Final version published as: Roth, W.-M., & Mavin, T. J. (2015). Peer assessment of aviation performance: Inconsistent for good reasons. *Cognitive Science, 39*, 405–433. DOI: 10.1111/cogs.12152

Peer Assessment of Aviation Performance: Inconsistent for Good Reasons

Wolff-Michael Roth[1,2] & Timothy J. Mavin[2]

[1] Faculty of Education, University of Victoria

[2] Griffith Institute of Educational Research, Griffith University

Corresponding author: Wolff-Michael Roth, Lansdowne Professor, Applied Cognitive Science, Faculty of Education, MacLaurin Building A567, University of Victoria, Victoria, BC, V8P 5C2. Tel: #1 (250) 721-7764; FAX: #1 (250) 721-7598; Email: mroth@uvic.ca

RUNNING HEAD: **Experts Assessing Experts**

Keywords: Assessment; Peer; Performance; Expertise; Fuzzy Logic

Peer Assessment of Aviation Performance: Inconsistent for Good Reasons

Abstract

Research into expertise is relatively common in cognitive science concerning expertise existing across many domains. However, much less research has examined how experts within the same domain assess the performance of their peer experts. We report the results of a modified think-aloud study conducted with eighteen pilots (6 first officers, 6 captains and 6 flight examiners). Pairs of same-ranked pilots were asked to rate the performance of a captain flying in a critical pre-recorded simulator scenario. Findings reveal (a) considerable variance within performance categories, (b) differences in the process used as evidence in support of a performance rating, (c) different numbers and types of facts (cues) identified, and (d) differences in how specific performance events affect choice of performance category and gravity of performance assessment. Such variance is consistent with low inter-rater reliability. Because raters exhibited good, albeit imprecise, reasons and facts, a fuzzy mathematical model of performance rating was developed. The model provides good agreement with observed variations.

1. Introduction

There now exists a long history of studies of expertise in very different domains focusing on such issues as problem solving (e.g., Arts, Gijselaers, & Segers, 2006), decision making (e.g., Connors, Burns, & Campitelli, 2011), categorization (e.g., Chi, Feltovich, & Glaser, 1981), and complex systems in real worlds and microworlds (e.g.,

Güss, Tuason, & Gerhard, 2010; Hmelo-Silver & Pfeffer, 2004). Many studies focus on differences between experts and novices (e.g., Rottman, Gentner, & Goldwater, 2012). A much smaller number of studies focus on expert/expert differences in, for example, chess players at the international masters versus class A levels (Connors et al., 2011), historians' categorization of historical texts (Wineburg, 1998), English professors' interpretations of academic arguments about poems (e.g., Warren, 2011), designing experiments (e.g., Schraagen, 1993), or research biologists' interpretation graphs (e.g., Roth & Bowen, 2003). In such studies, the research investigates, assesses, and evaluates the real-time performance of experts within or out of their field. Much less frequent, if these exist at all, are studies (a) of the cognitive processes involved in assessing the performance of peer experts and the form and quality of reasoning underlying performance assessment and (b) that investigate cognitive processes and reasoning in the workplace. The present study was designed to understand how pairs of same-ranked pilots with different levels of (a) piloting experience (captains vs. first officers) and (b) assessment experience (flight examiners vs. captains and first officers) assess the performance of peers at work. Specifically, this study was designed to answer four research questions: (a) What is the degree of variability in peer performance assessment within and between levels of assessment and job experience? (b) Do peers with different levels of assessment and job experience engage the same performance assessment process? (c) How do the number and type of perceptual distinctions made (facts isolated) determine performance assessment? and (d) How are specific performance events mapped onto performance categories and levels of performance ratings?

2. Study background: expertise in performance assessment

This study investigates performance assessment experts' assessment of their peers in a context where peer assessment (evaluation) is part of the regular work. A review of

studies on judgment expertise reviewed nine traditional approaches used to distinguish levels of judgment expertise: within reliability (internal consistency), inter-rater reliability (consensus), experience, certification, social acclamation, discrimination ability, behavioral characteristics, knowledge tests, and rater training (Shanteau, Weiss, Thomas, & Pounds, 2002). The review concluded that each of these approaches to distinguishing expertise in making judgments has flaws. For example, whereas within-rater reliabilities are high in some domains—e.g., weather forecasters (r = .98), livestock judges (r = .96), or auditors (r = .90)—these tend to be low in other domains—e.g., pathologists (r = .50), clinical psychologists (r = .44), and stockbrokers ($r \leq .40$). The problem is that even if the consistency is high, it may be achieved by following consistent but incorrect rules. Consensus-based reliabilities in the same fields were lower but clustered the same professions with highest levels of performance among weather forecasters (r = .95), auditors (r = .76), or grain inspectors (r = .60) and lowest levels of performance among pathologists (r = .55), clinical psychologists (r = .40), and stockbrokers (r = .32). Previous studies of assessment in aviation vary between reporting good (O'Connor et al. 2002), moderate (Flin et al., 2003) and even low or non-existent inter-rater reliabilities (Mavin, Roth, & Dekker, 2013; Smith, Niemczyk, & McCurry, 2008). There are suggestions, however, that increasing the number of judges of the same performance episodes would increase the reliability (Brannick, Prince, & Salas, 2002; Hung, Chen, & Chen, 2012). In the present study, performance assessment sessions involved pairs of assessors asked to discuss performance levels until they reached agreement; this allows answering the first research question concerning the variability in peer performance assessment conducted by pairs of experts.

In aviation, peer assessment is a regular part of work. Flight examiners are experienced pilots (captains) who conduct the performance evaluations that a pilot undergoes twice per year. In the present study, the non-flight examiner participants had received some training in performance assessment as part of their professional

development. The second research question investigates whether experienced performance assessors differ from less experienced pilots in their method of assessment. Some previous studies suggest that there are no correlations between the amount of experience in making judgments and judgment expertise (Shanteau et al., 2002). However, rater training generally does increase expertise but may take years to do so. Increases in the quality of performance assessment with experience have been reported—e.g., trained raters of writing samples tend to be more reliable than untrained raters (Shohamy, Gordon, & Kramer, 1992); and as the number of ratings conducted increases, assessments tend to align with those of acclaimed expert raters (Lim, 2011). In aviation, some studies report achieving up to 80% consistency by means of rater training (Flin et al., 2002). However, whereas intensive training and calibration periods have been recommended in the aviation context (O'Connor et al., 2002) or nuclear power plant operations (O'Connor, O'Dea, Flin, & Belton, 2008), at least one study shows that even a three-year training program produced only moderate inter-rater reliability (Holt, Hansberger, & Boehm-Davis, 2002). These findings are consistent with the suggestion that the observed variability in performance assessment arises from the complexity of performance in real-life setting; some authors therefore use complexity theory to explain why and how performance in critical safety areas is too complex to be reliably and accurately assessed (Bergström, Henriqson, & Dahlström, 2011). On the other hand, studies of rater performance that varied the task constraints and therefore task complexity did not lead to differences in the quality and accuracy of ratings (Hung et al., 2012). A study of workplace-based assessment in the medical field does conclude that there are differences in the cognitive processes between experienced and non-experienced raters of performance in complex environments (Goevarts, Schuwrith, van der Vleuten, & Muijtjens, 2011).

The third research question concerns the relation between number and type of perceptual distinctions assessors make and performance assessment. Discrimination

ability has been identified as a necessary but insufficient criterion of judgment expertise (Weiss & Shanteau, 2003). For example, to judge whether a patient has or does not have a *myocardial infarction*, the doctor has to identify specific symptoms and their gravity among a range of stimuli that the patient presents. Research shows that assessors tend to differ considerably both in the nature and gravity of the symptoms they observe and in the precision of their categories-in-use, which leads to considerable variations in reasoning and categorization during medical diagnosis (e.g., Esogbue & Elder, 1983) and during assessment in aviation (Rigner & Dekker, 2000). One study in aviation showed that performance raters frequently fail to discriminate and assess particular cues, for example, how a pilot responded to air traffic control's loss of radar contact or to an electrical fire in the aircraft (Brannick et al., 2002). Although discriminating relevant information or processes is integral to being a good judge of performance, great variability in both diagnoses and cue utilization has been reported among expert surgeons (Skåner, Strender, & Bring, 1998) and the cues actually used tend to be imprecise (Esogbue & Elder, 1979, 1980). More experienced raters appear to generate more inferences whereas non-experienced raters generated more literal descriptions of behaviors while assessing workplace-based performance in the medical field (Goevarts et al., 2011).

