors that should be taken into account in the assessment and, in the process, improve inter-rater reliability (e.g., MacDonald & Sulsky, 2009). The assessment instrument is similar to those developed and used in other contexts (e.g., O'Connor et al., 2008), with the difference that rather than having a simple rating scale, the airline of our participant pilots had developed descriptors for each performance level. This study shows that the variability occurred despite the use of a 20-item rating sheet separated into 6 larger categories, which themselves divided equally into enabling and essential skills. Moreover, variability existed despite the fact that the assessors had the option to return to the scenario—which they frequently did—and ascertain the factual nature of specific observations. In line checks, where the assessor has to complete the checklist or rating sheet in real time without the option of going back, the reliability of the assessment can be expected to decrease.

A good understanding of the variability in the reasons on which assessment is based is important because performance assessment, especially when negative (i.e., “fail”), has considerable consequences for the lives of the assessed in terms of their careers and in terms of the psychological impact. When there is a pilot with difficulties, some airlines will make arrangements for an extensive retraining program. At the end of the training, the pilot will be examined again. To ascertain that the assessment is less subject to (intended or unintended) bias on the part of the examiner, two examiners collaborate in the assessment that follows training. Whereas this is an advance over having only one examiner conduct the assessment, the present study shows that even in the case of pairs of examiners, there is considerable variability—though the variability within the flight examiner rank is smaller than that within the other two ranks. This shows that variability does not go away when multiple evaluators become involved—which mitigates common assessment wisdom that larger numbers of raters and group assessment sessions will improve scoring reliability and lead to calibration (e.g., Kubiszyn & Borich, 2010). Further improvements in the process may be possible, however, given that many airlines now have the possibility to record the simulator session to be used during debriefing (i.e., “debriefing tool”). This record could be examined, in critical cases, by a larger group of examiners, who, in the process, are forced to externalize the reasons for their assessment. Variation could thereby be made the object of a conscious decision about the apparent weaknesses in performance and about subsequent decisions to be made. However, in some settings, the contracts with pilot union mandate the recordings of simulator assessment sessions to be deleted after the debriefing. In such settings, this avenue for improving the assessment reliability is not feasible.

A previous study had shown flight examiners to rate consistently with captains (Roth & Mavin, in press). However, it also had shown that the process of reasoning between flight examiners and captains differed. The captains tended to create an overall, holistic narrative of the performance. In one pair of flight examiners (FE3), this approach was explicated in the context of the current task by means of using gut feelings and then mapping these onto the assessment grid:

01 R: Well, I just tend to go gut feeling and then look into the metrics afterwards.
I’ll be the same. I’ll use gut feeling to know which word pictures to head for. And then work through and go, “Hey look, here, yea, crew members tasked effectively or adequate.” And then I go, “Yea, it was more adequa...
the pilots (Dekker & Lundström, 2007) with the possibility of more resilient solutions from the variability evident in joint cognitive systems (Bergström et al., 2011). This training would capitalize on the situation, because the variability is likely going to give rise to considerably more discussion in large-group assessment sessions. The assessment model and assessment grid used by the airline would then be a tool that stimulates and organizes discussions about performance aspects. Pilots of different ranks could learn in such environments where their peers of the same, lower or higher rank experienced a particular scenario and built a stronger peer culture around the notion of expertise. As our research in progress shows, such discussions tend to lead pilots to reflect on their own performances through the lens of the ongoing discussion about the performances seen in the scenarios. At times, a pilot realizes that s/he tends to act or behave in the same way but, seeing such behaviour from the outside, realizes that it could be improved or modified.

Acknowledgments

This work was funded by a joint grant from the Griffith University Industry Collaborative Scheme and the airline involved. The views expressed in this paper are those of the authors and do not necessarily reflect those of the university or the airline. We thank all participating pilots and the airline for their support in this research.

