Exploring the Use of Categories in the Assessment of Airline Pilots' Performance as a Potential Source of Examiners' Disagreement

David E. Weber, Timothy J. Mavin, Griffith University, **Wolff-Michael Roth**, University of Victoria, **Eder Henriqson**, Pontifícia Universidade Católica do Rio Grande do Sul, and **Sidney W. A. Dekker**, Griffith University

It is a current trend in aviation to use categories of technical (e.g., knowledge) and nontechnical skills (e.g., situation awareness) to assess airline pilots' performance. Several studies have revealed large disagreement between assessors when airline professionals use these categories to assess the performance of their peers. The aim of the present study is to investigate whether the categories themselves are at the source of disagreement. We explore the reasoning of flight examiners who assess an engine fire scenario in pairs. The results provide insight into the overlap of topics that constitute certain categories. Implications are drawn in regards to the use of assessment categories and their influence on pilot performance assessment.

Keywords: performance assessment, pilots' performance, validity, situation awareness

INTRODUCTION

It is current practice in aviation to use a range of technical and nontechnical skills (NTS) to assess the performance of airline pilots (Mavin & Dall'Alba, 2010). Particularly the assessment of NTS (i.e., situational awareness, teamwork, or management) has proven difficult (Orlady & Orlady, 1999). (In the present study, we refer to "assessment" as the flight examiners' act of making judgments about a crew member's performance, which are reflected in the examiners' reasoning and scores provided.) Pursuing the aim of creating a valid and reliable assessment tool, behavioral markers have been developed that split performance into measurable components (Helmreich & Foushee, 1993).

A similar approach to performance assessment, which also builds on behavioral markers, is the NOn-TECHnical Skills (NOTECHS) project (Flin et al., 2003; O'Connor et al., 2002). NOTECHS provided assessors with a framework that can be used to assess pilots' NTS. The key categories are decision making, situational awareness, leadership, and teamwork (O'Connor et al., 2002). [The same terminology will be used as in NOTECHS, yet the categories will slightly differ from this framework (see "Methods," "Procedure"). In the following, the term *category* will specifically be used to refer to the assessment dimensions that the participating airline uses in its formal assessment process: Situation Awareness, Decision Making, Aircraft maintained within tolerances, Knowledge, Management, and Communication. Note that the analysis required the addition of one further category, called "Initial" (see "Methods," "Transcription and Coding").] The usability and (interrater) reliability of the NOTECHS framework was evaluated in the JARTEL project (JARTEL, 2002; O'Connor et al., 2002), which was probably one of the largest investigations into the validity of assessment categories. Investigating a number of hypotheses and associated assumptions, JARTEL (2002) found NOTECHS "capable of providing itself a valid and reliable method for assessing NTS" (p. 17) and "a useful and usable tool for the instructors" (p. 20).

Studies that seek to validate commonly used categories to assess pilot performance largely

Address correspondence to David Emanuel Weber, Safety Science Innovation Lab/Aviation, Griffith University, Nathan Campus, 170 Kessels Road, Nathan, Brisbane, QLD 4111, Australia, david.weber@griffithuni.edu.au.

Journal of Cognitive Engineering and Decision Making
201X, Volume XX, Number X, Month 2014, pp. 1–17
DOI: 10.1177/1555343414532813

focus on the scores provided by the assessors. Current research, however, indicates that assessors' reasoning behind the same or very similar scores can largely vary (Weber, Roth, Mavin, & Dekker, 2013). Assessors seem to arrive at the same scores for entirely different reasons. The resulting question is whether assessors' disagreement in terms of both the scores and reasoning might be the result of the low (discriminant) validity of the categories themselves. The literature lacks qualitative investigations that examine the reasoning behind the scores.

The aim of the present effort is twofold: (1) we examine whether the categories used to assess the performance of airline professionals are specific, in the sense that flight examiners build their assessment on the same observations and justifications; (2) we investigate what might distinguish the sort of observations that are specific to a single assessment category from observations that are stated in multiple categories.

METHODS

Participants

The present study forms part of a larger series of full-flight simulator and debriefing trials over a period of 2 years, involving a total of 18 assessment sessions (assessment of three scenarios per session) and $N = 36$ pilots, including flight examiners, captains, and first officers (FOs). For the results of the present study, we focused in on six participating flight examiners, who are not only formally qualified and the most experienced at assessing pilot performance but who also have the power to append career progress consequences to their assessments. Constraints in regards to the sample size are not unique to this study alone. Given constraints and opportunities on aviation research today, there is a growing legitimacy and acceptance of the use of single-case research designs (Whitehurst, 2013).

The mean age of the flight examiners was 49.2 years (*SD* = 6.2) and they had a mean of 25.3 years (*SD* = 5.9) of commercial flying experience with a mean of 14,250 flight hours (*SD* = 4,910). All participants worked for the same airline and were randomly picked among those with free slots during the 1-week data collection period. To encourage the verbalization of their reasoning, all participants assessed performance in pairs while using a single rating sheet.

Video Scenario

Each pair watched, discussed, and assessed a videotaped scenario of a captain and FO flying in a Dash-8 simulator. The two pilots wore company uniform. Another employee acted as the tower ATC controller. The video was scripted in advance and recorded from the position of the flight examiner. A first camera captured the pilots from behind, whereas a second and third camera provided pertinent close-up recordings of important instrument displays. At several points during the video, relevant charts were superimposed.

The length of the video was 09:16 min. It showed an approach to an airport, with which the pilots of the participating airline were familiar. During the approach, when the aircraft was already configured for landing (flaps set, landing gear down), the left-hand engine (#1) caught fire. The pilots followed their procedures to extinguish the fire and continued on the approach. They informed the cabin before touchdown, landed the aircraft, and evacuated the passengers on the runway after having received the clearance from the tower controller.

Procedure

The pairs watched the video on a large LCD TV screen, which they controlled via a computer with a mouse (Figure 1). The assessment sessions were video-recorded from three perspectives: CAM1 was positioned in front of the flight examiners, recording their interaction, together with the laptop that presented the scenario currently being watched. CAM2 provided a closer view of the flight examiners. In contrast, CAM3 recorded the area of work from above, capturing, for example, the notes taken, scores provided, and the pilots' pointing gestures toward certain word pictures on the assessment forms.

Flight examiners were given a booklet to take notes. Furthermore, each pair received two sheets from their company-training manual: the model the participating airline uses to assess their pilots and an assessment form derived from this model (one for assessing the captain, and another for the FO).

