Understanding Variance in Pilot Performance Ratings

Two Studies of Flight Examiners, Captains, and First Officers Assessing the Performance of Peers

Timothy J. Mavin,[1] Wolff-Michael Roth,[1,2] and Sidney Dekker[1]

[1]Griffith University, Nathan, QLD, Australia, [2]University of Victoria, BC, Canada

Abstract. Two studies were designed to investigate how pilots of different rank evaluate flight-deck performance. In each study, the pilots were asked to assess sets of three different videotaped scenarios featuring pilots in a simulator exhibiting poor, average, and good performance. Study 1, which included 92 airline pilots of differing rank, was aimed at comparing how individuals rate performance. The subjects used a standardized assessment form, which included six criteria, each having a 5-point rating scale. Analysis of the first study revealed that there was considerable variance in the performance ratings between flight examiners, captains, and first officers. The second study was designed to better understand the variance. Eighteen pilots (six flight examiners, six captains, and six first officers) working in pairs evaluated performances, in a modified think-aloud protocol. The results showed that there were good reasons for the observed variances. The results are discussed in relation to inter-rater reliability.

Keywords: nontechnical skills, assessment, performance assessment, inter-rater reliability

Original Article

Assessment of pilot performance has traditionally focused on flying skills and related aircraft knowledge that generally are referred to as "technical skills." Focusing on these skills both in training and assessment aims to build up an inventory of techniques for the operation of an aircraft. Such an approach is quite context-specific: It is set in, and tightly anchored to, the local technical environment (of an aircraft flight deck – which in itself operates within the wider aviation system) in which various kinds of problem-solving activities are to be carried out. The assessment of technical skills is fundamental to pilot certification and qualification (Johnston, Rushby, & Maclean, 2000). However, in the emphasis on technical skills, generic abilities often referred to as "nontechnical skills (NTSs)" – including human–human and human–machine coordination, communication, problem solving, management of crew member tasks, and problem escalation (Woods & Patterson, 2000; Woods, Patterson, & Roth, 2002) – have long been left to develop from the exercise of context-specific work. Yet it is the prevalence of aircraft accidents due to failures in the domain of NTS that have brought the importance of decision making, risk assessment, management, communication, and teamwork to the fore (Fischer, Orasanu, & Montvalo, 1993; Flin, Goeters, Hoermann, & Martin, 1998; Maurino, Reason, Johnston, & Lee, 1995; Nagel, 1988; Orasanu, 1990, 2001; Orasanu & Martin, 1998).

NTS, part of an area in aviation often referred to as crew resource management (CRM), has been defined as the "timely and proficient use of aircraft resources by operating crew" (Johnston, 1993, p. 371). With mounting evidence that NTSs play a significant role in aviation accidents – and, conversely, an important role in the creation of resilience in the face of novel threats (Dekker & Lundström, 2007; Nijhof & Dekker, 2009) – the International Civil Aviation Organization (ICAO) has called for an increased integration of NTS training for airline flight crew (Maurino, 1996). Some airlines have more or less successfully incorporated NTS training. But developing meaningful assessment of such skills, and assuring the quality and fairness of its assessment, has proven elusive for many (Flin et al., 1998; Munro & Mavin, 2012; Rigner & Dekker, 2000). In fact, there appear to be substantial global variations in the implementation and maturity of NTS teaching and assessment. For instance, the use of the NOTECHS behavioral markers enables examiners to assess NTSs such as cooperation, leadership and management, situational awareness, and decision making (Flin & Martin, 2001). In Europe, NOTECHS is well integrated into airline practice, whereas in countries like Australia, assessing NTS has only been implemented recently; and, in New Zealand, such implementation has yet to be mandated.

Although regional differences exist between how NTS is assessed, a common trend is to separate technical skills from NTSs. However, the increasingly complex automation onboard modern flight decks has eroded the strict separation between technical and nontechnical skills (Dekker & Orasanu, 1999; Dekker & Woods, 1999; Sarter & Woods, 1995, 1997). Automation skills in a modern cockpit are as technical as they are nontechnical, requiring the careful coordination of flight parameters, navigation, associated

DOI: 10.1027/2192-0923/a000041

call-outs, tasks, and double-checks. Successfully operating an automated airliner is as much about the technical knowledge of the system and how it works, as it is about how to work that system in concert with the other crewmember's goals, knowledge, and mindsets. In reprogramming an aircraft's flight path for an approach in the flight management system, for instance, and doing this in a way that involves both crewmembers as well as air traffic control, it is hard to say where technical skills end, and nontechnical skills begin.

In addition to the capacity to operate under normal conditions in the modern technically advanced aircraft, the adaptive capacity required by crew cannot be overlooked. The ability to recognize threats and errors does not just come from rehearsing separate packages of technical and nontechnical skills (Rouse & Morris, 1986). It emerges from crew activities directed at recognizing and adapting to these threats and errors. It requires crews both to recognize any shortfall in the system's existing expertise and to develop subsequent strategies to deal with the problem (Rochlin, LaPorte, & Roberts, 1987). Any separation of NTSs and technical skills, then, runs the risk of missing those aspects of flight-deck performance that arise from the interaction of technical and nontechnical skills.

This separation of technical skills from NTSs for assessment (or flight safety) purposes does not reflect how pilots assess performance (Mavin & Dall'Alba, 2010, 2011). A recent critique by an airline outlined that separating technical skills and NTSs can be confusing to pilot-examiners who are not entrenched in academic jargon, and whose prime function is to unpack the performance of other pilots for appraisal and debriefing purposes (Munro & Mavin, 2012). For this reason, an alternative model has been proposed (Mavin & Dall'Alba, 2010). It integrates six criteria related to technical (e.g., aircraft flown within tolerances and aviation knowledge) and NTSs (including situational awareness, decision making considerate of risk, management of crew, and communication) (see Figure 1).