The fourth research question asks how specific performance events—e.g., a pilot observing and commenting on the bad weather ahead—influence the assessors' choice of performance categories and levels. Previous studies show that performance assessment is mediated by the level of assessment expertise and the particulars of the situation assessed, such as the number and types of categories created, the particular nature of the cues that figured into the generated categories, and the function of these cues during the process of categorization (Stains & Talanquer, 2008). Variance in assessment therefore may be controlled by the surrounding situation within which performance occurs rather than by pilot performance itself (Brannick et al., 2002). To locate the possible sources of these

variations in performance assessment, we draw on a model from assessment in aviation (O'Connor et al., 2002). It suggests that performance assessment moves from elements of a category (e.g., perception, comprehension, and projection), which are collected into performance categories (e.g., situation awareness), which, when combined, result in a pass/fail rating: elements > categories > pass/fail. However, to make an assessment with respect to a particular element, an assessor of pilot performance draws on one or more specific observations (cues) that become the evidence on which assessment is based (e.g., "the pilot perceived the bad weather"). We therefore expand the assessment model: facts (cues) > elements > categories > pass/fail. This model suggests sources of variance of the type reported in the literature: (a) performance assessors identify different cues or differ in the importance they assign to a specific cue (fact) when assessing category elements and categories and (b) the performance categories themselves are fuzzy rather than sharply defined (Rigner & Dekker, 2000; Dekker & Hollnagel, 2004). Variation also may occur because in complex fields, such as aviation or schools, any scenario provides a number of situations that can be mapped onto the same performance category so that assessors use cues from different situations to assess it (Holt et al., 2002; Horng, Klasik, & Loeb, 2010).

3. Methods

This study was designed to investigate how experts in the aviation field (airline pilots) evaluate the performance of peers and the reasons they use. For maximum ecological validity, the research was conducted in the normal workplace setting of the participating pilots, where they had access to the same tools that were part of their crew resource management training and the company's assessment for each pilot on a six-month basis.

3.1. Participants

Eighteen pilots from one airline participated in this study, including three pairs each of flight examiners (FE, who are also captains who continue to serve the company in this function), captains (CAP), and first officers (FO). Participants were selected randomly from those with free slots on the roster during the eight-day period that the researchers were on site—i.e., the pilots were on duty but were not flying. The means for the variables of age, total flight hours, years as commercial pilot, years as training captain, and years as flight examiner together with standard deviations are listed in Table 1.

«««««« Insert Table 1 about here »»»»»»

Participants differed along two dimensions: years as pilots and rank (captains vs. first officers) and experience in performance assessment (flight examiners vs. non-flight examiners). Although somewhat older than the non-flight examiner captain peers, the flight examiners do not differ from the former in terms of years as commercial pilot or flight hours (Table 1). They differ from the former, however, in that they have served as training captains or flight examiners and therefore have conducted regular performance assessments. Two captains and one first officer (in the air force) had experience as training captains (who train pilots new to a type of aircraft), which involves performance assessment; the same first officer also had served 6.5 years as flight examiner during his previous experience in the air force.

Assessment of pilot (crew) performance is integral to the work of the participants, and the activity of assessing the performance of peers is part of their career trajectory. All pilots had been previously trained—in their company training courses—to use the company's assessment tool to assess pilot performance in pre-recorded scenarios. The airline had developed this form consistent with the recommendation to use "grade sheets" for achieving adequate ratings of pilot performance (Goldsmith & Johnson, 2002). All participating pilots had assessed at least three videos prior to this study as part of their

company training. Flight examiners had additional training and on-the-job experience in using the company's assessment model and rating tool.

3.2. Scenario

In the case of ephemeral stimuli, such as performance at work, the use of recordings have been recommended, as these preserve the information required for experts to make judgments (Weiss & Shanteau, 2003). In this research, the performance of the captain in a pre-recorded videotaped scenario was assessed. The scenario, recorded in a high-fidelity, full-motion flight simulator, was set in a company aircraft while flying into an airport that is part of the regular flight schedule of the airline. All participants were familiar with flying into this airport. Poor weather, consisting of low cloud and rain, was simulated at the airport. To accomplish a safe landing, the conditions required the pilots to fly a specific *instrument approach*, and, for this scenario, a *circling approach* was conducted. The circling approach has the aircraft descending safely on a predetermined route, using cockpit instruments to a specified altitude, called the *minimum descent altitude*. Once reaching this altitude, if visual reference to the runway is obtained, the pilot maneuvers (called circling) the aircraft around the airport while keeping it in view, and prepares the aircraft for a landing on the chosen runway. If *visual reference* to the runway cannot be maintained, generally due to aircraft having entered *instrument meteorological conditions* such as rain, low cloud or snow, then the pilots conduct a *missed-approach* (also "go-around") procedure—i.e., abort the landing attempt and climb the aircraft back to a safe altitude. The *circling approach* differs from a *runway-aligned approach* where visual maneuvering is not required. The circling approach is, as several participants noted, possibly the most hazardous kind of approach.

The research below shows that even flight examiners differed with respect their perceptions of the facts. However, certain facts about the flight in the scenario are clearly established:

- The aircraft is consistently too high and too fast on the initial part of the approach (as per standard operating procedure) and, therefore, arrives at the minimum descent altitude later than the standard operating procedures require.
- The captain notes that there is bad weather where the aircraft would be making the final turn of the approach and that the crew had to watch out. However, he does not state that they might lose sight of the runway or discuss alternatives.
- The captain (as the first officer) appears to be surprised when the aircraft enters the cloud upon engaging the final turn to land.
- The gyroscope shows that the plane, which had been inclined to the left, had been returned to a slight inclination to the right and towards the mountains (where there had been a fatal crash in the past).
- The "go-around" procedure, which would take the plane back up to 3,100 feet and out of danger, was not executed according to the sequence stated in the standard operating procedures.

The detail of the visual information available to the raters is apparent in Fig. 1. There are close-ups of the instrument panels (Fig. 1a), overview shots of the left or right sides (Fig. 1b), and overview shots of the cockpit as a whole (Fig. 1c). Although such videotaped scenarios do not have the same high fidelity that an assessment in a simulator or aircraft would have, the scenario provided sufficient information for the raters to make and justify their ratings. For example, although the indicated airspeed in the following close-up shot is very low, raters tended to notice that the *required* speed, which is indicated by a *white speed bug* corresponding to Flap 15 setting, was too high (Fig. 1a). In the same context, one assessor noted what was counted as two facts: (a) the flight

director (center) is "showing very high altitude" and (b) the "course deviation indicator isn't pointing in the direction it should be."

«««««« Insert Fig. 1 about here »»»»»»

3.3. Procedures and instructions

The think-aloud protocol as an approach to researching cognition combines a relatively strong theoretical basis with very low difficulty (Dekker & Hollnagel, 2004). For several reasons, pilots in the present study were paired rather than engaging in a standard think-aloud protocol (Ericsson & Simon, 1993). First, because the company has a *debriefing tool* (a camera contained within the simulator for the purpose of video-recording pilots flying in the simulator), pilots are familiar with the process of discussing with a flight examiner selected video clips of their own simulator-based performance. Pair-wise analysis of video therefore bears higher ecological validity than individual think-aloud protocols. Second, because one assessment form was to be completed, the assessors were forced to discuss and justify their differences. The recordings of the sessions therefore constituted *natural (verbal) protocols* of the assessment situation (Suchman, 2007). In the case of a natural protocol, the thought processes are assumed to be the same as when participants participate in a debriefing session using a debriefing tool.

The pairs assessed performance in terms of their company's normal assessment model and rating tool:

- The model is in the form of a pyramid in which three "enabling skills" (*communication*, *knowledge/procedures*, and *management*) form the base from which arrows point upward to the essential skills. The three "essential skills" are arranged in pyramidal form, with *decision making* and *flying the aircraft within tolerances* as the base to which arrows point from *situational awareness* at the top.

- The assessment tool consists of a table with the six main "skill" categories subdivided into a total of 20 elements constituting the rows and the five possible scores constituting the columns. The resulting 100 cells of the rating tool contained descriptions that mapped each performance element, such as "projection," onto 5 possible performance levels (e.g., "difficulty predicting future events" = 2; "some difficulty predicting future events" = 3).