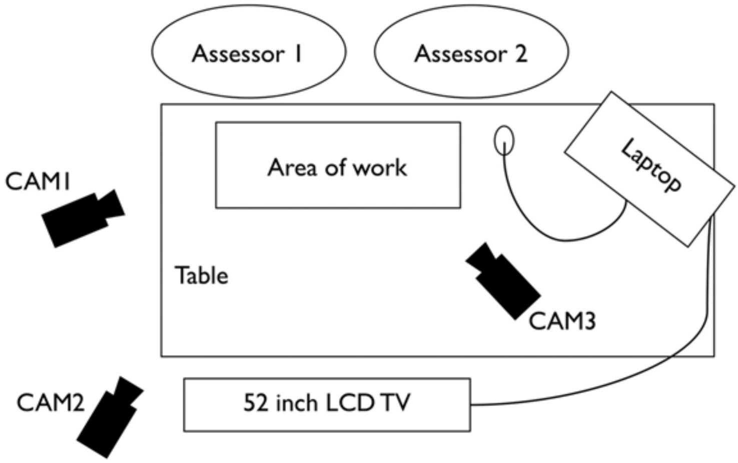

Figure 1. Study setting.

As part of their CRM training courses, all flight examiners had been extensively trained to use the model and assessment form. A PowerPoint presentation was shown to explain the six assessment categories in detail. Time was given for questions and discussion. Prior to participating in the study, all flight examiners had assessed the performance of pilots featured in at least three videos other than those they were rating for the present study.

The assessment categories of the model used by the participating airline are slightly different from the NOTECHS framework, including "Situational Awareness" (SA), "Decision Making" (DM), "Aircraft maintained within tolerances" (AC), "Knowledge" (KN), "Management" (MN), and "Communication" (CM). In contrast, the assessment form provided word pictures, which are descriptions of performance for each of the six categories and scores (Table 1), ranging from 1 to 5 (1 being the worst performance; 5 being the best performance).

Task

The pairs were informed that the study aimed at understanding the reasoning behind their assessment of pilot performance. The researchers emphasized the importance of explicitly articulating the reasons and thoughts for a specific assessment. During the assessment, flight examiner pairs were free to pause, replay, or go back to the video at any time and how often they wanted. The decision of where to start the discussion and how to assess performance was entirely left to the flight examiners. Two researchers attended the assessment sessions. Following a fixed protocol, they only intervened when it was necessary—for example, to encourage participants to speak louder or clarify a comment (i.e., to provide reasons for a decision when these were not made explicit).

Transcription and Coding

After the data collection, the assessment sessions were transcribed by a commercial transcribing service. Two of the authors checked and corrected the transcriptions. All pairs assessed one category of the assessment model after the other. This allowed a distinction between seven categories: before pairs started to use the assessment model, they all had an initial discussion about the scenario. Subsequently, the code "initial [INI]" was used as a marker for items that appeared in the assessment talk prior to a pair's actual assessment of one of the six categories of the assessment model. Following the initial discussion, pairs assessed the six categories of the assessment model: SA, DM, AC, KN, MN, and CM.

Several supplements were added to the transcripts: (a) which pilot (captain or FO) the pairs were addressing, (b) which word picture was

TABLE 1: Excerpt of the Assessment Form, Including the Word Pictures for SA and DM

	Scores				
	1	2	3	4	5
SA • Perception • Comprehension • Projection	• Lacked awareness of clearly obvious systems or environmental factors. • Misinterpreted or did not comprehend factors affecting flight safety. • Did not predict future events, even those obvious to flight safety.	• Missed some minor systems or environmental factors not critical to flight safety. • Comprehended some factors and implications on flight safety. • Difficulty predicting future events.	• Perceived significant systems or environmental factors affecting flight. • Comprehended significant factors and implication on flight safety with few errors. • Some difficulty predicting future events.	• Perceived all systems or environmental factors affecting flight. • Comprehended the implication of all factors. • Predicted future events and impact on flight safety.	• Perceived all systems or environmental factors, with an active approach to seeking further information. • Clearly comprehended the meaning of all factors. • Actively considered future events and impact on flight safety.
DM • Time • Facts and Diagnosis • Option generation • Risk assessment • Plan + contingency	• Poor grasp of time available. • Facts not considered, leading to ill-informed or wrong identification or diagnosis. • Inadequate range of options considered. • Inappropriate risk assessment. • Unable to develop plan.	• Limited time appreciation, led to rushed or delayed decision. • Correct diagnosis, though some facts not considered creating initial difficulty. • Limited range of options. • Some risks assessed. • Difficulty developing a plan. • Limited contingency planning.	• Decisions made in time available. • Most facts taken into account and problem correctly identified or diagnosed. • Obvious options considered. • Most risks taken into account. • Developed a plan. • Adequate contingency planning.	• Decision made within clearly established time constraints. • All facts taken into account; problem correctly identified or diagnosed. • All suitable options considered. • Sound risk assessment. • Developed a sound plan. • Sound contingency planning.	• Consistently identified problem, choosing best possible option in time available, appearing almost intuitive. • Clearly outlined the plan, with no doubt to intentions from any crew member. • Thorough risk assessment. • Detailed contingency planning.

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

currently spoken about or pointed to by each of the flight examiners, and (c) which categories (INI, SA, DM, AC, KN, MN, CM) the pairs were talking about.

Data Analysis

In order to determine whether the categories are specific, the key statements made by the pairs had to be extracted. Key statements are the arguments flight examiners put forward, which subsequently are referred to as "justifications". (The term *justification/s* is used to indicate that assessors justify their reasoning after the viewing of each scenario, instead of applying some predefined criteria in the assessment.) The nature of the justifications was not limited to observable behavior. It was entirely left to the flight examiners to decide what aspect of performance they wanted to talk about. Similar justifications were summarized under a "topic," which we defined as the thematic summary of multiple justifications. The summary of justifications into topics was made necessary by the large number of justifications, which often related to very similar issues. For example, the justifications "The cabin call was a full blown conversation at a very late stage, when the crew has got 10–15 seconds until touchdown" (stated by Flight Examiner Pair 1, subsequently referred to as FE1) and "The first officer gave a long-winded briefing under 500 ft" (FE2) were summarized under the topic "Cabin call negative: position, altitude, length" (Table 2). (Subsequently, we use the abbreviations "FE1," "FE2," and "FE3" when we refer to the Flight Examiner Pairs 1, 2, and 3 who participated in the present study.) The methodology best suitable to extract the justifications from the flight examiners' discourse and arrive at the topics is Grounded Theory (Glaser & Strauss, 1965, 1967).

When doing Grounded Theory, the researcher aims at "building theory from data" by using "techniques and procedures for gathering and analyzing data" (Corbin & Strauss, 2008, p. 1). From the data analysis, inductive theories are built (Charmaz, 2008; Corbin & Strauss, 2008; Glaser & Strauss, 1965, 1967). Concepts are developed that conceptualize and synthesize data to explain what the data indicate. By doing so, properties are discovered and relationships identified.