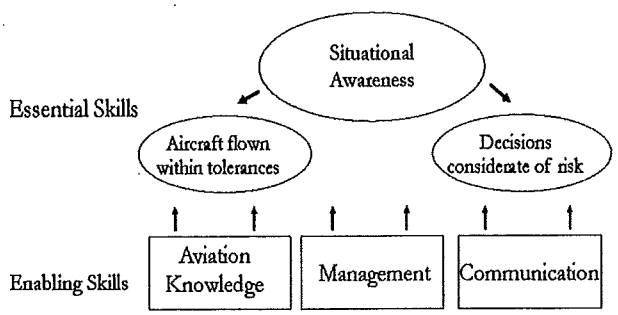


Figure 1. Model for assessing a pilots' performance. Adapted from *A Model for Integrating Technical Skills and NTS in Assessing Pilots' Performance,* by T. J. Mavin & G. Dall'Alba, 2010, Paper presented at the 9th International Symposium of the Australian Aviation Psychology Association, Sydney, Australia. Copyright 2010 by T. J. Mavin.

Rather then treating skills separately along technical and nontechnical dimensions, this model for assessing a pilots' performance (MAPP) shows performance to integrate "essential" and "enabling" dimensions, each consisting of a mixture of technical and nontechnical skills. Essential skills are those skills required by crew for minimum performance. It is suggested that final decisions by examiners are made on the basis of essential criteria. A notable problem under any of these criteria would indicate to an examiner that a pilot is below a minimum standard. For example, viewing the performance of a pilot demonstrating evidence of losing situational awareness, not being able to maintain the aircraft within specific limits laid down by the airlines (e.g., bank, speed, altitude, configuration, etc.), making decisions considered too conservative/risky, or being unable to make a decision would be considered *not proficient.* Conversely, a pilot who exhibits sound technical knowledge, good management skills, and outstanding communication skills (all enabling skills), yet continues to exceed company limits during flying exercises, would also be deemed not proficient. As such, the MAPP provides a hierarchy of integrated skills, which has essential skills at the top.

A number of airlines have adopted the MAPP and use it as a framework during debriefing. It gives pilots and examiners a clear understanding of how deficiencies in areas such as enabling skills can affect essential skills, thus determining areas of need. One airline in particular has taken performance assessment training one step further by including not only examiners but also all of the other pilots it employs. Pilots are trained in using the model and an associated assessment form to rate the performance of pilots at work (e.g., shown in videotaped simulator scenarios) and, in so doing, improve their own practice by considering it through the lens of the model/assessment instrument. During this initial trial, the airline and participants permitted the researchers to gather data during the assessment process, for the purpose of identifying any variations that existed between performance assessments by those of different rank. The practical aim of the study was to generate insights for improving training benefits during debriefing. For example, having first officers viewing performance differently from examiners may set up an initial difficulty in arriving at a shared understanding of required performance. This may be exemplified during promotion training when first officers undergo preparation to become captains. It is on this difference in performance assessment that this paper focuses.

The present study was designed to better understand (a) whether there are differences in how pilots of different rank assess the performance of peers shown in video-recorded simulator performances and (b) what the sources of variation are underlying any observed within- and between-rank differences in performance assessment.

Study 1

Study 1 was designed to answer the research question,

	1	2	3	4	5
Essential Skills					
Aircraft maintained within tolerances • Manual flight • Automation • Monitoring	• Manipulative skills resulted in frequent or sustained deviations outside allowable tolerances. • Automatic system use led to aircraft exceeding tolerances. • Frequent mistakes or missed calls made in monitoring.	• Aircraft manipulated to limit of tolerances, or, slightly exceeded tolerance, immediately corrected. • Inappropriate use of automated systems, though tolerances maintained. • Significant mistakes or lapses in monitoring.	• Manipulated with some deviation from target parameters, though quickly recovered. • Appropriate use of automated systems with few errors. • Minor lapses or mistakes in monitoring.	• Manipulated accurately, with only occasional variation from target parameters, quickly corrected. • Correct and appropriate use of automatic systems. • Appropriate and timely monitoring.	• Manipulated accurately, with no deviations from target parameters. • Totally appropriate use of automated systems at all times. • Monitoring well carried out with timely calls.
Enabling Skills					
Management • Workload • Control • Cooperation • Threats & Errors	• Ineffective organisation of crew tasks. • Inability to control self or crew member performance. • Interaction was negligible, or disrupted team effectiveness. • Serious threats or errors not mitigated or managed.	• Inefficient organisation of crew tasks. • Controlled self or crew member actions, though with difficulty. • Interacted with crew member, but provided limited support. • Threats or errors not well mitigated or managed.	• Adequate organisation of crew tasks. • Controlled self or crew members performance; disagreements resolved. • Interacted with crew member. • Most threats managed; most errors trapped.	• Crew member tasks effectively organised. • Effective control of self or crew to achieve expected performance. • Considered other crew to improve team performance. • Threats identified and managed; errors trapped.	• Tasks organised so challenging aspects of flight appeared easy. • Effective control of self or crew, even in a challenging situation. • Interaction with and consideration of crew maximised performance. • TEM well integrated.