The pilot pairs were asked to (a) rate the performance shown in the scenarios and (b) provide reasons for their ratings. Pairs were told that the research focused on the *reasons* for attributing a score and the facts that substantiate a rating. The pairs were encouraged to keep notes. They used a mouse to play, stop, and replay the scenario to confirm observations or seek further information, especially when they remembered an aspect differently.

When participants did not justify a particular assessment but simply selected a particular rating, one of the two authors would encourage them to articulate reasons or examples. To ascertain that their company's assessment model and tool had not limited them in conducting the performance assessment, each pair was asked at the end whether there was anything else that they had noticed but that was not taken into account. In all cases, the pairs confirmed that they had articulated everything that could be said about the scenario. They were also asked about their confidence in the overall rating of pass/fail to which their scores had led following company procedure (one 1 or three 2s = fail). In all cases, the pairs confirmed that the overall rating (pass or fail) corresponded to their general sense of how well the pilot had done.

3.4. Data collection

The airline made available the same seminar room in which the pilots normally take their professional development and training workshops. The assessment sessions were

recorded using three cameras. The two main cameras focused on (a) the pair of participant experts together with a copy of the videotape on a laptop computer monitor for tracking what they were currently viewing and (b) the forms and notes in front of the experts all placed within the work area marked by black electrical tape. A third, micro-computer-based camera was used for backup purposes. The camera placement allowed cross-coordination with the monitor on which the scenario was played. The database also included (a) the aircraft manufacturer's training and flight operations services manual, (b) the airline-specific standard operating procedures of how to fly a circling approach and a missed approach, (c) the aerial and geographical maps, (d) raters' handwritten notes, and (e) the "approach plate" that the scenario pilots used for this particular landing that provides them with all relevant information required for the approach. When necessary, additional information about standard operating procedures was obtained from the chief training officer.

3.5. Data analysis

Understanding and theorizing human practices requires researchers to be competent in the phenomena that are the objects of research, especially in a complex work environment such as flying a modern aircraft (Hutchins, 1995b). The second author (TJM) of this paper had been a commercial airline pilot for 22 years with a total of over 10,000 flying hours. He continues to work as a training pilot for a major company in the industry. He conducts regular workshops for several airlines and the air force related to the assessment and training of pilots.

We began the data analysis by transcribing the scenario, mapping the scenario onto the relevant standard operating procedures, and extracting all objectively available facts of the flight given in the video. A flight examiner was recruited from the participating airline to add to the transcription performance- and flight-related facts. All assessment

sessions were transcribed by a commercial service and checked for accuracy by a graduate student and by both authors. Data analysis began during the 8-day fieldwork, when the authors discussed specific observations. For example, we stated hypotheses, such as "Raters build a case description and then map this description onto the assessment grid." Subsequent joint analyses of the entire data set revealed support for the hypothesis among flight examiners; but for captains and first officers, a new hypothesis had to be created: "The assessment process is driven by the assessment tool, whereby assessor pairs seek evidence for or against a performance level of a particular assessment element," which was then tested in the data set. All gestures were coded by type (iconic, deictic, abstract, and beat) and listed together with time code; 20% of the time codes were randomly selected and the associated gestures were coded a second time, which revealed perfect agreement. We conducted analyses individual and periodically met for joint interaction analysis and calibration. One of us created tables of (a) all the facts assessor pairs stated (e.g., "Flap 15 call missing" or "Weather not flash out east") and (b) reasons and evidence mobilized in support of each category (e.g., "good perception of where they were" or "Didn't work out how weather would affect aircraft"). All tables were crosschecked for accuracy and completeness by the second author. Any disagreement was discussed until consensus was achieved. The training manager of the participating airline conducted a final check of accuracy.

Classical statistics are useful to test the degree to which data falsify the assumption of a null hypothesis; this approach, however, constitutes "a bias to overstate the evidence against the null" (Rouder et al., 2009, p. 227). In this study we use a Bayesian approach, which evaluates the support the data provides for null and alternative hypotheses—i.e., $p(H_0|data)$ and $p(H_A|data)$—rather than the probability $p(data|H_0)$ of observing the data given H_0. We chose this approach because it provides us with an indication of the strength of support for null and alternate hypothesis, which is especially important given our small sample sizes. The Bayesian approach is suitable for theory building because it

(a) allows collecting evidence *for* the null hypotheses, (b) yields results that "*may be interpreted with confidence for all sample sizes*" (Rouder et al., 2009, p. 234, emphasis added), and (c) provides better evidence than *p* values of the strength of support for null and alternative hypotheses in the form of an odds ratio (Wetzels et al., 2011). Odds ratios $1 \leq \Omega < 3$ provide "anecdotal," $3 \leq \Omega < 10$ "substantial," $10 \leq \Omega < 30$ "strong," $30 \leq \Omega < 100$ "very strong," and $\Omega \geq 100$ "extreme" evidence for one (null or alternative) hypothesis over the other (Wagenmaker et al., 2011). Thus, for example, an odds ratio of $\Omega = 10$ in favor of the alternative hypothesis means that it is 10 times more likely than the null hypothesis. The posterior odds reported here were calculated with equal prior probabilities for null and alternative hypotheses, in which case the Bayes factor is equal to the posterior odds (Rouder et al., 2009).

4. How experts with different levels of experience assess peer performance

This study was designed to investigate, in the context of aviation, how experts of different rank and experience assess other experts in their own field. Specifically, the study was to provide answers to four questions: (a) What is the degree of variability in peer performance assessment within and between levels of experience? (b) Do peers with different levels of experience engage the same performance assessment process? (c) How do the number and type of perceptual distinctions made (facts isolated) determine performance assessment? and (d) How are specific performance events mapped onto performance categories and levels of performance ratings?

4.1. Evidence for the degree of variability in performance levels

Existing research suggests that the variability of performance assessment should decrease when two or more assessors collaborate (Brannick et al., 2002; Hung et al.,

2012). The first research question pertains to the variability when pairs of pilots assess peers. A previous study with 92 pilots assessing performance individually showed consistency across ranks with respect to different scenarios: When flight examiners rated the performance of a pilot as low, so did captains and first officers (Mavin et al., 2013). The assessment profiles across the six assessment categories did not vary considerably. However, the study showed considerable within-rank variance. In the current study, the assessment profiles vary considerably between groups on each performance category. This can be seen in Fig. 2, which presents assessment profiles by rank. For example, the ratings of the *decision making* category ranged from a low of 1 (FE2), which constitutes an automatic failure in terms of company policy, to a high of 4 (FO1, FO2). Similarly, the category *knowledge* was scored from a low of 1 (CAP1) to a high of 4 (FO2). Five of the six captain pairs (FE is also a captain) rated the six categories in a way that led to a fail (one 1 or three 2s rate a fail); the sixth pair (CAP2) rated only two 2s, requiring another 2 for a failure. All first officer pairs passed the captain in the scenario. The evidence in favor of the alternate hypothesis of a difference between captain pairs (3 FE + 3 CAP) and first officer pairs (3 FO) is strong for *knowledge* (Ω = 27.8) and *management* (Ω = 10.4) and anecdotal for *situation awareness* (Ω = 1.85) and *decision making* (Ω = 1.85). There is anecdotal evidence in favor of the null hypothesis in the case of *flying the aircraft within tolerances* (Ω = 1.68) and *communication* categories (Ω = 1.24). Fig. 2 shows that the assessment profiles across categories differ considerably. This is reflected in the fact that the correlation matrix of all pairs shows that 9 of the 36 values are associated with odds ratios $\Omega \geq 3$ (substantial or greater evidence) in favor of the alternative hypothesis ($r \geq .63$); two values exhibited substantial evidence for an inverse correlation ($r \leq -.63$).

«««««« Insert Fig. 2 about here »»»»»»

4.2. Assessment experience determines assessment process

Flight examiners assess their peers on a regular basis as part of the airline industry's effort to ascertain quality and safety and, therefore, are experienced assessors of performance. Whereas some research suggests that the level of assessment experience does not correlate with level of expertise (e.g., Shanteau et al., 2002), other studies suggest cognitive differences between experienced and non-experienced assessors of workplace performance (e.g., Govaerts et al., 2011). The second research question asks whether there are differences in the assessment process between flight examiners and regular pilots (captains, first officers) that may have their origin in the different levels of assessment experience. In this section, we show (a) that flight examiners begin by establishing narratives of the video scenarios before mapping the narrative onto the categories of the assessment tool, whereas captains and first officers move step-by-step through the items of the assessment tool seeking evidence from the video for attributing a particular score; and (b) that associated with the different nature of the assessment process were differences in the use of gestures.