References


Legends to tables

Table 1. Background information on participants (means, standard deviations)
Table 2. Knowledge assessments [facts, procedures] made by 9 assessor pairs
Table 3. Performance assessment of the decision-making component
Table 4. Assessment of the timing of the retry/divert discussion
Table 5. Assessment of power-over mitigation
Table 6. Actual ratings of knowledge and procedures and most likely ratings based on a fuzzy logic model of assessment (boxes)

Captions

Fig. 1. Model of Assessment of Pilot Performance used by the participating airline.
Fig. 2. Assessments of the low-fuel-related scenario by three pairs each of (a) flight examiners, (b) captains, and (c) first officers.
Table 1. Background information on participants (means, standard deviations)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Age</th>
<th>Years as Commercial Pilot</th>
<th>Flight Hours (hrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flight examiners</td>
<td>49.2</td>
<td>25.3</td>
<td>14,250</td>
</tr>
<tr>
<td></td>
<td>(6.2)</td>
<td>(5.9)</td>
<td>(4,910)</td>
</tr>
<tr>
<td>Captains</td>
<td>45.3</td>
<td>24.7</td>
<td>15,420</td>
</tr>
<tr>
<td></td>
<td>(6.7)</td>
<td>(7.0)</td>
<td>(6110)</td>
</tr>
<tr>
<td>First officers</td>
<td>30.5</td>
<td>11.3</td>
<td>4,900</td>
</tr>
<tr>
<td></td>
<td>(3.3)</td>
<td>(3.9)</td>
<td>(1,960)</td>
</tr>
</tbody>
</table>
Table 3. Performance assessment of the decision-making component

<table>
<thead>
<tr>
<th>Rating</th>
<th>Justification</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>“The decision should have been made earlier. There wasn’t a line in the sand there.” “I mean it was a delayed decision. Wasn’t it?” “And that’s where it all went wrong. And I don’t mind people delaying a decision till some time in the missed approach, but so long, because it’s proformalized and they both know, and it’s made where there will be no discussion. And there was no line in the sand for me.” “The risk, the timeliness of it put them at risk. Inappropriate risk assessment.” “They had two options, didn’t they? But they hadn’t resolved which one they were going to use.” “That decision should have been made prior to commencing the approach, and it wasn’t. And in that, again, the context, if you bring that delayed decision in to that context, where it’s close to the ground, a very busy phase of flight, he should have known the decision needed to be made earlier.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“Limited time appreciation, led to rushed or delayed decisions. The fuel, as you said, I thought it was a hell of a time to be discussing fuel plan. When you’re just about to establish on the [instrument landing]. He gave himself a limited range of options. They did assess some risk, yes, that would be fair. . . . He did assess that with a 300 [kilogram] reserve he could make it back to [alternate], so you could say he did a risk assessment. The fact that it was not by our company procedures, that’s not mentioned there. It just says some risk assessed. So he did have some idea of what he was going to do for gas. As a captain he definitely had difficulty developing a plan with his crewmember.”</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>“Should have done it earlier. They hadn’t taken the effect in consideration in that particular case to go down to 600 [kilograms].” “The options were to give it another crack or to go home.” “Inappropriate risk assessment if you’re taking into consideration, if you come back to his knowledge and procedures.” “It is not really appropriate us actually doing risk assessment in that particular case. We have clearly laid down procedure. So did he have any right to be even thinking about it.”</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>“Appropriate at the right time, in option generation, and the risk assessment, maybe the areas that didn’t go so well, possibly. All facts not considered. I mean the things I’m thinking about are the weather back in [alternate], creating the initial difficulty. As the pilot monitoring, you would have been reviewing the weather well prior to that point, so thinking about what you would be suggesting. A risky option suggested, at an inappropriate time.” “I don’t think it was a safe way to, a safe option to suggest.” “Inappropriate risk.”</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>“The time there was good, they didn’t rush in to making a decision. But they did it in a timely manner. Talked about all the facts. Did a good diagnosis. They actually generated options this time, they had two options and tossed them up. Risk assessment. The first officer clearly put forward why his plan was less risky and explained that to the captain. They had a plan, which was a contingency worked out.” “Even though he probably wanted to go to something in [target city] that night, he said ‘that’s fine. We’ll go home.’ So you’d mark him up for that and say yep, he made a good decision.” “If fuel was this much of a problem, why are they still making the decision at the end of a missed approach? They could have said if we miss with one shot, then we’re straight off to [alternate].”</td>
<td></td>
</tr>
</tbody>
</table>
“He didn't kind of really develop a sound plan, he was kind of led by the FO.”