Figure 2. Assessment criteria sheet depicting essential skills of aircraft flown within tolerance, and enabling skill of management. TEM = threat and error management.

Research Question (RQ1): Does pilot rank in an airline determine the assessment of other pilots' performances?

Method

Participants

All 92 pilots of a regional airline participated in this study. This included 9 flight examiners (FEs), 42 captains (CAPs), and 41 first officers (FOs). The pilots had no prior inter-rater reliability training.

Materials

Over a 12-month period, the airline reviewed its assessment program, where the assessments of technical skills and NTSs were integrated using the MAPP. To facilitate greater detail regarding the assessments, a single-sided assessment form (A4) was designed based on the MAPP. The assessment form consisted of a 6 × 5 grid where the six criteria from the MAPP made up the vertical dimension (situational awareness, decision making, aircraft maintained within parameters, aviation knowledge, management, and communication). Conforming to the airline's preference, each criterion had a performance rating from 1 (= *very poor*) to 5 (= *very good*). The ratings formed the horizontal dimension of the assessment grid. To assist users of the grid, word pictures for each dimension were developed. For example, the criterion "situational awareness" included "perception," "comprehension," and "projection." A rating of 1 included the word pictures "Lacked awareness of clearly obvious systems or environmental factors," "Misinterpreted or did not comprehend factors affecting flight safety," and "Did not predict future events, even those obvious to flight safety." Figure 2 shows details of the assessment grid for the essential skill of aircraft flown within tolerance and the enabling skill of management.

In addition to the assessment grid, 15 flight scenarios were recorded in the company's simulator. Two company pilots in the scenarios wore company uniform. They took turns in acting as captain or first officer. Each video was scripted and recorded using three cameras. A cabin crewmember also participated in some videos. For Study 1, the 15 videos where analyzed to identify three flight scenarios that depicted poor, good, and average performance (see Table 1 for video description).

Design

Study 1 took place as part of the company's annual NTS/CRM training day, serving as a refresher for the modules situational awareness, decision making, management, and communication. Class size was 6–10 pilots, with a mix of captains and first officers. The morning saw PowerPoint presentations featuring descriptions of the MAPP and a review of the six dimensions of the assessment form. Classroom discussion of the theory underlying each dimension occurred. During the afternoon session, each pilot assessed three previously chosen videos (Table 1). Pilots were asked to assess overall performance of the crew in each video, without discussion with other pilots. On completion of each video, the classroom instructor led a discussion. The aim of the discussion was to obtain the grades awarded by each pilot and deliberate on the reasons for their assessment. All assessment forms were collected for analysis.

Table 1. Video description for Study 1

Video	Description
1A Poor	Aircraft positioned at bay in low visibility operations (fog – daylight). Pilots given specific taxi instructions by air traffic control. Whilst taxiing, cabin crew advises of ill passenger. Crew become distracted and taxi onto wrong taxiway. (Time: 3:14)
1B Good	Aircraft descending through 5,000 feet in instrument meteorological condition (IMC) at night. Captain is pilot flying (PF) and first officer pilot monitoring (PM). Aircraft suffers electrical malfunction with crew managing the malfunction well. (Time: 10:14)
1C Average	Aircraft flying at 17,000 feet in daylight. Aircraft suffers same electrical malfunction as in Video 1B, however, crew have difficulty diagnosing malfunction. (Time: 4:18)

Two extra modules for flight examiners were requested by the company: *briefing* and *debriefing*. The 2-day course included exactly the same material as the normal 1-day course except that on the afternoon of Day 1, there was a module on briefing techniques; on the afternoon of Day 2, there was a debriefing module. For the purpose of this study, apart from the briefing module on Day 1, all pilots received the same training prior to assessing the videos.

Results

We began this study with the hypothesis that experience would lead to differences in rating pilot performance. Given that no pilot in the company had undergone previous training in assessment, the only intervention was the training provided. Figure 3 shows means and standard deviations for the three scenarios by rank, which shows flight examiners being consistently harder than either the captains or first officers on all performance factors. Still, a comparison across the three scenarios showed that the ratings were, for all three ranks, relatively close together (within standard deviation) and, for the most part, exhibited the same profiles across the six performance factors. Moreover, the assessments differentiated between scenarios: The ratings of the pilots in Video 1B are substantially higher than those in Video 1A.

To find out whether flight examiners, captains, and first officers differed significantly in rating the three video clips (1A, 1B, and 1C), a multivariate analysis of variance (MANOVA) was used. In this procedure, all six areas – situational awareness (SA), decision making (DM), aircraft flown within tolerances (AC), aviation knowledge (KN),