4.2.1. Narrative- versus assessment-tool-driven assessment

Pairs spent between 41 and 95 minutes with the scenario (X = 65.6 min, SD = 19.3) including the 6:45 minutes for watching the scenario. In some pairs, individuals began commenting while the video was still running. Following an initial playing of the flight scenario, all but one pair described in general terms what they had seen; pair FO2 immediately began the assessment. Two general types of assessment process were observable. In the first type, the assessors went from the category element in the cells of the assessment tool to seek supportive evidence (i.e., rating > fact). Frequently, the assessor pairs then took the corresponding item in the neighboring cells to the left and right, checking its fit against the video. In the second type, pairs first constructed a

holistic narrative, either (a) of the whole scenario, or (b) for each of the 6 categories. They then asked themselves how the narrative maps onto the rating scores of 1 to 5. When there was a question regarding facts, participants navigated to the pertinent sequence in the scenario and ascertained the response to their question. Overall, the pairs replayed aspects of the scenario a mean number of $X = 6.6$ ($SD = 3.2$) times (median = 5; mode = 5; min = 3; max = 12). The hypothesis of a difference in number of times replayed between those pairs failing the captain in the scenario versus those who do not is 5.88 times more likely (substantial evidence) than the null hypothesis. One might anticipate differences between the pairs in the approaches used (assessment tool-driven vs. narrative). But the present data constitute no more than anecdotal evidence in favor of a difference ($\Omega = 1.79$).

During the process of assigning scores, the company's assessment model or the order of performance categories on the assessment tool (situational awareness, decision making, tolerances, knowledge, management, communication) generally structured the rating process. However, there was considerable variation in this process between flight examiners and the other participants. The flight examiner pairs first elaborated an overall narrative and, then, moved to map their narrative onto the assessment tool. In the narratives, they described the essence of what happened in their own words before rating the performance. Within each category, they continued using their own narratives. The flight examiner pairs began with what the model describes as "enabling skills" and then talked about the "essential skills." The captain and first officer pairs generally began with very brief narrative account of what they had seen in the flight scenario and then discussed it in the order provided by the assessment tool. In two captain pairs (CAP1, CAP2) and two first officer pairs (FO1, FO3), there was a brief overall narrative, and then the pairs commenced working through the assessment form. CAP2 stood out from all other pairs. After producing a short narrative, they jumped back and forth between the categories in an attempt to score these. FO2 approached the task similarly to the flight

examiners, except that they established narratives within each of the six performance dimensions. They began with the "enabling skills" from bottom up (like FE1 & FE2), and then assessed the "essential skills" top down (as did FE3).

In five of the pairs, the evaluation process was driven by the structure of the assessment tool. Most pairs followed the main performance categories from top to bottom in the order situational awareness, decision making, flying the aircraft within tolerances, knowledge, management, and communication. CAP1 began with situational awareness, and then did the three factors from the enabling skills (knowledge, management, communication) prior to completing the essential skills in the order tolerances, decision making, and situational awareness. The pair pointed to a particular performance element and then provided evidence for or against giving it a particular rating. For example, while discussing perception, CAP1 identified the fact that the flying pilot had missed the minimum descent altitude and the direction in which to fly during the missed-approach procedure. A lower rating was suggested. The pair suggested that in neglecting the wind, the captain had lacked awareness of clear and obvious systems or environmental factors; and the wind direction from behind should have entered the captain's deliberations. The pair first suggested attributing a score of 1. They then considered possible evidence for assigning a score of 2. Asserting that the captain had missed a crucial aspect of the flight (setting the altitude), the pair then "headed down," assigning a 1 as the score most consistent with the observed performance. The process has the structure {category > element > rating > fact}.

Narration of what is happening was an important part in many assessment sessions—though it was much more prevalent in the three FE pairs than in the CAP and FO pairs. In these latter pairs, this form of encoding the events was brief and occurred at the beginning and prior to orientating towards the assessment tool. In the case of the FE pairs, creating a coherent narrative dominated the assessment session. The pairs built a case description to encode the scenario, characterized it as a pass or fail, and then (a) mapped

individual observations onto elements that were combined to categories or (b) used these observations to holistically score the category. They might conclude, "I don't consider decision making as an area that caused the issues they had there. It's more related around their knowledge, workload" (FE3), and then engage in rating accordingly. The narrative was used as the source for the data that the assessors used in support of a rating. Although some of the category names might appear in the narrative, the use was not related to the assessment tool. Following the building of a case scenario in their own words, the examiners then shifted to the tool and worked backwards from there to the case that they had been building. For example, FE2L proposed to start with *communication* and then described it to be "bull's-eye." This process was also used, and therefore of special importance, where the captain was noted as being confused, and announced his confusion, which allowed the problem to be captured and dealt with. The *communication* was judged to be "nice and clear," but there were deemed to be "some issue with [its] timeliness" and the group provided a clear example from the scenario. The process has the structure {narrative > pass/fail > category > (element >) fact > rating}.

4.2.2. Prevalence of (iconic) gestures in narrative-driven assessment

The preceding subsection shows that there were two very different assessment processes at work, one organized around the production of an overall narrative capturing the essence of the performance seen in the scenario, the other one driven by the assessment tool. Given the research on the deep integration of gestures with narratives (McNeill, 2005; Nûñez & Cornejo, 2012) and other cognitive capacities such as spatial orientation (Haviland, 1993; Widlok, 1997), we might anticipate difference in gesture use between the two types of processes. Such an expectation is reasonable because the narration of stories and situation descriptions tends to be accompanied by coproduced gesticulations and independent of culture or of the actual presence of the interlocutor (e.g.,

on-telephone talk) (Kendon, 1997; Levinson, 1997). Fewer gestures may be anticipated when the assessment process is oriented towards matching particular observational facts (cues) to a given assessment category. In this section, we describe the use of gestures in the assessment process and evaluate the evidence for different hypotheses about gesture use.

In this study, gestures were used to a different extent by pilots with different levels of assessment experience, a difference associated with the distinction between narrative- and assessment-tool-driven process. As part of the situation description (narrative), the flight examiners especially played out the scenario as if they were sitting in the cockpit, with gestural movements directed towards the places where the switches and buttons normally would be located. The gesture would be pushing if the corresponding button needed to be pushed or would be a flicking gesture if a switch had to be thrown. Thus, for example, one flight examiner talked about the situation whereby the missing push of the go-around button led to the flight director telling the pilot in the scenario to fly down when in fact he was to raise the nose of the aircraft and fly up. He had quite frequently seen that pilots followed the instructions of the flight director instrument when it pointed down (Fig. 3a). He accompanied the account using a gesture that is holding the control column (Fig. 3b), which he pushed forward (Fig. 3c), which would have made the aircraft head downwards. The flight examiner gestured the result of these actions by showing how the aircraft would fly from the perspective of the flight deck: an approximately parabolic flight path that comes to an end with the "bomp" (Fig. 3d). He then returned to talking about the scenario, explicating that it is dangerous to leave the command bars pointing down while gesturing the position of what the pilot would have seen on his instrument panel (Fig. 3e) following his engagement in the final turn (Fig. 3f). As a result, the flight examiner not only narrated events but also re-enacted the specific flight procedure as if he was currently flying.

«««««« Insert Fig. 3 about here »»»»»»

In the present dataset, there are three types of iconic (depicting) gestures: (a) reenacting movements specific to the operation of the aircraft, such as pushing buttons (e.g., go-around button), turning knobs, and moving the control column; (b) featuring the orientation or flight path from the perspective of the pilots in the cockpit; and (c) modeling flight path as seen from the outside. Striking differences existed in the use of gestures associated with the assessment process. For the three pairs employing a narrative-driven approach, there was a mean of $X_B = 82.3$ ($SD_B = 42.5$) iconic gestures in the course of assessing the captain's performance in the scenario. There was a mean of $X_A = 22.7$ ($SD_A = 17.5$) iconic gestures among the six groups that followed the structures of the assessment model or tool. Given $\Omega = 6.67$, there is substantial evidence against the null hypothesis. One might assume that the number of gestures is related to the amount of time talked, the number of facts talked about, or the number of times the assessor pairs actually watched the video (intensity of engagement). Evidence for the alternative hypothesis regarding the number of flight-related facts is substantial ($\Omega = 3.88$); the alternative hypothesis regarding the number of times the assessors returned to the video to watch particular segments is 1.98 times as likely as the null hypothesis (anecdotal); there is no evidence for the alternative hypothesis in the case of amount of time spent in the analysis.