The methodology was implemented in the following way. We began by watching each assessment video, looking for the justifications each pair stated. As shown in the following example, in which FE1 questioned the content and timing of the FO's cabin call, we used curly brackets to highlight and summarize the justifications in the transcripts. To identify justifications at any time, each justification was given a number (here 78):

> FE1–Assessor 1: {It's a full blown conversation with the cabin at a very late stage.
>
> FE1–Assessor 2: Very late stage.}^{78}

In total, 360 justifications were extracted and listed in an Excel spreadsheet. The columns of the list included (a) the justifications, (b) the numbers of the justifications, (c) the pair that stated the justifications (FE1, FE2, or FE3), (d) during the assessment of which category (INI, SA, DM, AC, KN, MN, CM) the justifications were stated, and (e) which pilot (captain, FO, or both) the pairs addressed. Subsequently, justifications were grouped together that thematically addressed the same issue. Each group of justifications received a name, which is referred to as the topics (e.g., "High workload and pressure" [Table 2]). Eventually, the 360 justifications were condensed into 70 topics (this number was not previously defined). Two researchers independently coded the transcripts and grouped the justifications into topics. Any disagreements were discussed until agreement was achieved. Table 2 shows the topics the pairs stated during the assessment.

The list of topics (Table 2) was used to investigate the specificity of the assessment categories. We analyzed which pairs addressed the topics and in which categories. Each category was compared with each other in terms of the topics (e.g., INI-SA, INI-DM, INI-AC, INI-KN, INI-MN, INI-CM, SA-DM, SA-AC, etc.). The results are depicted in the form of a hexagon (Figure 2).

The authors are aware of the risk involved in using topics in the analysis. Summarizing flight examiners' statements (justifications) into topics is a researcher-depended interpretation of the

TABLE 2: The Topics Stated by the Pairs in the Assessment of the Categories

- The engine fire (FE1: INI, SA, DM, KN, MN, CM; FE2: INI, DM, MN; FE3: INI, KN, CM)
- The fire handle issue (FE1: INI, SA, KN, MN, CM; FE3: INI, KN, CM)
- The FO verbalizing the wrong fire handle (FE1: INI, SA, KN, MN, CM; FE3: INI, KN, CM)
- The severity of the fire handle issue (FE1: INI, SA, KN, MN, CM)
- The FO had his hand at the correct fire handle (FE3: INI, KN)
- The simplicity of diagnosing an engine fire (FE2: DM)
- The illumination of the fire handle as the saving grace (FE1: INI, MN)
- The FO correcting himself in regards to pulling the correct fire handle (FE3: KN)
- How the crew dealt with the fire drills (FE1: DM; FE2: INI, DM, MN)
- How the captain checked the FO in regards to the fire handle (FE1: INI, KN; FE3: INI, CM)
- The issue of the running engine on the ground (FE3: SA, DM, AC, KN, MN)
- Not shutting down both engines (FE3: AC, KN, MN)
- The severity of the running engine (FE3: SA)
- FO falsely reporting task completed (FE3: AC, KN)
- Questioning the engine shut down procedure (FE3: AC, KN, MN)
- The crew not picking up the running engine (FE3: SA, DM, KN)
- Evacuating PAX into a running engine (FE3: SA, DM, KN, MN)
- Prompting each other (FE2: KN, MN, CM)
- Splitting tasks on the ground (FE3: KN, MN)
- The captain's delegation of tasks to the FO (FE2: MN; FE3: INI, MN)
- Issues related to the collaboration as a crew (FE1: DM, MN, CM; FE3: MN)
- Planning: negative. No or incorrect plan (FE2: INI, DM, MN, CM; FE3: DM)
- Planning: positive. Plan available (FE3: DM)
- The procedures were good, accurate, appropriate (FE2: DM, KN; FE3: INI)
- The procedures were poorly executed (FE3: DM, KN)
- Issue of disengaging the autopilot (FE2: INI, DM, AC; FE3: INI, AC, CM)
- The cabin call in general (FE1: INI, KN; FE2: KN; FE3: INI)
- The cabin call was appropriate (FE3: INI)
- Cabin call negative: position, altitude, length (FE1: INI, KN; FE2: KN; FE3: INI)
- Questioning whether the crew became VMC by pure chance (FE2: DM, MN)
- Chance of a go around due to not becoming visual (FE2: DM, MN)
- The risk of the fire and a disaster (FE2: DM)
- There were no threats (FE1: MN)
- Wrong display up on the PF/captain's side (FE1: INI)
- The approach profile—in general (FE1: INI, AC; FE2: AC, MN; FE3: INI)
- The approach profile was on track (FE2: AC, MN; FE3: INI)
- The approach profile was off track (FE1: INI, AC; FE2: AC, MN)
- Aircraft flown within the limits/tolerances (FE1: AC; FE2: SA, AC; FE3: INI)
- The landing was off the centerline (FE2: AC; FE3: AC)
- The need to keep flying the aircraft safely (FE2: DM)
- Good decision making (FE2: DM; FE3: DM)
- Poor decision making (FE2: DM)
- The crew made a decision to send PAX out without having shut down both engines (FE3: DM)
- The crew made the decision to continue with the approach and land (FE1: DM; FE2: SA, DM, MN, CM; FE3: INI, DM)
- The crew didn't make a decision to continue or discontinue the approach and land (FE2: INI, DM, MN, CM)
- Good communication—not specified which crew member addressed (FE1: SA; FE2: CM; FE3: CM)
- Good communication of the crew (FE2: CM; FE3: CM)
- Good communication of the FO (FE1: SA; FE2: CM; FE3: CM)
- Poor communication—not specified which crew member addressed (FE1: INI, KN, MN; FE2: INI, SA, DM, MN, CM; FE3: INI, DM, CM)
- Poor communication of the FO (FE2: SA)
- Poor communication of the captain (FE1: INI; FE2: DM, CM)
- Poor communication of the crew (FE1: INI, KN, MN; FE2: INI, SA, DM, MN, CM; FE3: INI, DM, CM)
- Good knowledge—not specified which crew member addressed (FE1: SA, KN; FE2: KN, MN; FE3: KN)
- Good knowledge of the crew (FE2: KN, MN; FE3: KN)
- Good knowledge of the FO (FE1: SA, KN)
- The pilot's knowledge of what is going to happen (FE1: SA, DM; FE3: INI, DM, CM)
- Questioning whether crew considered facts (FE2: DM)
- Captain's SA was good (FE1: SA)
- Captain's SA was poor (FE2: CM)
- Crew's SA was good (FE2: SA; FE3: SA)
- FO's SA diminished due to not knowing what the captain's plan was (FE2: SA)
- Good management of the crew (FE1: MN; FE2: MN; FE3: MN)
- Poor management of the crew (FE2: MN; FE3: MN)
- The crew handled the workload well (FE3: KN)
- High workload and pressure (FE1: KN; FE2: DM, MN, CM; FE3: INI, SA, KN)
- Overload of the captain (FE2: CM)
- Good initial performance (FE1: INI; FE2: SA, DM; FE3: INI, DM, AC, MN)
- The crew is competent and dealt well with a lot of things (FE1: INI, MN; FE2: DM; FE3: INI, SA, DM, KN)
- Good overall flight and outcome (FE1: INI; FE2: SA, DM; FE3: DM)
- Questioning the FO's performance in the future (FE1: MN)

Note. INI = Initial; SA = Situation Awareness; DM = Decision Making; AC = Aircraft flown within tolerances; KN = Knowledge; MN = Management; CM = Communication; FE = Flight examiner (pair); FO = First Officer; PAX = Passengers; VMC = Visual Meteorological Conditions; PF = Pilot Flying.