Well [the decision] certainly was made in the time available. He didn't kind of sit around in the holding pattern for a few hours, did he?” “Even before they got anywhere near [turn around for 2nd try] they were saying, you know, the decision to go was made pretty promptly.” “At least he was, he was good in agreeing with the pilot flying.”

“It wasn't a unanimous immediate decision. However, they did make a decision and they made it relatively quickly. So I think that’s probably made within clearly established time constraints.” “They could have said, ‘Right o, we've got one shot here, if we get to the end, we've got to go back.’ Unless it's obviously clearing from the south.” “The FO had made that decision. He had his plan. If they went around, he said, ‘No, another miss.’ Yea. Maybe the captain not quite so much.” “But really it was a delayed decision, it was, potentially a bit rushed towards the end as well.”

“They didn’t, I mean part of planning, trying to work something out, is getting all the information you can, you know?” “They never came together and explored the reasons why they both felt that way. And come to some agreement. It was obvious that [the captain] wanted to give it another shot. But just went along with what [the FO] said.” “There wasn’t much of a plan at all.” “Difficulty developing a plan.”

“Some very good decision-making. Good assessment done throughout, although perhaps a little bit at the incorrect time. . . . as far as the risk assessment goes, he was making quite a good assessment of it, because, I mean getting in to there, well we’ll take 80kgs off, we can still do this. . . . He was telling him his mind, I mean coming up with the cold hard facts. . . . So they made the decision to divert without actually having hard and fast figures. Can we divert? Because there was still quite a discrepancy there between what the FO said and what the captain said.”
Table 2. Knowledge assessments [facts, procedures] made by 9 assessor pairs