When the referents of the gestures are considered separately, there is strong evidence ($\Omega = 10.8$) for the hypothesis of differences between the flight examiners ($X = 33.7$, $SD = 15.8$) and the other pilots ($X = 8.7$, $SD = 5.9$) in terms of the numbers of gestures directly reenacting the operation of the aircraft or the perception of instrument readings. When the number of these gestures was evaluated relative to the total number of gestures within each pair, it was evident that flight examiners and other pilots employed gestures embodying the operation of the aircraft in the same proportion of the total number of gestures ($X_{FE} = .46$, $SD_{FE} = .24$; $X_{CAP,FO} = .52$, $SD_{CAP,FO} = .29$). Overall, there were few abstract gestures referring to the flight generally: of a total of 385 gestures in the 9

sessions, only 14 were of this kind. That is, about half of the gestures directly pertained to what the pilot did or should have done in flying the aircraft, especially during the crucial moments when the aircraft entered the cloud.

4.3. How number and type of perceptual distinctions relate to performance ratings

Discrimination and cue utilization are deemed core dimensions of judgment expertise (Esogbue & Elder, 1983; Weiss & Shanteau, 2003). The third research question concerns the relationship between number and type of perceptual distinctions made (facts observed) and performance ratings. In the present study, for each of the six categories, there was considerable variation with respect to (a) the nature of the fact used as evidence to support an assessment, (b) the number of facts used as pieces of evidence, and (c) the gravity with which a fact counted against the pilot. These variations are exemplified with respect to the *knowledge/procedure* category (Table 2). Most pairs made some general reference to the go-around/missed-approach procedure as having been problematic and that the scenario captain displayed confusion, minor errors, or lack of familiarity. However, with respect to specific facts supporting a particular score, the five pairs (i.e., FE1, FE2, FE3, CAP2, & CAP3) who rated the captain 2 on *knowledge/procedure*, used different facts to support the score. Furthermore, there is considerable variation in the number of specific facts articulated, ranging from 1 (CAP2) to 8 (FO3). Third, there is considerable variation in how specific facts influence the final score. For example, there is considerable overlap in the 7 and 8 pieces of evidence used by CAP1 and FO2, respectively. Whereas CAP1 decided that the performance should be rated as an automatic fail (the lowest of all assessment scores), FO2 attributed to the captain the highest score (4) for similar facts. Pairs FO2 and FO3 were influenced by positive mediating facts—e.g., giving a higher rating because "the procedure was correctly flown."

«««««« Insert Table 2 about here »»»»»»

To test whether there were any differences in performance assessment between pairs who failed the pilot in the scenario versus pairs who arrived at a pass according to company policy, we counted the number of specific facts articulated by each pair. For the pairs rating the performance as a fail (3 FE and 2 CAP), there were X_F = 18.6 facts (SD_F = 3.6), whereas for the four groups (1 CAP and 3 FO) rating the captain's performance as a pass there were X_P = 13.3 facts (SD_P = 3.6). The evidence in favor of the alternate hypothesis is anecdotal (Ω = 2.63). If the two outlier pairs are not considered, the means are X = 20.0 (SD = 2.2) and 11.7 (SD = 2.1), respectively; the evidence for the alternative turns out to be strong, the alternative hypothesis being 21.3 times more likely than the null (Ω = 21.3). In the groups assessing the performance as a fail (n = 5), FE1 stands out in that their assessment process differed from other pairs, the number of facts considered was smaller, the comments were more generic, and they had to be encouraged to complete the assessment form. Among those passing the scenario pilots (n = 4), FO2 stands out by the number of facts considered (n = 18). An investigation of the facts being articulated in the course of justifying a particular assessment score shows that across the scenario as a whole, only one fact was articulated by all groups: that the captain was uncertain about the turn direction upon entering the cloud in the final turn. Eight groups either saw that the captain failed to push the go-around button or inferred the failure to do so from the instruments. Only 2 groups stated that the "go-around call", which would have been the obligatory first step in flying the go-around/missed-approach procedure, was missing.

4.4. How events are mapped onto performance categories and ratings

Previous research shows that assessors may use the same performance event to make judgments about different performance categories (Holt et al., 2002; Horng et al., 2010).

We may anticipate that how pilot pairs assess the performance of peers is a function of what they observe to be happening. Studies of cue utilization suggest that if a particular demonstrable fact or event is not salient then it will not enter the assessment (Skåner et al., 1998). At the same time, when a fact is salient and considered, the role it plays in assessment of a particular performance may differ between assessors because (a) raters weight the evidence differentially or (b) the fact is not or cannot be stated in a precise manner. In the following two subsections we exemplify how the same segments from the scenario (raw data) become evidence for different performance categories and lead to different performance ratings (gravity). Two issues in the scenario turned out to be determining for justifying the ratings: (a) the captain had noted the rain cloud in the area where they would be circling and did nothing about it and (b) pilot behavior during the initial period after the airplane had entered the cloud.

4.4.1. Consequences (or lack thereof) of noticing bad weather

The first scenario segment that heavily weighed in the assessments pertained to the instant when the captain sees the area where he had to make his final turn to align the aircraft with the runway. The captain notes, "the rest doesn't look too crash hot," and follows this with a conclusion, "So um out to the east there, we'll have to be careful." However, the captain does not review what to do in case the aircraft would enter the cloud. Different assessor pairs attributed the segment to problems in different performance categories.

For some assessors, the segment was part of the evidence that the pilot did not have good *situational awareness*; others suggested that it was an instance of poor *decision making*. Three pairs (FE1, CAP1, FO2) did not make any reference to the statement about weather; the captain's reaction to the weather did not enter their performance ratings. Six assessor groups (FE2, FE3, CAP2, CAP3, FO1, FO3) made explicit reference to the

comment about the weather. In five groups, this weighed heavily against the captain, because his (non-) action led to a major threat to the safety of the aircraft. In two of these groups (FE3, CAP2), the issue was attributed to a "major weakness" in *situational awareness*. In three groups (FE2, CAP3, FO3) it was the *decision making* that was deemed problematic. In each case, the score attributed was the lowest that the assessor group assigned (down spikes in Fig. 2).

In the case of FE3, the captain's statement was evidence that he had *perceived* the bad weather but had failed to *comprehend* its implication for the future flight path ("it hasn't actually sunk in"), and he did not *project* ahead to anticipate having to fly a missed-approach procedure. This was evident in the very early part of the assessment, where the pair worked on elaborating a general narrative of the scenario. As the protocol shows, the descriptions this pair used pertain to the *situational awareness* category. They noted that the captain realized what was happening outside the cockpit but then did not project the implications of the outside situation onto the inside, that is, the captain did not engage in actions to assure flight safety. The pair did not talk about the weather in the context of other human factors. They asserted that the pilot did not lack awareness of clearly obvious systems or environmental factors and chose to rate the performance as a 2.

For three pairs, the key factor that brought about the dangerous situation was faulty decision making: the captain saw the poor weather to the east but did nothing about it. Options that had been considered included going lower than the cloud to avoid it; but in any event, the possibility of losing visual contact should have led to a quick reconsideration of what to do *if* the aircraft was entering the cloud. For example, FE2 summarized the assessment in this way: "It was just *almost completely hopeless in the decision making*. And it led to a, sort of led on to a situation where they got confused, they got themselves into a dangerous situation." The weather incident heavily weighed on the performance, which reflected "not very good decision making." It was so bad that it could "not be dressed up," because there was "no contingency planning," in fact, "there

wasn't any planning." The pair listed the absence of alternative action plans and noted the lack of planning for contingencies. The pair noted that the flying pilot had a grasp of the fact that the situation was dangerous, and therefore took into account facts and showed evidence of diagnosis. If the captain had planned a missed approach, he would "not [have] buggered it up" in the way he did. The pair noted the lack of appreciation for time in the sense that as a result of the failure to make a decision earlier, the captain had to make a rushed decision when flying into the cloud. The pair concluded that the failure to make a decision with respect to the poor weather and the possibility of having to fly a missed-approach procedure would mean a 1 or 2 rating. With respect to situational awareness, this pair concluded that despite the weather issue, there was sufficient evidence that the situational awareness "actually wasn't too bad."