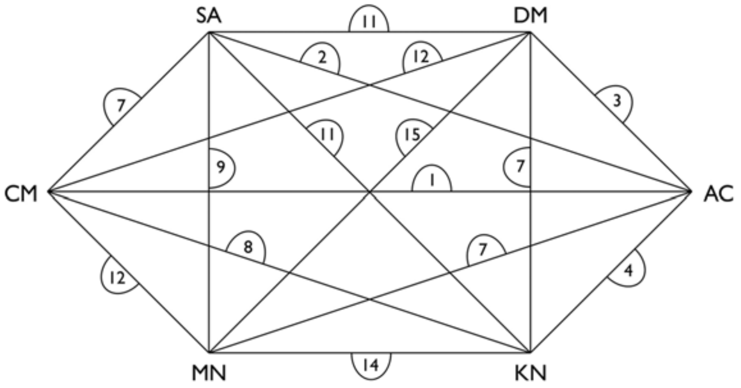

Figure 2. Overlap of the topics that constitute certain categories.

data that evokes the question of the level of abstraction. For instance, we summarized the justifications that the "FO had to prompt the captain with various facts" and that the "Captain prompted the FO to do the PA (public announcement) call" in the topic "Prompting each other" (Table 2). Yet researchers might want to further summarize similar topics—for example, "Prompting each other" and "Splitting tasks on the ground" in a higher topic like "Teamwork." However, by doing so the researchers move further away from the assessors' statements and increasingly introduce their own interpretations that do not necessarily reflect assessors' thinking. The more abstract the topics are chosen (high level of abstraction), the more interpretations the researchers make. To minimize the researchers' influence, we aimed at keeping the level of abstraction low and the topics as close as possible to what the pairs stated. Depending on the statements made by the pairs, Table 2 thus involves a mix of specific (e.g., FO's SA diminished due to not knowing what the captain's plan was) as well as more abstract topics (e.g., crew's SA was good).

The analysis does not include how often a pair addressed each topic because these numbers would be distorted. In their dialogue, flight examiners sometimes stated a topic repeatedly, sometimes they returned to a topic at a later stage of the assessment, and sometimes only one flight examiner spoke about a topic and kept repeating it—for example, to convince his partner. Hence, a high number of repetitions would neither reflect why a topic was reiterated nor say anything meaningful about its importance.

Based on the findings, we extended the analysis and examined the time spent by the pairs in regards to watching the scenario and assessing each category, which subsequently will be referred to as the assessment process (Table 7). [The "assessment process" is defined as the time from when a pair started to watch the video scenario until they (verbally or nonverbally [e.g., putting down the pens, leaning back on the chairs, and starting to talk about something not related to the scenario]) declared the assessment to be completed (without any influence from the researchers).] The assessment process may provide further insight into the differences found between the pairs in terms of the scores. In order to illustrate how each pair conducted the assessment of the scenario, a distinction was made between different phases (Table 7). A phase (P) was defined as a period of time in the assessment process in which the pairs either (1) watched the video scenario or (2) discussed and assessed it in regards to the categories (INI, SA, DM, AC, KN, MN, CM). The assessment process began when the pairs started to watch the scenario (Phase 1; P1). The decision of where to start their discussion and how to assess the scenario was entirely left to the pairs. We defined the transition from one phase to another to occur when a pair (a) went back within the scenario to watch parts of it again (e.g., FE1, P1 to P2), (b) stopped watching the scenario and started to discuss and assess the categories (e.g., FE1, P2 to

TABLE 3: The Scores the Flight Examiner Pairs Provided to the Captain and First Officer Performing in the Scenario

	FE1		FE2		FE3	
	Captain	FO	Captain	FO	Captain	FO
SA	4	4	3	3	1	1
DM	4	4	2	2	2	2
AC	4	4	3	3	2	2
KN	3	4	4	4	2	1
MN	4	4	3	3	1	1
CM	3	4	3	3	4	3
Pass/fail	pass	pass	pass	pass	fail	fail

Note. SA = Situational Awareness; DM = Decision Making; AC = Aircraft maintained within tolerances; KN = Knowledge; MN = Management; CM = Communication; FE = Flight Examiner (pair); FO = First Officer.

P3), (c) stopped their discussion and assessment and went back to rewatch the scenario (e.g., FE3, P9 to P10), or (d) started to discuss and assess a different category (e.g., FE1, P12 to P13). The results related to the assessment process are shown in the subsection "Other Relevant and Interesting Findings."

RESULTS

In the following, we present our findings under four subchapters. First, an overview is given of the scores provided by the pairs. Second, we investigate the specificity of the assessment categories (first aim). Third, we look into the topics addressed in single versus multiple categories (second aim). Finally, we provide further relevant and interesting findings that pertain to (a) the pass/fail assessment in regards to the issue of the spinning engine, (b) the assessment process, (c) how pairs focus on different aspects of performance when they "miss" evidence, (d) disagreement within and between the pairs, and (e) the flight examiners' expression of uncertainty during the assessment.

Scores Provided by the Pairs

The pairs were tasked to mark each pilot in regards to the six categories (SA, DM, AC, KN, MN, CM) with a score from 1 to 5. Within the participating airline, a "pass" reflected the fact that the pilot maintained the aircraft in a safe state. One score of 1 on any of the six categories means a "fail," making a repetition of the flight exam unavoidable. The same is the case when given three 2s. The scores provided are shown in Table 3.

Specificity of Assessment Categories

The first aim of the present effort was to examine whether the categories used to assess the performance of airline professionals are specific, in the sense that flight examiners build their assessment on the same observations and justifications. One could assume that the topics used to assess a certain category are distinctive for each category, in the sense that all pairs articulate the same topics when they, for example, assess SA. However, as outlined in the following, our findings indicate that flight examiners largely apply a different set of topics when they assess a certain category.

The analysis of the topics in regards to the categories in which they were stated (see "Data Analysis") is shown in Figure 2. The lines and numbers between the categories indicate how often topics that were stated by all pairs in regards to a certain category also showed up in the discussion of other categories. Subsequently, the occurrence of a topic in multiple (two or more) categories is referred to as *overlap* of the categories in terms of the topics. For instance, 11 of the topics stated in SA also came up when the pairs assessed DM, 9 in MN, 7 in CM, and so forth. Flight examiners thus often use a very similar set of topics when they assess different categories.