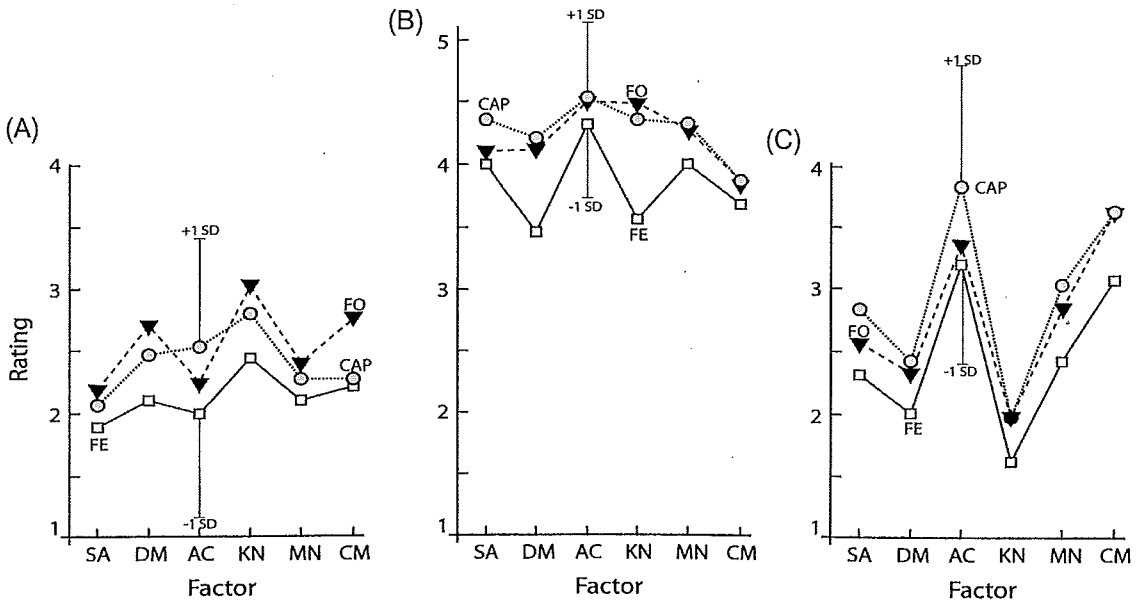


Figure 3. Mean performance ratings provided by flight examiners (FE), captains (CAP), and first officers (FO) using the model for assessing a pilots' performance (MAPP) assessment form in three flight scenarios: (A) Video 1A (poor performance), (B) Video 1B (good performance), (C) Video 1c (average performance). AC = aircraft flown within tolerances; CM = communication; DM = decision making; KN = aviation knowledge; MN = management; SA = situational awareness.

Table 2. Video 1B crosstab "decision making" (1BDM) against "experience"

	Decision-making (DM) grade					
	1.00	2.00	3.00	4.00	5.00	Total
Count	2	5	2	0	0	9
%	22.2%	55.6%	22.2%	0.0%	0.0%	100%
Count	5	16	15	5	1	42
%	12.0%	38.1%	35.6%	11.9%	2.4%	100%
Count	7	18	15	1	0	41
%	17.1%	43.9%	36.6%	2.4%	0.0%	100%
Count	14	39	32	6	1	92

management (MN), and communication (CM) – are treated as dependent variables that are tested simultaneously. Only if this test showed significant differences ($p < .05$) would the six variables be investigated individually. Significant differences between the ratings of the flight examiners, captains, and first officers were detected only for Video 1B, Wilks $\Lambda_{1A} = .817$, $F_{1A}(6, 83) = 1.473$, $p_{1A} = .139$; Wilks $\Lambda_{1B} = .751$, $F_{1B}(6, 84) = 2.513$, $p_{1B} = .016$; Wilks $\Lambda_{1C} = .868$, $F_{1C}(6, 83) = 1.013$, $p_{1C} = .439$. When the six dependent variables for Video 1B were investigated individually using an analysis of variance (ANOVA), significant differences existed on the factors of decision making ($p_{1BDM} = .007$) and knowledge ($p_{1BKN} = .01$). It is noted that situational awareness was $p_{1BSA} = .056$.

To better understand these findings, a crosstab procedure was applied to investigate the distribution of scores in each of the six categories for the three videos. The results showed that for almost all categories, there was a considerable distribution of scores. For example, when the 92 pilots rated Video 1A with respect to decision making (DM), 14 pilots rated the performance as 1, 39 as 2, 32 as 3, 6 as 4, and 1 as 5. That is, there was a wide variation of assessment, from failed to very good (Table 2).

When the ratings of flight examiners, captains, and first officers are compared for Video 1C, a slight trend appeared to exist for the first officers to be "harder" in the ratings. Thus, about 62% of the flight examiners and captains thought the performance on this factor was 4 or 5, whereas only 51% of the first officers thought that the performance was 4 or 5. We also note that the percentage of first officers rating this factor as 2 was more than double that of the flight examiners and captains. This pattern was repeated on the other criteria for this video and for the other two videos used in the training session. In some instances, all scores fell by on 3 scores (e.g., situational awareness Video 1B, all scores fell in the 3–5 range, with only 7.6% of the pilots giving a 3). On the other five criteria, one assessor apparently gave 1s or 2s, whereas all other ratings fell between 3 and 5 and the predominant number of scores fell between 4 and 5 (decision making 86.9%; aircraft within tolerances 96.8%; knowledge 93.5%; management 90.2%; and communication 76.1%). The ratings of the performances of the six criteria wandered even farther apart on Video 1C.

Study 2

To better understand the source of the variations observed in Study 1, we designed a think-aloud study to answer the question,

Research Question 2 (RQ2): What are some of the possible sources of variation that occur between pilot raters of differing experience levels?

Methods

Participants

The same airline as Study 1 was used. Three pairs of pilots at each rank of flight examiners, captains, and first officers ($N = 18$) participated. The selection of pilots was a function of the company roster. Pilots were randomly picked among those with free slots during the 7-day data collection period. Pilots had no prior inter-rater reliability training, but all had participated in the training workshop to rate performance using the MAPP assessment form.