4.4.2. Entering the cloud: Missed approach and go-around

The second segment that heavily weighted on scoring was the captain's execution of the missed-approach procedure. The video shows how upon entering the cloud, the captain announces that he has lost visual contact with the airport. The captain levels the plane while the first officer comments that there is a "rain shower there," and just as the plane is level, the captain says, "Missed approach is right hand," followed by saying, "Wasn't it?" The first officer says "negative," and repeats "left hand, left hand," emphasizing the word "left." The captain queries, "it is left?" and the first officer confirms, "Left hand."

This fragment played a critical role in the performance assessments. However, because the raters differed in what they saw happening, they came to very different conclusions about the performance. Three assessor pairs (FE1, FO1, FO3) saw the aircraft turning right, "rolling the wrong way," noting that the captain "was initially going to turn the wrong way." One assessor pair described what was happening as the pilot

"going to turn the wrong way [right]" and "wanting to go right" without actually stating that the aircraft had started rolling to the right. Finally, five assessor pairs (FE2, FE3, CAP1, CAP2, FO2) noted that there was "confusion" and that "rather than just do the wrong things," the captain sought confirmation about whether the turn was right before actually turning the wrong way. Those pairs who saw the aircraft rolling to the right or the captain wanting to go right heavily counted their observation against the captain. The pairs that saw the captain seeking confirmation about the direction to fly noted the confusion, which ought not to have existed, but counted the confirmation-seeking question heavily in favor of the captain and the positive communicative climate in the cockpit. This climate allowed lack of *knowledge* and confusion to be acknowledged, brought into the open, and corrected prior to engaging in procedures. Three pairs of assessors (FE1, FO1, FO3) agreed that the captain actually was turning to the right—the angle of the wings to the horizon is about 3 degrees to the right—before being corrected and then turning to the left. This kind of assessment is clearly evident when the pair FO3 talked about that instant of the scenario. The pair stated that there *was* a brief turn to the right "beyond tolerances," but that this was corrected. The pair CAP3 agreed that there was evidence for an intention to go right in the missed-approach procedure without an actual turn being executed. Because this error was trapped, an unpleasant outcome was only a potential danger that was actually avoided.

5. A fuzzy logic model of performance assessment

In the preceding section we provide answers to four research questions concerning how pilots with different levels of assessment experience assess the performance of peer experts. The results presented show that

- there exists considerable variance with respect to specific ratings along 6 performance categories, even though the ultimate assessment of whether a performance constitutes a pass or fail is relatively consistent within ranks;
- the assessor pairs differed with respect to the number, type, and gravity of facts that were identified and used as the basis of performance assessment; and
- different segments of the flight scenario were used to arrive at an assessment of the same performance dimension, and the same segment was used to evaluate different performance dimensions.

That is, although the assessor pairs stated good reasons—i.e., sufficiently good to convince peers—there was considerable variation in their performance assessments. These observations led us to hypothesize that assessment—although airlines treat it as a measurement task with associated interests in inter-rater reliability—may at best contain ordinal information (Goldsmith & Johnson, 2002). More likely, performance assessment is a judgment and decision-making task poorly modeled by standard psychometric approaches (Govaerts, van der Vleuten, Schuwirth, & Muijtjens, 2007). To address the problems of inter-rater reliability approaches to performance assessment, several recent studies suggest the use of fuzzy set theory (Ciavolino, Salvatore, & Calcagnì, 2013; Özdaban & Özkan, 2010). Consistent with this suggestion, we provide in this section a fuzzy mathematical approach to modeling performance assessment.

The findings showed how descriptions used to encode observations tended to be inexact, fuzzy, and in degrees. For example, the captain was observed (a) to have turned right, (b) wanting to go right, and (c) seeking confirmation whether to turn right. These observations are consistent with research findings that inter-rater reliabilities in aviation tend to be moderate to low (Brannick et al., 2002; Mavin et al., 2013). The present study shows that the variance is not random: assessors tended to provide (good) reasons for their assessments, thus, the variance appears to be grounded rather than due to error as assumed in the inter-rater reliability model. In fact, the descriptions provided so far are

consistent with the contention that the assessors (a) are engaged in a classification rather than measurement and (b) use ambiguous, fuzzy inputs (facts) and fuzzy rating categories. We therefore adopted a fuzzy logic approach to mathematically model performance assessment similar to the ways in which medical diagnosis is modeled, because these ways distinguish between fuzziness and random error (Esogbue & Elder, 1980; Klir & Yuan, 1995). We model three stages in performance assessment, how facts combine to elements and categories.

For each performance level {unsatisfactory, minimum standard, satisfactory, good, very good}, two fuzzy sets are created specifying the lower ($\mathbf{B}_L$) and upper ($\mathbf{B}_U$) boundaries, thereby providing a map that translates between fuzzy rating levels and observed (imprecise, fuzzy) facts ($B_L(i), B_U(i) \in I; I = \{0,1\}$). As a previous study with 92 pilots showed that there was a preference for ratings to be in the mid-range (2–4) and a dispreference for extreme values (1, 5) (Mavin et al., 2013), we defined the boundaries for the categories with a larger middle section as {0, .15, .35, .65, .85, 1} (i.e., unsatisfactory is associated with $\mathbf{B}_{L,u} = 0$ and $\mathbf{B}_{U,u} = 0.15$). A fuzzy relation **W** specifies the weight of a specific (fuzzy) rater observation (fact) on the performance level. For example, in considering an "unsatisfactory [u]" rating for situational awareness, the fact that the scenario pilots were considerably off the aircraft's glide path was deemed as irrelevant, meaning that the weight would be $\mathbf{W}_{u,GP} = 0$. In a "very good [vg]" rating, however, the glide path would need to have been perfect, which means the weight of glide path assessment and rating would be $\mathbf{W}_{vg,GP} = 1$.

A fuzzy set of observations **A** encodes everything an assessor uses as evidence in support of a rating. For example, considering communication to be "nice and clear," but with some evidence that it has been slow or delayed in an instance, would be associated with a corresponding value in **A** of .65 (between satisfactory and good). The observation that a performance was seen as "bang on" is coded as .9. Table 3 provides an example of assigning fuzzy codes to observations. When a category (e.g., risk consideration) was not

addressed—corresponding to undiagnosed symptoms in the medical field—the corresponding rows in the matrixes were eliminated.

«««««« Insert Table 3 about here »»»»»»

For each performance category, clustering technique is used based on the (Euclidean) distance between a rater pair's fuzzy set of *n* observations and each fuzzy performance category:

$$D=\left[\sum_{j=1}^{n}[\mathbf{W}_j(\mathbf{B}_{L,j}-\mathbf{A}_j)]^2+\sum_{j=1}^{n}[\mathbf{W}_j(\mathbf{B}_{U,j}-\mathbf{A}_j)]^2\right]^{1/2}$$

In its assessment of decision making, the pair FE2, for example, stated that there were 2 main decision points: one when the bad weather was detected; the other when the aircraft entered the cloud. They considered the first to be the important one. The pair stated evidence that the captain was aware of the bad weather and rated his diagnosis of the weather as "pretty good." We chose a fuzzy observation of $\mathbf{A}_2 = .7$. However, the statements "completely hopeless," "got confused," "[the decision] could have led to the world's biggest disaster" were used to describe all other aspects, and the performance aspects considered were "all bad." We therefore coded these aspects (timeliness, options created, risk considered, and contingency planning) with fuzzy observations $\mathbf{A}_j = .1$. The resulting distances for the 5 categories from unsatisfactory to very good were D = 0.76, 0.82, 1.33, 1.93, and 2.40. The fuzzy logic model, therefore, predicts a rating of unsatisfactory, consistent with the score of 1 that FE3 assigned to decision making. Table 4, which maps the actual ratings provided against those that the fuzzy logic model predicted, shows good agreement. In the cases of misclassification (FE1, FO2), the distances for neighboring categories are close or identical; in the case of FE3, although the classification was correct, the differences in distances between a satisfactory and good rating are close, consistent with the rater pair's overall comment that they could have considered rating decision making a 4, but for the "overall result of what was going on there."