<table>
<thead>
<tr>
<th>Rating</th>
<th>Justification</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>“Basically he’s talking about compromising the actual reserve. The only time to actually do this is prior to departure, when you’ve got a payload issue. The way the SOPs read, though, is quite categorical. In the sense that’s the only time that you’re allowed . . . you’re not supposed to do it.” “Correct procedure not identified, major deficiencies or excessive time taken in execution.” (O_3 = 0; O_4 = 0) (^{1})</td>
<td>CAP2</td>
</tr>
<tr>
<td>2</td>
<td>“Deviated from the SOP [standard operating procedure] slightly. They were not meant to do that within a thousand feet.” “There’s two things again, there’s what he actually did in the cockpit at the time [late], the suggestion of what they were going to do. I thought the actual logic of the, of his role, was good. But our procedures should say that you shouldn’t be trying to land below minimal reserve in that instance.” “Lack of awareness of procedure, that’s our fuel procedure.” (O_3 = .25; O_5 = .25)</td>
<td>FO1</td>
</tr>
<tr>
<td>3</td>
<td>Well done. Just their procedures, except hitting that [missed approach altitude] he did call that, but a bit of a delay, he didn’t get a response from the pilot, correct procedures. ((O_1 = .5; O_2 = .35))</td>
<td>FE1</td>
</tr>
<tr>
<td>3</td>
<td>Good knowledge of company procedures. So missed approach procedures, good knowledge all around. I suppose the captain’s was a little bit down in one area, but he, what he was actually proposing wasn’t actually legal and it was legal but in terms of the company procedures it’s not allowable, what the captain was proposing.” “He was proposing to recalculate their fuel to use a smaller reserve. We can really only use that procedure when payload, so prior to departure. We do a fuel analysis. That’s something you do before you get airborne. Whereas he was proposing to get the paper out and do a hand amendment. Yea. Being creative I suppose. At least he knew the procedure. He had the knowledge of it.” ((O_3 = .50; O_4 = .80))</td>
<td>FE2</td>
</tr>
<tr>
<td>3</td>
<td>“Factual and procedural knowledge. My only question there is the captain thinking at that point that he can actually go below the 600 kilos. I mean it is, that depends how you interpret that SOP isn’t it?” “He’s reading between the lines there.” “He’s definitely greying the boundary, blurring the boundary [of the procedure] “the only issue with the captain as well was just that, can he drop down to the 300 kilos. But otherwise, all of the procedures for both of them, was going really well.” ((O_3 = .50))</td>
<td>FE3</td>
</tr>
<tr>
<td>3</td>
<td>“The fact that he was aware that there is a provision to lower the reserve to 300 [kilos], I would suggest he probably knew what he was doing. And was applying a bit of poetic license to the intent of, with the SOP. So I wouldn’t like to say that he was ignorant of the SOP. He was aware that there was a way to fudge it, which in some cases is legal and in some cases isn’t. Correct procedure performed with minor errors.” ((O_3 = .50))</td>
<td>CAP1</td>
</tr>
<tr>
<td>3</td>
<td>“Knowledge of what, fuel, is what their minimum fuel was a bit of an issue” “They weren’t certain about what they were doing. Was it the correct procedure to go back at [flight level] 17 or go back at [flight level] 15 to [alternate]? And not give [target] another shot?” “We’ve got a [SOP] book there that tells us exactly what we need to do. And how much we need to have. And so we shouldn’t deviate beyond that. So at that point, we go back.” ((O_3 = .20; O_4 = .85))</td>
<td>FO3</td>
</tr>
<tr>
<td>4</td>
<td>“In facts and procedures, he did make errors.” “It was just a [wrong] height call,</td>
<td>CAP3</td>
</tr>
</tbody>
</table>
but that was about it. It's not as if he was lacking anything as far as his knowledge is concerned, at all. He's just offering an alternative process.” (O₂ = .75)

5  “They recalled, I mean they knew all the correct procedures. You could perhaps make reference to the fact that the captain’s, just, you know, the option of dropping the reserve and coming back in with what was technically less than.” “he mentioned in there something about dropping the 80 kilos. Which, yeah, of course, if you use the point to point fuel tables, they do include the 80 kilos, which is the climb to 1,500 feet and the descent from there. Which you can do for the alternate. They knew where they could cut that 80 kilos out of the fuel figure. That's actually in the [SOPs], they did that quite well. The actual flying of the approach and the missed approach was textbook.” (O₄ = .95)

¹Fuzzy codes for the 5 flight aspects (codes O₁ to O₅ in the fuzzy logic model, Section 5) that entered the knowledge assessments (O₁ = ALT*; O₂ = 1000-ft limit; O₃ = fuel; O₄ = missed approach; O₅ = weather).
Table 4. Assessment of the timing of the retry/divert discussion