Design

As in Study 1, pilots rated the performance of pilots in the videos in pairs. In addition, the pairs were asked to provide reasons for their assessments. Although a traditional think-aloud protocol (Ericsson & Simon, 1993) could have been chosen, we know from experience that practitioners tend to find tasks more natural when they give their reasons while talking to peers (Roth, 2005). Each pilot pair was required to assess (and provide reasons for the assessment) the flight-deck performances of pilots displayed on video scenarios not seen before (see Table 3 for video description). To stimulate verbalization, each pair was asked to come to an agreement about their assessment of the captain and first officer shown. To increase verbalization, only one assessment form was provided for each pilot in the video. Pilot pairs were encouraged to (a) discuss their assessment, with explicit references to their reasons behind the score and (b) provide evidence from the video that was linked with this reason. At the end of each assessment, pilots were asked whether there was anything in the performance that was relevant but not being assessed using the assessment form. Each pair was asked whether participants were confident that the rating reflected a fair judgment and whether the overall assessment reflected the quality of the performance viewed holistically. These questions were asked in light of the fact that within the participant airline, three 2s or one 1 on any of the six criteria means "fail," which requires a repeat of the assessment exercise.

Each session was video recorded using three cameras. One camera was above the pilot pair recording written notes and scores on the assessment form. A second camera was in front of the pair recording their faces, interactions,

Table 3. Video description for Study 2

Video	Description
2A	Captain (PF) and FO (PM) conduct instrument approach by day, becoming visual close to airport. During visual maneuvering to land, aircraft encounters rain, missed approach conducted, captain initially turning wrong direction, though corrected by FO. (Time: 6:45)
2B	FO (PF) and captain (PM) conducting instrument approach. Due to poor weather, crew conducts a missed approach. Fuel is low, requiring diversion to alternate airport. Captain attempts to convince FO to conduct one more approach. FO is reluctant. (Time: 3:30)
2C	Aircraft on descent in instrument meteorological condition (IMC) suffers engine fire prior to becoming visual. Captain (PF) elects to continue approach and land aircraft. Crew evacuates passengers after landing. (Time: 9:16)

Notes. FO = first officer; PF = pilot flying; PM = pilot monitoring.

and gestures. The third camera, like camera two, recorded the pair interaction, plus the flight videos being currently viewed. The final video was cut to a picture-in-picture view for analysis (see Figure 4).

The picture-in-picture think-aloud protocols were transcribed word for word to capture the content of the talk, with both being used to conduct a fine-grained level of analysis. Video and transcript serve as natural protocol of the assessment sessions (Roth, 2007). Analysis was based on interaction analysis (Jordan & Henderson, 1995), a method whereby teams of researchers interactively analyze the interactions on the videotape (pilots assessing the video).

Results

This think-aloud study was conducted to understand the variation in the assessments observed during Study 1. To answer our research question, we sought answers to two subordinate questions: (a) What are the phenomena that different pilot pairs identify and rate? and (b) What are the reasoning patterns with which they arrive at a conclusion?

As in Study 1, ratings provided during our think-aloud protocol varied considerably even within level of expertise (FE, CAP, and FO). Thus, for most factors on all three videos, the minimum and maximum score differed by 2 to 3 points. For example, in Video 2A, scores for the criterion situational awareness for all nine pairs of flight examiners, captains, and first officers, mean ratings were $X_{FE} = 2.33$ ($SD_{FE} = 0.58$), $X_{CAP} = 2.00$ ($SD_{CAP} = 1.00$), and $X_{FO} = 3.00$ ($SD_{FO} = 0.00$), respectively. Overall, the mean ratings of the captains' performance, for each of the six criteria, in Videos 2A, 2B, and 2C, tended to be lowest for flight examiners and highest for the first officers.

The analysis of the transcripts revealed that there were considerable variations between the groups both within and across ranks about what had gone wrong or how one performance aspect had led to a chain of events that had caused problems elsewhere in the video scenario as a whole. For example, Video 2C featured an engine fire,

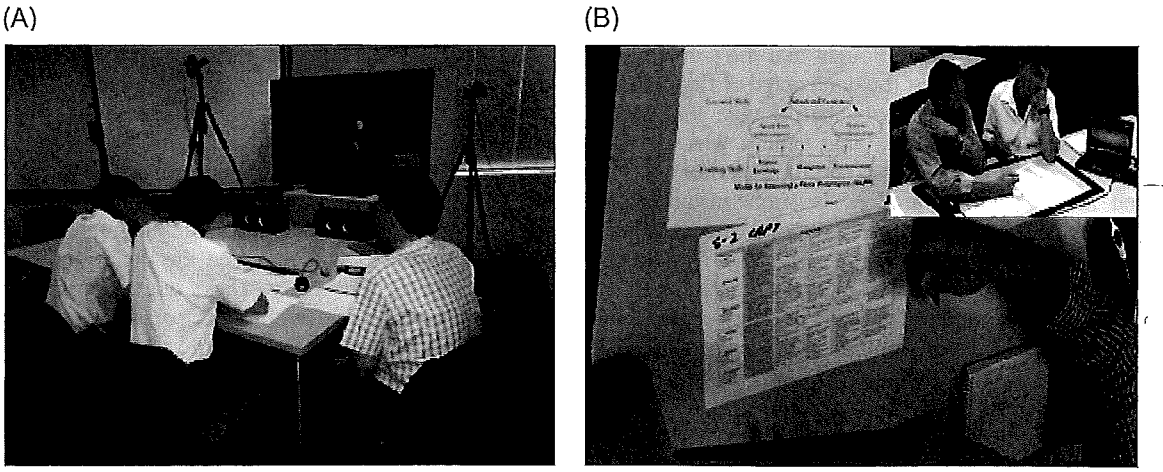


Figure 4. Pilot pair assessing a video in the left image (A) with the researcher sitting on the right. Picture on the right (B) showing picture-in-picture image used for analysis.