TABLE 4: Ratio of the Number of Overlaps of Each Category in Terms of the Topics to the Number of Times Each Category Was Addressed by All of the Pairs

	Overlaps[1]	Total[2] (100%)	Number Of Times Category Overlapped At Least Once	Number of Times Category Did Not Overlap With Any Other	Ratio Overlaps[a] to Total[b]
SA	40	26	18 (69%)	8 (31%)	1.54 (40/26)
DM	48	42	27 (64%)	15 (36%)	1.14
AC	17	16	11 (69%)	5 (31%)	1.06
KN	44	36	25 (69%)	11 (31%)	1.22
MN	57	41	29 (71%)	12 (29%)	1.39
CM	40	29	20 (69%)	9 (31%)	1.38

Note. SA = Situational Awareness; DM = Decision Making; AC = Aircraft maintained within tolerances; KN = Knowledge; MN = Management; CM = Communication.

[a] Number of times the same topic was addressed in multiple (two or more) categories by all pairs. For example, 40 times, the pairs noted the same topics that they used to assess SA in other categories (multiple overlaps per topic included).

[b] Total number of times topics were addressed by all pairs in regards to a certain category. For example, the number of times all of the pairs referred topics to SA was 26.

All categories overlap, yet some do more than others (Figure 2). An example is the topic “poor communication of the crew” (Table 2). One could assume that this topic was explicitly related to the category CM, yet it appears to be crucial in the assessment of other categories. FE1, for instance, spoke about the poor communication of the crew in their INI, KN, and MN discussion; FE2 in INI, SA, DM, MN, and CM; and FE3 in INI, DM, and CM.

Figure 2 further indicates the total number of times the topics stated in the assessment of one category reappeared in the discussion of all the other categories. For example, the topics stated when assessing SA also emerged 40 times in the discussion of other categories. This number is the sum of all connections between SA and the other categories (Figure 2 and Table 4, “Overlaps”). All pairs together related topics to the SA category for a total of 26 times (column “total”).

Table 4 shows the number of times a topic at least once showed up in multiple categories (e.g., 18 for SA), as well as the number of times the topics were only used to assess one category (i.e., 8 for SA). The ratio between “overlaps” to “total” highlights that SA had the most overlaps (1.54) in regards to the number of times the category was addressed (26 times) by all pairs. These rates are greater than 1.0 because the number of times a topic showed up in multiple categories (column “Overlaps”) was always higher than the number of times a topic was addressed by all pairs in regard to a certain category (column “Total”). The reason was that the categories could overlap multiple times in terms of the topics.

A closer examination of the number of topics addressed by the pairs is provided in Table 5. FE1 spoke about 10 topics in regards to SA (column “SA”, row “Number of topics stated by FE1”), FE2 about 9, and FE3 about 7. Out of 25 different topics (100%) stated by all pairs when assessing SA, FE1 addressed 40%, FE2 36%, and FE3 28% of the topics. Furthermore, the pairs spoke about 25 *different* SA topics (in contrast to Table 4, one SA topic was stated by two pairs, thus 25). However, 24 out of 25 different SA topics were only addressed by a *single* flight examiner pair (Table 5, column “SA,” row “Topics addressed by 1 pair”). One topic was shared between two pairs (row “Topics addressed by 2 pairs”), whereas none was stated by all of the pairs (row “Topics addressed by all 3 pairs”). Hence, the pairs largely applied an entirely different set of topics when they assessed each category, particularly in regards to SA.

TABLE 5: The Number of Topics Addressed by the Pairs

	Initial	SA	DM	AC	KN	MN	CM
Total number of *different* topics the pairs stated in each category (100%)	30	25	30	11	27	33	21
Topics addressed by 1 pair	18	24	19	6	19	26	13
Topics addressed by 2 pairs	9	1	10	5	7	6	8
Topics addressed by all 3 pairs	3	0	1	0	1	1	0
Percentage of topics addressed by 2 or 3 pairs in regards to the total number of different topics	40% [(9+3*100)/30]	4%	37%	46%	30%	21%	38%
Number of topics stated by FE1[a]	17 (57%)	10 (40%)	5 (17%)	3 (27%)	12 (44%)	12 (36%)	5 (24%)
Number of topics stated by FE2[a]	7 (23%)	9 (36%)	22 (73%)	6 (55%)	6 (22%)	19 (58%)	13 (62%)
Number of topics stated by FE3[a]	21 (70%)	7 (28%)	15 (50%)	7 (64%)	18 (67%)	10 (30%)	11 (52%)

Note. SA = Situational Awareness; DM = Decision Making; AC = Aircraft maintained within tolerances; KN = Knowledge; MN = Management; CM = Communication; FE = Flight examiner (pair).

[a] Percentages are in regards to the total number of *different* topics the pairs stated in each category (yet the numbers themselves relate to column "Total" in Table 4, because a topic could be addressed by more than one pair).

Topics Addressed in Single Versus Multiple Categories

The second aim of the present study was to investigate what might distinguish the sort of topics that are specific to a single assessment category from the topics that are stated in multiple categories. This required an analysis of the 70 topics in regards to the seven categories and pairs, which made us distinguish between three cases (note that many topics were not addressed by all pairs; see Table 2): In the first case, all pairs stated a specific topic in only a *single* category (either in INI, SA, DM, AC, KN, MN, or CM), which occurred in relation to 27 out of the 70 topics. Examples are the topics "There were no threats" or the "Overload of the captain" (see Table 2 for the pairs who stated these topics and in which categories). Two of the 27 topics were only stated in the INI category.

In the second case, all pairs spoke about a certain topic in *multiple* categories. FE1, for instance, spoke about "The fire handle issue" in regards to the categories INI, SA, KN, MN, and CM; FE3 in INI, KN, and CM. This case was found in 25 out of 70 topics.

In the third case, a *mix* occurred between the first and second cases: some of the pairs addressed a topic in only a single category, whereas others did in multiple categories. Examples are the "Good overall flight and outcome" or "The captain's delegation of tasks to the FO." This case occurred in 18 out of 70 topics.

We extensively compared the topics related to the first and second cases. Yet the analysis revealed no clear pattern of what might distinguish the topics that are specific to one versus two or more categories. Some of the topics that the pairs only addressed in regards to a single category (first case) were similar to some of the topics that came up in multiple categories (second case). However, no clear distinction could be drawn between topics that relate to a single versus multiple categories.