<table>
<thead>
<tr>
<th>Fuzzy Rating. Justification</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Too late</strong></td>
<td></td>
</tr>
<tr>
<td>“There’s no point, you could almost argue it’s almost too late to have that 300 kilogram argument or discussion then. It should have been had before, when they do.” (.35)¹</td>
<td></td>
</tr>
<tr>
<td>“That decision should have been made prior to commencing the approach, and it wasn’t. And in that, again, the context, if you bring that delayed decision in to that context, where it’s close to the ground, a very busy phase of flight, he should have known the decision needed to be made earlier.” (.20)</td>
<td></td>
</tr>
<tr>
<td>“They probably should have had that decision made before they started the approach.”</td>
<td></td>
</tr>
<tr>
<td>“Leaving it until the missed approach is not that great.” (.30)</td>
<td></td>
</tr>
<tr>
<td>“If you knew you going in to approach that gas would have been tight, the timely time to have this discussion about whether you want to have another go is before you actually commence the approach. “It’s a bit late in the piece to start talking about having another go on minimum gas, when you’ve been there and done it. It’s probably time then to just have a plan in place and work through it.” (.25)</td>
<td></td>
</tr>
<tr>
<td><strong>Timely, completely appropriate</strong></td>
<td></td>
</tr>
<tr>
<td>“They’ve done the missed approach, they’ve cleaned it all up, they’ve got through all their checks, which is fine. And then they started to talk about the fuel. Bang, straight in to fuel. So it’s obviously on both their minds, isn’t it?” (.65)</td>
<td></td>
</tr>
<tr>
<td><strong>Too early</strong></td>
<td></td>
</tr>
<tr>
<td>“Talking at an inappropriate time about things they could have left another minute or two. . . 700 feet to go here [to missed approach altitude] and, as you say, they’re still talking about something that’s got to happen later.” (.15)</td>
<td></td>
</tr>
<tr>
<td><strong>Too early, too late</strong></td>
<td></td>
</tr>
<tr>
<td>“And it’s interesting that the captain says oh there’s a thousand to go, and then keeps the discussion going, so...” “So that’s still, it’s probably an appropriate time because he doesn’t want to level off he wants to keep the climb going, to safe fuel, so he can get back to [alternate].” “[It was timely] given the fact that they were aware that there was a thousand feet to go, the Alt STAR call was still made, it was quite appropriate, yea.” “The only thing that I saw during the whole thing was, and it’s very minor, was that the discussion was had within a thousand feet of the level off.” (.45)</td>
<td></td>
</tr>
<tr>
<td>“Discussing fuel, at that stage they did, I don’t think that was timely. That was a call that should have been discussed at top of descent.” “You should, you should already know whether you’re going to have another shot prior to commencing the approach. What your fuel status is. And as you say, it’s too late to be discussing a 300 or 600 reserve. After you’ve already missed.” CAP1L “The second reason was they did it [fuel discussion] before they even told the tower they’d been round.” (.30)</td>
<td></td>
</tr>
</tbody>
</table>

¹The fuzzy codes provided in the modelling as suggested in Roth and Mavin (2013).
**Table 5. Assessment of power-over mitigation**