Table 4. Assessment of Video 2A by pilot rank

Pilot pair	SA	DM	AC	KN	MN	CM	Pass/Fail
FE 1	2	2	2	2	2	3	F
FE 2	3	1	3	2	3	4	F
FE 3	2	3	2	2	2	4	F
M	2.33	2.00	2.33	2.00	2.33	3.67	
SD	0.58	1.00	0.58	0.00	0.58	0.58	
CAP 1	2	3	3	2	2	3	F
CAP 2	1	2	3	1	2	3	F
CAP 3	3	2	4	2	3	3	P
M	3.00	2.00	4.00	2.00	3.00	3.00	
SD	1.00	0.58	0.58	0.58	0.58	0.00	
FO 1	3	4	3	3	3	4	P
FO 2	3	4	3	4	4	4	P
FO 3	3	2	3	3	3	3	P
M	3.00	3.33	3.00	3.33	3.33	3.67	
SD	0.00	1.15	0.00	0.58	0.58	0.58	

Notes. Captain pair 2 (CAP2) and CAP3 (*shaded area*) differed in final scores. Their pass/fail reasons for SA are taken up in Table 5. AC = aircraft flown within tolerances; CAP = captain; CM = communication; DM = decision making; FE = flight examiner; FO = first officer; KN = aviation knowledge; MN = management; SA = situational awareness.

landing, and evacuation. Whereas the video had been picked because of the general highly competent performance – reflected in the assessments across ranks – one group of flight examiners discovered that the crew had failed to close a small lever on the throttle quadrant of the functioning engine; which would impede the evacuation. This had not been identified by any of the other pairs; even the researchers had failed to note this failure. Based on this fact, the pair ended up failing both flying and nonflying pilots whereas all other pairs rated the performance of the crew as in the good to very good range.

As another example of the generally good reasons for varying assessments, we considered the following aspect from Video 2A. The flight examiners all considered the performance of the captain to be a fail; two of the captains agreed, though one pair considered it a pass (though there were two of the necessary three 2s for a failure rating). However, all of the first officer pairs considered the performance a pass (see Table 4). To illustrate the source of variance in this assessment, we drew on the protocols from two pairs of captains (CAP2 and CAP3) rating situational awareness of the captain (see Table 4, shaded area), which featured the aircraft flying into a rain shower on final turn before landing. For the pair CAP2, there were "two gaping holes" that led to the ultimate decision to score a 1 on situational awareness: a missed minimum descent altitude call, and no plan-of-action following a missed approach call (Table 5). The pair identified a number of associated issues: (a) did not push go-around button, (b) no go-around Flap 15 call, (c) likely failure to plan during brief, and (d) missed altitude capture. In contrast, pair CAP3 made note of the surprise and the likelihood of a failure to plan a possible missed approach. The pair noted some loss of situational awareness but felt that the captain was aware of the poor weather, noting only "some" difficulty predicting future events. In Video 2A, all flight examiners and two captain pairs tended to evaluate in the direction of pair CAP2, all of the FO groups rated the performance as pair CAP3 had.

Discussion and Conclusion

Studies 1 and 2 investigated how pilots within an entire airline assess the performance of other pilots. Both studies used assessment criteria that integrated technical and nontechnical skills (or NTSs). Study 1 showed differences between flight examiners, captains, and first officers, though only Video 1B was considered significant in the ratings of decision making and knowledge. However, no significant difference existed in the context of other scenarios. Given that a discussion did occur after each video, some polarization or group thinking may have occurred, though this would have not been evident in Video 1A.

Study 2 showed that in some instances, the difference was linked with the experience of the pilot making the assessment. For instance, in Table 4 three flight examiner pairs considered the captain's performance in Video 2A to be below standard. However, within the captain group, two pairs agreed with the flight examiners' judgments, with one captain pair considering the performance to be acceptable. On the other hand, all three first officer pairs considered the performance of the captain in the video to be acceptable. These findings, though they may appear surprising, replicate other studies in aviation (Brannick, Prince, & Salas, 2002) that also show that this variation can differ across the criteria measured (Holt, Hansberger, & Boehm-Davis, 2002).

Study 2 provided a better understanding of where between-rank differences may appear and why. Thus, when a performance is held to be unproblematic, such as in the case of Video 2C where all but one pair of flight examiners considered the performance to be good to very good, then the ratings of flight examiners, captains, and first officers tend to be aligned. When there are problems in the course of a flight scenario, however, pilots differ with respect to the facts they pick out, how these facts are assessed in terms of skill level, and how these facts explain the overall performance. Thus, the flying pilot's utterance about the direction in which to turn following the announcement of a missed approach may be interpreted as evidence of (a) the quest for affirming the direction and therefore desirable communicative environment, (b) lack of situational awareness, or (c) the result of lack of planning and debriefing. In each case, the pairs provided good reasons for the particular assessment they made.