Other Relevant and Interesting Findings

As shown in Table 3, FE1 and FE2 passed the crew, whereas FE3 failed both the captain and FO. It was also only FE3 who made the critical observation that the right-hand engine was still spinning when the crew evacuated the passengers on the runway to this side (Table 2, "The issue of the running engine on the ground"). (The issue of the spinning engine was not scripted prior to the production of the video-scenario. It was only noticed after its completion. The notion of unintentionally evacuating passengers into a spinning engine increased the scenario's originality in the sense that this could happen in a similarly hectic real emergency situation. This in turn was the reason to use the scenario to study pilot performance assessment.) In the discussion after the assessment, FE3 noted that the issue of the spinning engine had strongly influenced their scores provided.

Given the critical observation of the spinning engine, one might expect FE3 to share a lower number of topics with FE1 and FE2 than these latter pairs share among each other. Yet the comparison of the topics stated by the pairs (Table 6) shows that FE1 and FE2 share 27% of the 70 topics, which is lower than the topics shared between FE1 and FE3 (30%), and FE2 and FE3 (39%). Consequently, even the pairs who did not notice the spinning engine built their assessment on different topics (73%). To make the spinning engine responsible for all the variance within the scores (Table 3) appears too simple of an explanation.

The observation of the spinning engine influenced the way FE3 assessed the scenario, which is reflected in the process of how the pair assessed the crew's performance (Table 7). FE3 watched the scenario in its entirety (Phase P1) for 9 minutes, 16 seconds. Then they went straight back (P2) and watched the video again for 01:36. Subsequently, the pair had a conversation for 28 seconds (P3), then another look at the video (P4), and so forth. FE3 noticed the spinning engine when they watched the scenario again (P10, at 19:43). They paused using the assessment model and rewatched the scenario multiple times (P11-P14). Compared to the other pairs, FE3 rewatched the video more often (8 times) and spent more time watching the scenario (16:20). Furthermore, the analysis of the phases indicates that the pairs differ in the amount of time they spent to assess the scenario (Table 7, row "Total assessment time"), to watch the assessment video, and to assess each category. There is also a difference

TABLE 6: Comparison of the Number of Topics Stated by the Pairs

	FE1	FE2	FE3	FE1 and FE2	FE1 and FE3	FE2 and FE3	FE1, FE2, and FE3
Number of topics stated/shared	31 (44%)	43 (61%)	47 (67%)	19 (27%)	21 (30%)	27 (39%)	16 (23%)
Number of topics not stated/shared	39 (56%)	27 (39%)	23 (33%)	51 (73%)	49 (70%)	43 (61%)	54 (77%)
Total (100%)	70	70	70	70	70	70	70

Note. FE = Flight examiner (pair).

in how often each pair went back and watched parts of the scenario again (number of phases in the row "Time of watching the scenario").

FE1 and FE2 "missed" evidence in regards to the issue of the spinning engine. Instead, they focused more intensively on other aspects of performance. FE1 largely addressed the issue of the FO verbalizing the wrong fire handle during the engine shut down procedure (in INI, SA, KN, MN, and CM). The fire handle issue was also noted by FE3 but not discussed as extensively as by FE1. FE2 neither spoke about the fire handle nor about the spinning engine (Table 2). Instead, they largely found fault in the fact that the crew didn't make a decision to continue or discontinue the approach and land when the fire occurred (in INI, DM, MN, and CM). When pairs "miss" evidence, they more heavily focused on other aspects of performance.

The analysis of the topics revealed that there was disagreement between and within the pairs. Examples are whether the captain's SA was good versus poor (FE1 vs. FE2) or if the approach profile was on versus off track (FE1 vs. FE3). That the crew did not make a decision to continue or discontinue the approach and land was of major concern to FE2 and "the crux of [the scenario]." Yet the FE2 assessors disagreed about whether the captain made a decision without verbalizing his intentions, or if he did not make a decision on how to continue with the approach. "It tended to be the right decision, but as to actually making the decision to do that, maybe he did, maybe he didn't, but he didn't communicate it to anyone" (FE2–Assessor 2).

In the course of the assessment, the pairs expressed uncertainty and the need for more information about what the ratee said in the debriefing. Uncertainty is also reflected in the flight examiners' language, such as a statement made by FE2 (emphasis added):

> FE2–Assessor 1: *Maybe* [the captain] was in overload. *Maybe* his SA was screwed, down a wee bit because of the fire, and he was thinking a million things so he needs the other guy. So *maybe* from the [FO]'s point of view, *maybe* the communication was a bit better.

Furthermore, uncertainty was often expressed in regards to the pilots' SA. FE3 noted that because the crew did not speak much, it was hard to assess if the crew was considering the future state of the flight or if they were just acting on the here and now (projection of the future is deemed a vital part of SA [Endsley, 1995]; e.g., projecting the future state of the aircraft). FE2's disagreement about the captain's decision to continue on the approach remained unsolved. They ended their discussion, and expressed their difficulties with the assessment of SA, by saying: "You don't know what was going on in his head" (FE2–Assessor 1).

DISCUSSION

This study was designed to examine whether flight examiners build their assessment on the same observations and justifications and to investigate what might distinguish the sort of observations that are specific to a single assessment category from observations that are stated in multiple categories. The results show that the categories

TABLE 7: The Assessment Process: Phases of Watching and Assessing the Video Scenario by Each Flight Examiner Pair

	FE1[b]	FE2	FE3
Time of watching the scenario	Phase P1 (07:18), P2 (00:37), P4 (00:12), P6 (03:20); total time: 11.27	Phase P1 (09:16), P2 (01:45), P4 (02:33); total time: 13.34	Phase P1 (09:16), P2 (01:36), P4 (01:45), P6 (00:47), P10 (00:23), P12 (01:10), P14 (00:27), P16 (00:56); total time: 16:20
Time initial[a]	P3 (00:17), P5 (01:08), P7 (05:15); total time: 06:40	P3 (00:40 = total time)	P3 (00:28), P5 (01:35), P7 (01:48), P11 (00:18), P13 (00:50); total time: 04:59
Time to assess SA	P13 (00:53, Captain), P17 (00:43, FO); total time: 01:36	P10 (07:45 = total time)	P20 (02:30 = total time)
Time to assess DM	P12 (01:16, Captain), P19 (01:04, FO); total time: 02:20	P9 (11:59 = total time)	P21 (04:05 = total time)
Time to assess AC	P11 (01:04, Captain), P18 (01:02, FO); total time: 02:06	P8 (02:29 = total time)	P22 (07:51 = total time)
Time to assess KN	P10 (01:40, Captain), P16 (00:41, FO); total time: 02:21	P7 (01:55 = total time)	P8 (01:28), P15 (01:54), P17 (01:13); total time: 04:35
Time to assess MN	P9 (03:43, Captain), P15 (00:58, FO); total time: 04:41	P6 (04:51 = total time)	P9 (00:45), P18 (02:07); total time: 02:52
Time to assess CM	P8 (02:17, Captain), P14 (01:17, FO); total time: 03:34	P5 (04:37 = total time)	P19 (01:15 = total time)
Total assessment time	34:45	47:50	44:27

Note. The phases (P) in the assessment process are numbered for each individual flight examiner pair (P1, P2, etc.). The duration of each phase is the time shown in brackets (minutes: seconds). SA = Situational Awareness; DM = Decision Making; AC = Aircraft maintained within tolerances; KN = Knowledge; MN = Management; CM = Communication; FE = Flight Examiner (pair); FO = First Officer.