<table>
<thead>
<tr>
<th>Fuzzy Rating. Justification</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No pressure, non-threatening</strong></td>
<td></td>
</tr>
<tr>
<td>“[The captain] wasn’t just looking at his point of view. I mean he had the first officer’s interests at heart as well, in the fact that he wasn’t trying to persuade him in any way.” “[The captain] communicated appropriately, that’s fine, . . . because he, they talked all the way, back and forth.” “The first officer . . . was calling the shots.” “[The captain was] making sure that he wasn’t pressurizing the FO. He was telling him his mind. He wasn’t trying to persuade him [the copilot] in any way.” (.75)</td>
<td>FE1</td>
</tr>
<tr>
<td>“[The captain] did consider what the other crew was thinking with his suggestion. Timely he did voice and say well that’s my idea. You know, it may not be what you’re thinking, but you know, basically whatever you decide and you’re happy with, we’ll go with.” (.75)</td>
<td>FO2</td>
</tr>
<tr>
<td>“Each of them had a plan but they didn’t talk about it, didn’t develop one as a crew. Didn’t come to a consensus.” “They even struggled to generate the obvious options as a team, didn’t they?” “It was only resolved because the captain chose to agree sufficiently with the FO that there was the lowest common denominator. And if he didn’t agree there was going to be a right royal barney. I don’t think it was resolved amicably, but it was resolved.” “They just interacted, didn’t they.” (.60)</td>
<td>FO3</td>
</tr>
<tr>
<td><strong>Subtle pressure</strong></td>
<td></td>
</tr>
<tr>
<td>“Where do you fit that very subtly thing, ‘What I’m talking about here is,’ although you say, ‘You did the right thing,’ by your body language and by tacking on the end, ‘But I don’t want to put any pressure on you about.’ You can see the way you could frame these things. You could say, ‘Basically I want to do this,’ and make it look like a nice statement by saying, ‘I don’t want to put any pressure on you.’ To try and force people to do something that you want them to do. Which is what we talked about, isn’t it?” (.35)</td>
<td>FE2</td>
</tr>
<tr>
<td>“The captain should have just asserted himself there rather than initiate the discussion. Just go right, I want you to do the reversal and have another crack. Or, I agree, let’s go. Because it was that whole, ‘Oh, I don’t want to influence your decision, I don’t want to . . .’” (.45)</td>
<td>FE3</td>
</tr>
<tr>
<td>“[The captain] promoted the idea of dropping the fuel reserve to enable another approach, which I don’t think I would encourage. I mean, he tried to do it in sort of a non-threatening way, but he’s saying about something that we wouldn’t really want to see.” “But he wasn’t doing it in an overbearing way. But to his credit, he did accept the input.” (.45)</td>
<td>FO1</td>
</tr>
<tr>
<td><strong>More than subtle</strong></td>
<td></td>
</tr>
<tr>
<td>“He’s tried to talk the guy [FO] into taking a 300 reserve, which, that is a misconception that you can do that.” “He tried to encourage the co-pilot to run themselves into an illegal fuel state.” (.25)</td>
<td>CAP1</td>
</tr>
<tr>
<td>“It’s one thing making an error. But it’s basically making an error and then forcing someone else beyond their comfort zone. That’s another thing. You know, if he’d been with the right co[pilot]—or should be saying with the wrong—co[pilot], they would have been hanging around, which would have then seriously, you know, then could have seriously compromised the safety of the flight.” (.15)</td>
<td>CAP2</td>
</tr>
</tbody>
</table>
“[The captain was] trying to force the issue, well not force the issue, just suggest that they could come back around for another go. And being the pilot not flying he should have the idea of what's going on in aviation law.” (25)

1The fuzzy codes provided in the modelling as suggested in Roth and Mavin (2013).
Table 6. Actual ratings of knowledge and procedures and most likely ratings based on a fuzzy logic model of assessment (boxes) using fuzzy scores provided in Table 2

<table>
<thead>
<tr>
<th>Pair</th>
<th>Actual Rating</th>
<th>Distances¹</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unsatisfactory</td>
<td>Minimal</td>
</tr>
<tr>
<td>FE1</td>
<td>3</td>
<td>0.37</td>
</tr>
<tr>
<td>FE2</td>
<td>3</td>
<td>0.74</td>
</tr>
<tr>
<td>FE3</td>
<td>3</td>
<td>0.61</td>
</tr>
<tr>
<td>CAP1</td>
<td>3</td>
<td>0.61</td>
</tr>
<tr>
<td>CAP2</td>
<td>1</td>
<td>0.44</td>
</tr>
<tr>
<td>CAP3</td>
<td>4</td>
<td>0.48</td>
</tr>
<tr>
<td>FO1</td>
<td>2</td>
<td>0.27</td>
</tr>
<tr>
<td>FO2</td>
<td>5</td>
<td>1.14</td>
</tr>
<tr>
<td>FO3</td>
<td>3</td>
<td>0.49</td>
</tr>
</tbody>
</table>

¹ Grey shading indicates agreement
Legends to Tables and Figures

**Legends to tables**

Table 1. Background information on participants (means, standard deviations)
Table 2. Knowledge assessments [facts, procedures] made by 9 assessor pairs
Table 3. Performance assessment of the decision-making component
Table 4. Assessment of the timing of the retry/divert discussion
Table 5. Assessment of power-over mitigation
Table 6. Actual ratings of knowledge and procedures and most likely ratings based on a fuzzy logic model of assessment (boxes)

**Captions**

Fig. 1. Model of Assessment of Pilot Performance used by the participating airline.
Fig. 2. Assessments of the low-fuel-related scenario by three pairs each of (a) flight examiners, (b) captains, and (c) first officers.