Taken together, the results of Studies 1 and 2 may have a bearing on understanding inter-rater reliability. Overall, inter-rater reliability training may initially be considered as the panacea for the above problem. For instance, during the 1980s, the efficacy of rater training was questioned,

Table 5. Main points from the evaluation of situational awareness of performance in scenario of going IMC and missed approach

Situational Awareness	
Group CAP2	Group CAP3
– Two gaping holes in situational awareness. – Mention about strong tail wind, or cross wind. – Didn't hear a "go around positive." – He obviously didn't push the go-around button, because flight director bars weren't consistent with what he was flying – go-around procedure begins with go-around call. – No go-around Flap 15 call. – Did ask for check power. – Botched go-around. – Situational awareness issues barely acceptable for a simulator check situation. – Missed MDA and what way to turn on missed approach, important because on that approach there is high terrain, hence necessity to turn the correct way. MDA is critical. – He did ask what MDA was and did ask what way to turn. – Late setting missed approach altitude indicates that he wasn't predicting the go-around; indication of difficulty predicting future events. – As flying pilot, completely missed altitude capture. – Lacked awareness of clearly obvious systems or environmental factors. – Two things you must know are minimum altitude and what to do if missed approach.	– Got onto the downwind. – Had feeling that weather conditions weren't good. – Cloud [IMC] was a bit of surprise. – Started to lose situational awareness, had feeling weather might not be good, had a little bit of confusion. – Likely hadn't developed a plan during brief, for there was confusion when turning left (toward mountain) rather than right. – Some difficulty predicting future events. – Flight well handled, approach well flown, kind of aware that downwind it didn't look good. – Had an idea of what to do for missed approach, execution was somewhat problematic.

Notes. CAP = captain; IMC = instrument meteorological condition; MDA = minimum descent altitude.

citing little evidence that raters changed their ratings post training (Bernardin & Buckley, 1981; Bernardin & Pence, 1980; Borman, 1978). These views have slowly changed (Woehr & Huffcutt, 1994), with more positive findings in recent years (Brannick et al., 2002; Clauser, Clyman, & Swanson, 1999; Gorman & Rentsch, 2009). However, an important issue has been raised in these studies. If inter-rater reliability training has the aim of standardizing the way individuals assess performance, without regard to the underlying reasons for the judgments made, does this not question the validity of the inter-rater reliability measure itself? The present study showed not only that there are considerable variations in assessment *in those instances where there is trouble* but also that assessing pilots have good reasons for varying in their assessments. Any trouble in a cockpit is not a stand-alone fact but occurs in the context of a complex performance of an unfolding sequence of events. Thus, there are many aspects of the flight that eventually may become problematic. Any one aspect (instant) could have corrected the flight (parameters) and led to a different outcome. Pilots differ in how they themselves would have acted, at what time, and based on what fact. In our database, there is evidence for very good reasons there are variations in the performance ratings on any one-performance factor.

Whereas Study 1 did not find significant differences between how pilots of different rank view performance, it would be hard to ignore the considerable variations observable in Figure 3. Given these variations, any between-rank differences would fail to be significant given large error variances in the numerator of the associated statistical significance tests. We may therefore retain the question of whether pilots of different rank actually view performance differently. If they do so, should this be ignored as a function of a typical career trajectory of a normal pilot, or should training interventions be introduced in an attempt to reduce this perceived gap? Such training may be like the one presented in this study, where assessment training is introduced across all ranks in the attempt to align how pilots view performance. Study 2, an exhaustive analysis of which is still to be conducted, will provide some answers toward this question. An appreciation of how pilots of different rank view performance will greatly enhance how performance is discussed in the aviation classroom, debriefing room, or more broadly within an airline.

References

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6*(2), 205–212.

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology, 65*(1), 60–66.

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

Brannick, M. T., Prince, C., & Salas, E. (2002). The reliability of instructor evaluations of crew performance: Good news and not so good news. *International Journal of Aviation Psychology, 12*(3), 241–261.

Clauser, B. E., Clyman, S. G., & Swanson, D. B. (1999). Components of rater error in a complex perfomance assessment. *Journal of Educational Measurement, 36*(1), 17.

Dekker, S. W. A., & Lundström, J. T. (2007). From threat and error management (TEM) to resilience. *Human Factors and Aerospace Safety, 6*(3), 261–274.

Dekker, S. W. A., & Orasanu, J. M. (1999). Automation and situation awareness: Pushing the research frontier. In S. W. A. Dekker & E. Hollnagel (Eds.), *Coping with computers in the cockpit* (pp. 69–85). Aldershot, UK: Ashgate.

Dekker, S. W. A., & Woods, D. D. (1999). Automation and its impact on human cognition. In S. W. A. Dekker & E. Hollnagel (Eds.), *Coping with computers in the cockpit* (pp. 7–27). Aldershot, UK: Ahsgate.

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

Fischer, U., Orasanu, J. M., & Montvalo, M. (1993). *Efficient decision strategies on the flight deck.* Paper presented at the 7th International Symposium on Aviation Psychology. OH: Columbus.

Flin, R., Goeters, K., Hoermann, H., & Martin, L. (1998). *A generic structure of non-technical skills for training and assessment*, Paper presented at the 23rd conference of the European Association for Aviation Psychology, Vienna.

Flin, R., & Martin, L. (2001). Behavioural markers for crew resource management: A review of current practice. *International Journal of Aviation Psychology, 11*(1), 95–118.

Gorman, C. A., & Rentsch, J. R. (2009). Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. *Journal of Applied Psychology, 94*(5), 1336–1344.

Holt, R. W., Hansberger, J. T., & Boehm-Davis, D. A. (2002). Improving rater calibration in aviation: A case study. *International Journal of Aviation Psychology, 12*(3), 305–330.

Johnston, A. N. (1993). CRM: Cross-cultural perspectives. In E. L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 367–398). San Diego, CA: Academic Press.

Johnston, A. N., Rushby, N., & Maclean, I. (2000). An assistant for crew performance assessment. *International Journal of Aviation Psychology, 10*(1), 99–108.

Jordan, B., & Henderson, A. (1995). Interaction analysis: Foundations and practice. *Journal of the Learning Sciences, 4*, 39–103.