«««««« Insert Table 4 about here »»»»»»

6. Discussion

This study was designed to find answers to four research questions about how experts of different levels of assessment experience assess the performance of other experts in their field. Concerning the first question about variability of performance assessment when assessors work in pairs, the findings show that there were differences in the pass/fail ratings, and in the reasons provided in justifying the rating, along the lines of experience. These differences existed even though the assessors worked in pairs, discussing evidence and reasons for their assessment, which had been suggested to increase inter-rater reliability (Brannick et al., 2002). The variability suggests that there may in fact exist good reasons for the variations observed (Govaerts et al., 2007). Although it might be thought that the scenario provided assessors with objective evidence—what the scenario pilots said and what the aircraft has done is available for everyone to see, hear, and evaluate—the narrative encodings and the categories and elements of the assessment model and tool were imprecise and fuzzy. Consistent with previous research (Holt et al., 2002; Horng et al., 2010), what assessors actually took as facts from the scenario differed in kind and degree. The present study confirms observations in other studies that reported good item–total correlations but relatively poor agreements between raters (e.g., Brannick et al., 2002); but it is inconsistent with another study that reported high inter-rater agreement at the category level and low inter-rater agreement at the pass/fail level (O'Connor et al., 2002). Although there have been suggestions that performance categories currently used in the assessment of pilots are too fuzzy to be of explanatory value (Dekker & Hollnagel, 2004), the present data show that in the field, terms such as *situational awareness* are readily applied and used as explanatory resources.

The second research question concerns similarities and differences in the performance assessment process. There were differences with respect to the overall assessment process, whereby the flight examiners first produced a narrative, holistic encoding of the scenario, drawing, to a significantly larger extent than their peers, on embodied forms of knowledge (i.e., gestures, orientations) (e.g., McNeill, 2005; Núñez & Cornejo, 2012), and then mapped this encoding onto the assessment tool. The detail provided by elements of the assessment tool was less important than the holistic narratives they were producing together for the purpose of capturing the essence of the episodes. This is consistent with other information that was collected, for example, when flight examiners say that they "generally remember the big issue thing and the big issue is the revelation to [them] in what [they] determined as the problem, not the small specifics that were making [them] feel uncomfortable in the first place." It is also consistent with an ongoing ethnographic study of the 6-monthly evaluations where simulator performance is immediately followed by debriefing meetings (unpublished data). In these meetings, and without much preparation time, flight examiners present the evaluated pilot with their overall judgment, supported with episodes from the simulator session. The prevalence of iconic gestures in this group re-enacting the movements required to fly the aircraft and making salient where and what to read from the instrument panel, associated with the gestures about the flight path of the plane from the perspective of the pilots (which represents embodied coding of the situation), is in support of the holistic encoding and embodied forms of knowledge underlying the assessment process. That is, gestures appear to be integral to the assessment-related cognition, much in the way they are in spatial orientation and cognition more generally (Haviland, 1993; Widlok, 1997). Even though the flight examiner pairs differed in the specific facts and ratings attributed to the performance categories, they agreed in the overall narrative and consistently rated the performance as fail.

The assessment process of captains and first officers was mediated by the assessment model and tool such that these individuals first selected categories and elements and then looked for evidence in support of one or another rating. The process was less about establishing a larger, holistic picture and more about accumulating facts required to rate the individual assessment elements. Previous research had suggested that performance assessment was constituted by a sequence whereby identified facts are attributed to categories, which are then analyzed element by element (O'Connor et al., 2002). On the basis of the implicit or explicit combining rules, an overall assessment is obtained. In this study, a more differentiated process emerged. In the case of the less experienced experts, the process can be represented in the structure categories > elements > facts > rating > combing rules > category rating, where categories and their elements were selected, followed by a search for supporting facts (cues), which then led to a rating. Company combing rules were used to generate the pass/fail rating from the ratings of the categories.

The third research question investigated how the number and type of perceptual distinctions (cues, facts) affected performance assessment. Previous research suggested that more experienced assessors of performance in the workplace made more inferences whereas less experienced assessors produced more literal descriptions (Govaerts et al., 2011). Previous studies also show that as assessors develop greater expertise, they discriminate larger and more relevant amounts of information during situational appraisal (Klein, 2003; Weiss & Shanteau, 2003). In this study, there were differences in the way specific facts and specific events from the episode entered the assessment process. Compared to other pairs, those pairs who failed the pilots in the videos used a significantly greater number of observational facts to determine the grade. But simply using a greater number of facts did not clearly determine the pass/fail decision. Rather, further analysis yielded a far more complex scenario. For instance, once a fact had been identified, assessor pairs categorized the same facts differently (e.g. *situational awareness* versus *decision making*). Furthermore, the weighting that was assigned to each

fact differed between pairs. The picture is complicated by the structure of the assessment process. As part of constructing the overall, more or less coherent, intelligible narrative, the more experienced pilots (flight examiners) identified salient facts. These tended to be associated with a key aspect in the flight—e.g., noticing the bad weather and what was or was not done upon noticing, or the issue of querying the turn direction upon the missed-approach actions. That is, these observational facts fit in and supported the overall narrative—consistent with those theoretical analyses of narrative that establish a close connection between characters, plot, and the time-spaces of the rendered events (Bakhtin, 1981). In the assessment-tool-driven process, facts were identified to dis/confirm the current rating.

Associated with the third research question, the fourth question investigated how specific performance events influence the choice of category assessed and the level of performance rating. This study shows that the same performance events—e.g., the pilot noticing bad weather and his subsequent actions—were used as evidence for evaluating very different categories, such as situation awareness or decision making, consistent with arguments that such categories are too ill-defined to be used in performance assessment (e.g., Bergström et al., 2011; Dekker & Hollnagel, 2004). Moreover, the same performance events were perceived very differently between pairs, as exemplified in the pairs' noticing of the pilot intending to turn, making a turn, or seeking confirmation for the nature of a turn. How the assessor pairs perceived the event affected both the category evaluated and the performance rating (e.g., poor situation awareness vs. good communication). The same performance event therefore became a different type of cue in different rater pairs, consistent with claims that complexity and context-specific features mediate quality and consistency of performance assessment (Govaerts et al., 2007).

In sum, this study shows that rater pairs mobilized good, albeit imprecise and qualitative rather than quantitative, reasons to justify their performance assessments during the discussions within rater pairs. Because rater pairs used multiple pieces of

evidence (facts), their rating process might be considered similar to the process outlined by the constraint satisfaction model, which describes the formation of an interpretation on the bridge of a navy vessel (Hutchins, 1995a). However, Hutchins' model, as decision-making models more generally (Esogbue & Elder, 1979), assumes defined inputs and hypotheses. The present study shows, however, that the performance descriptions that the raters provide are not hard and fast—as the ratings themselves may suggest—but are imprecise, as evidenced by the fact that the same observation may be used as evidence for the quality of *situational awareness* or *decision making*. That is, the variance observed is not due to error, as modeled in inter-rater reliability calculations, but is based on categorical judgments and decisions (e.g., Govaerts et al., 2007). Such judgments can be explained with a fuzzy logic model (Ciavolino et al., 2013). This study shows high consistency between actual and predicted ratings and, thereby, confirms the results of another study (Roth & Mavin, 2013). The fuzzy logic model treats assessment as a categorization process based on fuzzy observations, elements, and categories.

This study has considerable practical consequences, because pilot assessment is a crucial element in increasing the safety of air traffic. To be useful, assessment has to be reliable and consistent. However, research shows that inter-rater reliabilities in aviation performance assessment, even with extended rater training, tend to be low to moderate (Brannick et al., 2002; Holt et al., 2002; Mavin et al., 2013; Smith et al., 2008). The results of the present study provide an explanation of the difficulty in achieving high inter-rater reliability. If there are good reasons for the observed variances, then airlines might be ill-advised spending resources on traditional training methods developed to improve inter-rater reliability—e.g., behavioral observation training or frame-of-reference training (Woehr & Huffcutt, 1994). Instead, such companies might capitalize on the observed variance as part of their training in a resilience engineering approach based on the idea that human error and minimum breakdown during training are required to improve safety (e.g., Amalberti, 2001; Saurin, Wachs, & Henriqson, 2013).

Acknowledgments

This research was funded by a grant from the Griffith University Industry Collaborative Scheme. We are grateful to the participating airline and pilots and to David Weber for contributing to the verification of the transcriptions.

References

Amalberti, R. (2001). The paradoxes of almost totally safe transportation systems. *Safety Science, 37*, 109–126.

Arts, J. A. R., Gijselaers, W. H., & Segers, M. S. R. (2006). Enhancing problem-solving expertise by means of an authentic, collaborative, computer supported and problem-based course. *European Journal of Psychology of Education, 21*, 71–90.

Bakhtin, M. M. (1981). Dialogical imagination. Austin, TX: University of Texas Press.

Bergström, J., Henriqson, E., & Dahlström, N. (2011). From crew resource management to operational resilience. In E. Hollnagel, E. Rigaud, & D. Besnard (2011), *The fourth resilience engineering symposium* (pp. 36–42). Paris: Presses des Mines.