[a] This category involves the conversation of the pairs before using the assessment model. It also includes the time when pairs made a crucial observation that resulted in a period in which they did not use the assessment model (e.g., FE3, P11 and P13).

[b] FE1 finished the assessment of the captain before they assessed the FO's performance. In contrast, FE2 and FE3 assessed both pilots simultaneously, which did not allow a distinction between the assessment of the captain and FO.

are hardly specific to any single category in terms of the topics that the pairs stated. Rather, the topics that constitute certain categories overlap in the sense that the same topics are used to assess various categories. It remains unclear what distinguishes the sort of topics that are specific to a single versus two or more categories. The problem seems to be related to the categories rather than their operationalization, because flight examiners were given word pictures to assess each category.

SA appears to be the most problematic category. The topics used to assess SA often show up in the assessment of the other categories. At the same time, pairs noted entirely different topics to assess SA: only one topic was addressed by two pairs, whereas not a single topic was shared among all of the three pairs. Despite the fact that word pictures were provided that break SA (and all of the other categories) down into more specific units for each of the scores (1 to 5), flight examiners build their scores on entirely different reasons. Using an indirect measurement of SA (e.g., Chauvin, Clostermann, & Hoc, 2008; Gonzalez & Wimisberg, 2007; Gugerty, 1997), flight examiners do not agree about the topics related to this category. No clear analytical trace seems to exist between the topics stated and the category assessed (e.g., Dekker, 2006).

Such disagreement evokes concerns about the reliability of assessment scores and the validity of the categories. [Validity: "the extent to which a measure really assesses the construct" (Flin, O'Connor, & Crichton, 2008, p. 277).] Current research indicates that variance in pilot performance assessment can be explained and that the assessors' actual scores can be predicted by using a fuzzy logic approach (Roth & Mavin, 2013). Undoubtedly, assessors may have good reasons for their scores and there might be a pattern on how they arrive at a decision about a pilot's performance. However, reliability and validity need to be questioned because assessments are based on entirely different observations made and topics addressed by the assessor pairs (e.g., Weber et al., 2013).

It goes without saying that we can never expect all assessors to use certain topics only in regards to assessing any particular category. Yet if validity were high, we would expect assessors to use a very similar set of topics to assess each category (i.e., Flin et al., 2008). The overlap and disagreement found in the present study requires questioning whether the categories used and the current distinction between the categories is of help to achieve agreement between assessors.

There currently are over 30 techniques to measure SA (Salmon, Stanton, Walker, & Green, 2006), such as freeze probe, self-rating, or observer rating techniques. The latter was used in both the present study and within the NOTECHS framework. Based on our findings, we share concerns expressed in the literature (e.g., Salmon et al., 2006) about whether observers are able to accurately assess ratees' SA. The observer rating technique seems problematic because assessors are expected to make sense of what is going on in practitioners' mind.

The disagreement in regards to SA is not surprising when reflecting on both the concerns discussed in the literature (e.g., Dekker & Hollnagel, 2004; Dekker, Hummerdal, & Smith, 2010; Dekker, Nyce, van Winsen, & Henriqson, 2010; Hollnagel & Woods, 2005; Salmon & Stanton, 2013; Salmon et al., 2008) and the uncertainty stated by the pairs. Flight examiners ask for more information and express the need to question the pilots in the debriefing. At several stages during the assessment, the flight examiners have to make assumptions of the pilots' thought processes (e.g., whether the captain made a decision in regards to dis/continue the approach). To assess the categories, assessors are required to speculate about the processes that go on in pilots' heads. Based on our findings, we thus question whether disagreement within and between the pairs is the result of the categories themselves and the uncertainty assessors face when trying to assess these categories. This conclusion is given weight by the fact that all flight examiners received extensive training on the use of the categories and assessed various other scenarios before participating in the present study. Furthermore, it can be ruled out that the categories wholly fail to capture what is important to the flight examiners, because when asked after the assessment, they all stated that the assessment model captured their arguments and did not force them to any conclusion that made them feel uncomfortable.

Large overlap was found in terms of the topics that constitute SA and KN (Figure 2). An explanation may lie in the very notion of SA itself. Numerous definitions of SA have been proposed (Endsley, 1995), which have in common to all point to "*knowing* what is going on" (p. 36, emphasis added). Endsley (1995) defined SA as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (p. 36). Consequently, pilots are said to have SA if they have *knowledge* of the elements, *knowledge* of their meaning, and *knowledge* of their status in the future. Thus, SA can be seen as the pilot's *knowledge of the situation*, which in turn makes a separation between SA and KN in the assessment questionable.

The validation of assessment categories that are widely used to assess the performance of airline pilots largely builds on quantitative measures. To validate the NOTECHS framework, for instance, JARTEL had 105 assessors rate the performance of a captain and FO in eight scenarios (O'Connor et al., 2002). Each assessor provided a score for each of the NOTECHS items (element, category, and pass/fail ratings). Based on these scores, and by using reference ratings from two expert groups, internal consistency, accuracy, and interrater agreement were calculated. The results showed high levels of agreement between the assessors.

Similarly to the NOTECHS validation, it is a common trend to assume that identical scores reflect assessors' agreement in terms of their reasoning. Yet the results of the present study indicate that flight examiners build their assessments on entirely different observations and topics. Any validation of assessment categories that is simply based on quantitative measures does not indicate whether assessors agree about the underlying reasoning.

FE3, who failed both crews, was the only pair who made the observation of the spinning engine on the ground when the captain and FO evacuated the passengers. The issue of the running engine led to low assessment scores. However, it remains unclear whether the observation was made due to experience or luck. An observation strongly influenced the assessment of almost all of the categories (FE3 spoke about the spinning engine in all categories except INI and CM; Table 2). Yet the opposite does not seem to apply: the use of categories, and even specific word pictures, is no guarantee for more similar observations made by the pairs.

CONCLUSION AND FUTURE RESEARCH

The findings show that flight examiners use a range of the same and different topics when assessing widely used categories to judge a pilot's performance. This is so even when they already have the opportunity to discuss the assessment with a peer prior to settling on a score. Assessment categories largely overlap in terms of the topics addressed. Assessors face uncertainty and difficulties because several categories require them to speculate about what is going on in the ratee's mind instead of focusing on observable performance. Depending on the evidence flight examiners find, they focus on different aspects of performance to draw conclusions about underlying cognitive characteristics of the pilot. In case flight examiners miss evidence that is observed and deemed critical by other pairs, they weigh other observations more heavily. The possibility for considerable divergence between assessors, suggested by the larger samples of preceding work (Mavin, Roth, & Dekker, 2013), has been confirmed in the study reported here. The findings underline concerns about the validity of some of the assessment categories and indicate that the key to understanding disagreement between assessors may lie in the unspecific nature of the categories themselves.