Maurino, D. E. (1996). Eighteen years of CRM wars: A report from headquarters. In B. J. Hayward & A. R. Lowe (Eds.), *Applied aviation psychology: Achievement, change and challenge* (pp. 99–109). Aldershot, UK: Avebury Aviation.

Maurino, D. E., Reason, J. T., Johnston, A. N., & Lee, R. B. (1995). *Beyond aviation human factors*. Aldershot, UK: Avebury Aviation.

Mavin, T. J., & 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. Australia: Sydney.

Mavin, T. J., & Dall'Alba, G. (2011). *Understanding complex assessment: A lesson from aviation.* Paper presented at the 4th International Conference of Edcuation, Research and Innovations. Spain: Madrid.

Munro, I., & Mavin, T. J. (2012, November). *Crawl-Walk-Run.* Paper presented at the 10th International Symopsium of the Australian Aviation Psychology Association. Australia: Sydney.

Nagel, D. C. (1988). Human error in aviation operations. In E. L. Wiener & D. C. Nagel (Eds.), *Human factors in aviation* (pp. 263–303). San Diego, CA: Academic Press.

Nijhof, M., & Dekker, S. W. A. (2009). Restoration through preparation, is it possible? Analysis of a low-probability, high-consequence event. In E. Hollnagel, C. P. Nemeth, & S. W. A. Dekker (Eds.), *Resilience engineering perspectives: Preparation and restoration* (pp. 205–214). Aldershot, UK: Ashgate.

Orasanu, J. M. (1990). *Shared mental models and crew decision making*. Princeton, NJ: Cognitive Science Laboratory, Princeton University.

Orasanu, J. M. (2001). *The role of risk assessment in flight safety: Strategies for enhancing pilot decision making.* Paper presented at the 4th International Workshop on Human Error. Sweden: Safety and Systems Development Linköping.

Orasanu, J. M., & Martin, L. (1998). *Errors in aviation decision making: A factor in accidents and incidents. Human Error, Safety and Systems Development Workshop (HESSD) 1998.* Retrieved from http://www.dcs.gla.ac.uk/~johnson/papers/seattle_hessd/judithlynnep

Rigner, J., & Dekker, S. W. A. (2000). Sharing the burden of flight deck automation training. *International Journal of Aviation Psychology, 10*(4), 317–326.

Rochlin, G. I., LaPorte, T. R., & Roberts, K. H. (1987). The self-designing high reliability organization: Aircraft carrier flight operations at sea. *Naval War College Review, 40*, 76–90.

Roth, W.-M. (2005). *Doing qualitative research: Praxis of method.* Rotterdam, The Netherlands: Sense Publishers.

Roth, W.-M. (2007). *Doing teacher research: A handbook for perplexed practitioners*. Rotterdam, The Netherlands: Sense Publishers.

Rouse, W. B., & Morris, N. M. (1986). On looking into the black box: Prospects and limits in the search for mental models. *Psychological Bulletin, 100*(3), 359–363.

Sarter, N. B., & Woods, D. D. (1995). "How in the world did we get into that mode?" Mode error and awareness in supervisory control. *Human Factors, 37*(1), 5–19.

Sarter, N. B., & Woods, D. D. (1997). Teamplay with a powerful and independent agent: A corpus of operational experiences and automation surprises on the Airbus A320. *Human Factors, 39*, 553–569.

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.

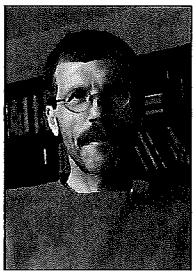
Woods, D. D., & Patterson, E. S. (2000). How unexpected events produce an escalation of cognitive and coordinate demands. In P. A. Hancock & P. Desmond (Eds.), *Stress, workload and fatigue* (pp. 290–304). Mahwah, NJ: Lawrence Erlbaum Associates.

Woods, D. D., Patterson, E. S., & Roth, E. M. (2002). Can we ever escape from data overload? A cognitive systems diagnosis. *Cognition, Technology & Work, 4*(1), 22–36.

Accepted for publication: May 3, 2013
Published online: November 29, 2013

Original Article

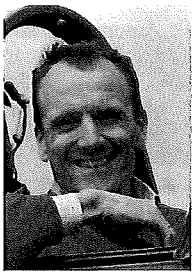
Timothy J. Mavin (EdD) is an associate professor at Griffith University, Brisbane. He has 10,000 hours of flight experience, including 7,000 hours of jet time. He continues to conduct type-rating endorsements on the Boeing 737. He is also a qualified high school teacher. He researches skills assessment and training in airline operations.



Wolff-Michael Roth (PhD, 1987) is Lansdowne Professor of Applied Cognitive Science at the University of Victoria, Canada. He studies knowing and learning across the lifespan from interdisciplinary perspectives. His work includes *Passibility: At the Limits of the Constructivist Metaphor* (Springer, 2011) and *Meaning and Mental Representation: A Pragmatic Approach* (Sense Publishers, 2013).

Correspondence Address

Tim Mavin
Aviation at Griffith
Griffith University
N44 Office: 3.24
170 Kessels Road
Nathan, QLD 4111
Australia
Tel. +61 7 3735-4404
E-mail t.mavin@griffith.edu.au



Sidney Dekker (PhD, The Ohio State University, 1996) is a professor at Griffith University in Brisbane where he runs the Safety Science Innovation Lab. Author of several best-selling books on human error and safety, he has recently been active as an airline pilot on the Boeing 737.

Original Article