Greater overall consistency between assessors who worked in pairs raises the question of whether getting expert evaluators to work in even greater groups will generate an even more rigorously defined set of categories and subcategories. This should be explored in subsequent research.

Furthermore, research needs to investigate the validity of assessment categories, involving airline professionals from different airlines and ranks (FOs, captains, flight examiners), who assess multiple video scenarios. For instance, the validity of assessment categories can be examined by having a large number of assessors rate whether they believe that there is high versus low evidence (e.g., using a Likert-type scale from 1 to 5) in an assessment scenario for each individual assessment category (e.g., SA, DM). Including a large number of both participants (airline professionals as well as, for example, aviation psychologists) and video scenarios, generalizations can be drawn to a broader population.

The analysis process revealed differences in the way the pairs assessed the scenario. Further research has to enquire whether there are (significant) differences between the time spent to watch the scenario, assess each category, and the scores provided. It needs to be questioned whether longer assessment times reflect assessors' difficulties to assess the categories, and how this is related to the scores. It is interesting to ask whether the issue of the spinning engine is only noticed by highly experienced assessors and how it influences the scores provided by a larger number of pairs. All of these questions are the subject of current research. At the present stage, the assessment categories that are currently used to assess airline pilots' performance have to be used with care when expected to increase agreement between assessors.

ACKNOWLEDGMENTS

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

REFERENCES

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

Chauvin, C., Clostermann, J. P., & Hoc, J.- M. (2008). Situation awareness and the decision-making process in a dynamic situation: Avoiding collisions at sea. *Journal of Cognitive Engineering and Decision Making*, *2*(1), 1-23.

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

Dekker, S. W. A. (2006). *The field guide to understanding human error*. Aldershot, UK: Ashgate.

Dekker, S. W. A., & Hollnagel, E. (2004). Human factors and folk models. *Cognition, Technology, and Work*, *6*, 79-86.

Dekker, S. W. A., Hummerdal, D. H., & Smith, K. (2010). Situation awareness: Some remaining questions. *Theoretical Issues in Ergonomics Science*, *11*(1-2), 131-135.

Dekker, S. W. A., Nyce, J. N., van Winsen, R., & Henriqson, E. (2010). Epistemological self-confidence in human factors research. *Journal of Cognitive Engineering and Decision Making*, *4*(1), 27-38.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32-64.

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

Flin, R., O'Connor, P., & Crichton, M. (2008). *Safety at the sharp end: A guide to non-technical skills*. Aldershot, UK: Ashgate Publishing Ltd.

Glaser, B. G., & Strauss, A. L. (1965). *Awareness of dying*. Chicago: Aldine.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.

Gonzalez, C., & Wimisberg, J. (2007). Situation awareness in dynamic decision making: Effects of practice and working memory. *Journal of Cognitive Engineering and Decision Making*, *1*(1), 56-74.

Gugerty, L. J. (1997). Situation awareness during driving: Explicit and implicit knowledge in dynamic spatial memory. *Journal of Experimental Psychology: Applied*, *3*(1), 42-66.

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

Hollnagel, E., & Woods, D. D. (2005). *Joint cognitive systems: Foundations of cognitive systems engineering*. Boca Raton, FL: Taylor & Francis.

JARTEL. (2002). *Final Report, JAR TEL, Consolidation of Results, WP7 Draft Report* (JAR TEL/WP7/D9_07).

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, Sydney, Australia.

Mavin, T. J., Roth, W.- M., & Dekker, S. W. A. (2013). Understanding variance in pilot performance ratings: Two studies of flight examiners, captains and first officers assessing the performance of peers. *Aviation Psychology and Applied Human Factors*, *3*(2), 53-62.

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

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

Roth, W.- M., & Mavin, T. J. (2013). Assessment of non-technical skills: From measurement to categorization modeled by fuzzy logic. *Aviation Psychology and Applied Human Factors*, *3*(2), 73-82.

Salmon, P. M., & Stanton, N. A. (2013). Situation awareness and safety: Contribution or confusion? Situation awareness and safety editorial. *Safety Science*, *56*, 1-5.

Salmon, P. M., Stanton, N. A., Walker, G. H., Baber, C., Jenkins, D. P., McMaster, R., & Young, M. S. (2008). What really is going on? Review of situation awareness models for individuals and teams. *Theoretical Issues in Ergonomics Science*, *9*(4), 297-323.

Salmon, P., Stanton, N., Walker, G., & Green, D. (2006). Situation awareness measurement: A review of applicability for C4i environments. *Applied Ergonomics*, *37*, 225-238.

Weber, D. E., Roth, W.- M., Mavin, T. J., & Dekker, S. W. A. (2013). Should we pursue inter-rater reliability or diversity? An empirical study of pilot performance assessment. *Aviation in Focus–Journal of Aeronautical Sciences*, *4*(2), 34-58.

Whitehurst, G. (2013). Dwindling resources: The use of single-case research designs as an efficient alternative for applied aviation research. *Aviation Psychology and Applied Human Factors*, *3*(2), 63-72.

David E. Weber (MSc, University of Basel, Switzerland, 2010) is a private pilot and PhD candidate at Griffith University, Brisbane, Australia. With a background in cognitive psychology, human computer interaction, and accident investigation, he focuses on performance assessment and aviation safety.

Timothy J. Mavin (EdD, University of Queensland, Australia, 2010) is an Associate Professor at Griffith University, Brisbane, Australia. He has been flying as a Captain for several major airlines on different aircraft types, including the Boeing 737. As a flight examiner, he is still instructing and assessing pilots.

Wolff-Michael Roth (PhD, University of Southern Mississippi, 1987) is a Lansdowne Professor at the University of Victoria, Canada. He has a broad background in teaching and had been working in research methods, science, mathematics, and technology education. He is a successful author of a large number of publications. One of his most recent books is *"Meaning and Mental Representation: A Pragmatic Approach"* (2013).

Eder Henriqson (PhD, Federal University of Rio Grande do Sul, 2011) is an Adjunct Professor in the School of Aeronautical Science at the Pontifical Catholic University of Rio Grande do Sul, Brazil. He has a commercial pilot license and experience as a flight instructor.

Sidney W. A. Dekker (PhD, Ohio State University, 1996) is a Professor in the School of Humanities at Griffith University, Australia, and Honorary Professor in the School of Psychology at the University of Queensland, Australia. Previously, he was a Professor at Lund University, Sweden, and Director of the Leonardo Da Vinci Center for Complexity and Systems Thinking. He has recently been flying part-time as a pilot on the Boeing 737NG.