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*, 241–261.

Chi, M., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121–152.

Ciavolino, E., Salvatore, S., & Calcagnì, A. (2013). A fuzzy set theory based computational model to represent the quality of inter-rater reliability. *Quality and Quantity*. doi:10.1007/s11135-013-9888-3

Connors, M. H., Burns, B. D., & Campitelli, G. (2011). Expertise in complex decision making: The role of search in chess 70 years after de Groot. *Cognitive Science, 35*, 1567–1579.

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

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

Esogbue, A. O., & Elder, R. C. (1979). Fuzzy sets and the modelling of physician decision making processes, part I: The initial interview-information gathering session. *Fuzzy Sets and Systems, 2*, 279–291.

Esogbue, A. O., & Elder, R. C. (1980). Fuzzy sets and the modelling of physician decision making processes, part II: Fuzzy diagnosis decision models. *Fuzzy Sets and Systems, 3*,1–9.

Esogbue, A. O., & Elder, R. C. (1983). Measurement and valuation of a fuzzy mathematical model for medical diagnosis. *Fuzzy Sets and Systems, 10*, 223–242.

Flin, R., Martin, L., Goeters, K., Hörmann, H., Amalberti, R., Valot, C., & Nijhuis, H. (2003). Development of the NOTECHS (nontechnical skills) system for assessing pilots' skills. *Human Factors and Aerospace Safety, 3*, 97–119.

Goevarts, M. J. B., Schuwirth, L. W. T., Van der Vleuten, C. P. M., & Muijtjens, A. M. M. (2011). Workplace-based assessment: Effects of rater expertise. *Advances in Health Sciences Education, 16*, 151–165.

Goevarts, M. J. B., Van der Vleuten, C. P. M., Schuwirth, L. W. T., & Muijtjens, A. M. M. (2007). Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. *Advances in Health Sciences Education, 12*, 239–260.

Goldsmith, T. E., & Johnson, P. E. (2002). Assessing and improving evaluation of aircrew performance. *The International Journal of Aviation Psychology, 12*, 223–240.

Güss, C. D., Tuason, M. T., & Gerhard, C. (2010). Cross-national comparisons of complex problem-solving strategies in two microworlds. *Cognitive Science, 34*, 489–520.

Haviland, J. B. (1993). Anchoring, iconicity, and orientation in Guugu Yimithirr pointing gestures. *Journal of Linguistic Anthropology, 3*, 3–45.

Hmelo-Silver, C., & Pfeffer, M. G. (2004). Comparing expert and novice understanding of a complex system from the perspective of structures, behaviors, and functions. *Cognitive Science, 28*, 127–138.

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*, 305–330.

Horng, E. L., Klasik, D., Loeb, S. (2010). Principal time-‒‐use and school effectiveness. *American Journal of Education, 116*, 491–523.

Hutchins, E. (1995a). *Cognition in the wild*. Cambridge, MA: MIT Press.

Hutchins, E. (1995b). How a cockpit remembers its speeds. *Cognitive Science, 19*, 265–288.

Hung, S.-P., Chen, P.-H., & Chen, H.-C. (2012). Improving creativity performance assessment: A rater effect examination with many facet Rasch model. *Creativity Research Journal, 24*, 345–357.

Kendon, A. (1997). Gesture. *Annual Review of Anthropology, 26*, 109–128.

Klein, G. (2003). *Intuition at work*. New York, NY: Doubleday.

Klir, G. J., & Yuan, B. (1995). *Fuzzy sets and fuzzy logic: Theory and applications*. Upper Saddle River, NJ: Prentice Hall.

Levinson, S. (1997). Language and cognition: The cognitive consequences of spatial description in Guugu Yimithirr. *Journal of Linguistic Anthropology, 7*, 98–131.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*, 543–560.

Mavin, T. J., Roth, W.-M., & Dekker, S. (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*, 53–62.

McNeill, D. (2005). *Gesture and thought*. Chicago: University of Chicago Press.

Núñez, R. E., & Cornejo, C. (2012). Facing the sunrise: Cultural worldview underlying intrinsic-based encoding of absolute frames of reference in Aymara. *Cognitive Science, 36*, 965–991.

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

O'Connor, P., O'Dea, A., Flin, R., & Belton, S. (2008). Identifying the team skills required by nuclear power plant operations personnel. *International Journal of Industrial Ergonomics, 38*, 1028–1037.

Özdaban, I., & Özkan, C. (2010). A fuzzy method on determining of job and personnel evaluation results, and matching them with suggested model. *International Journal of Industrial Engineering, 17*, 334–340.

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

Roth, W.-M., & Bowen, G. M. (2003). When are graphs ten thousand words worth? An expert/expert study. *Cognition and Instruction, 21*, 429–473.

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*, 73–82.

Rottman, B. M., Gentner, D., & Goldwater, M. B. (2012). Causal systems categories: Differences in novice and expert categorization of causal phenomena. *Cognitive Science, 36*, 919–932.

Rouder, J. N., Speckman, P. J., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237.

Saurin, T. A., Wachs, P. & Henriqson, É. (2013). Identification of non-technical skills from the resilience engineering perspective: A case study of an electricity distributor. *Safety Science, 51*, 37–48.

Schraagen, J. M. (1993). How experts solve a novel problem in experimental design. *Cognitive Science, 17*, 285–309.

Shanteau, J., Weiss, D. J., Thomas, R. P., & Pounds, J. C. (2002). Performance-based assessment of expertise. How to decide if someone is an expert or not. *European Journal of Operational Research, 126*, 253–263.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal, 76*, 27–33.

Skåner, Y., Strender, L-E., & Bring, J. (1998). How do GPs use clinical information in their judgments of heart failure? A clinical judgment study. *Scandinavian Journal of Primary Health Care, 16*, 95–100.

Smith, M. V., Niemczyk, M. C., & McCurry, W. K. (2008). Improving scoring consistency of flight performance through inter-rater reliability analyses. *Collegiate Aviation Review, 26*, 85–93.

Stains, M., & Talanquer, V. (2008). Classification of chemical reactions: stages of expertise. *Journal of Research in Science Teaching, 45*, 771–793.

Suchman, L. (2007). *Human-machine reconfigurations: Plans and situated actions* (2nd ed.). Cambridge: Cambridge University Press.

Wagenmaker, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Behm (2011). *Journal of Personality and Social Psychology, 100,* 426–432.

Warren, J. S. (2011). "Generic" and "specific" expertise in English: An expert/expert study in poetry interpretation and academic argument. *Cognition and Instruction, 29,* 349–374.

Weiss, D. J., & Shanteau, J. (2003). Empirical assessment of expertise. *Human Factors, 45*, 104–114.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6,* 291–298.

Widlok, T. (1997). Orientation in the wild: The shared cognition of Hai||om bushpeople. *Journal of the Royal Anthropological Institute, 3*, 317–332.

Wineburg, S. (1998). Reading Abraham Lincoln: An expert/expert study in the interpretation of historical texts. *Cognitive Science*, *22*, 319–346.

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

Figure Captions

Fig. 1. The scenarios make use of recordings made by three cameras simultaneously making available sufficient detailed information to be used in performance ratings: (a) a zoomed-in, close-up view of the captain's instrument panels, (b) wide-angle view over the shoulders of captain and first officer, and (c) wide-angle view of the first officer's side of the aircraft.

Fig. 2. Ratings on six performance categories—situational awareness (SA), decision-making (DM), tolerances (TOL), knowledge (KN), management (MN), and communication (COM)—by differently ranked pilot pairs. Three ratings in the light grey area or one rating in the dark grey constitute a fail according to company policy. a. Flight examiners. b. Captains. c. First officers.

Fig. 3. A flight examiner uses a series of gestures while accounting for the deviation between flight director (an instrument governed by the automatic pilot that tells the pilot in which direction to fly) and attitude during the early part of flying the missed approach procedure. The examiner gestures the relation between attitude and flight director bars as the captain would have seen them on the instrument panel (a). He then shows how a pilot in the situation would hold the control column (b) and, in response to the flight director bars, would have it forward (c) thereby taking the aircraft downward. As a result, the resulting trajectory of the aircraft would make it crash into the ground (d). The examiner returns to showing how the flight director bars direct the pilot to pull the aircraft in the wrong direction (e) while rolling aircraft to the left (f) to engage the final turn that would have aligned the aircraft with the runway.

Figure
Click here to download high resolution image

Figure

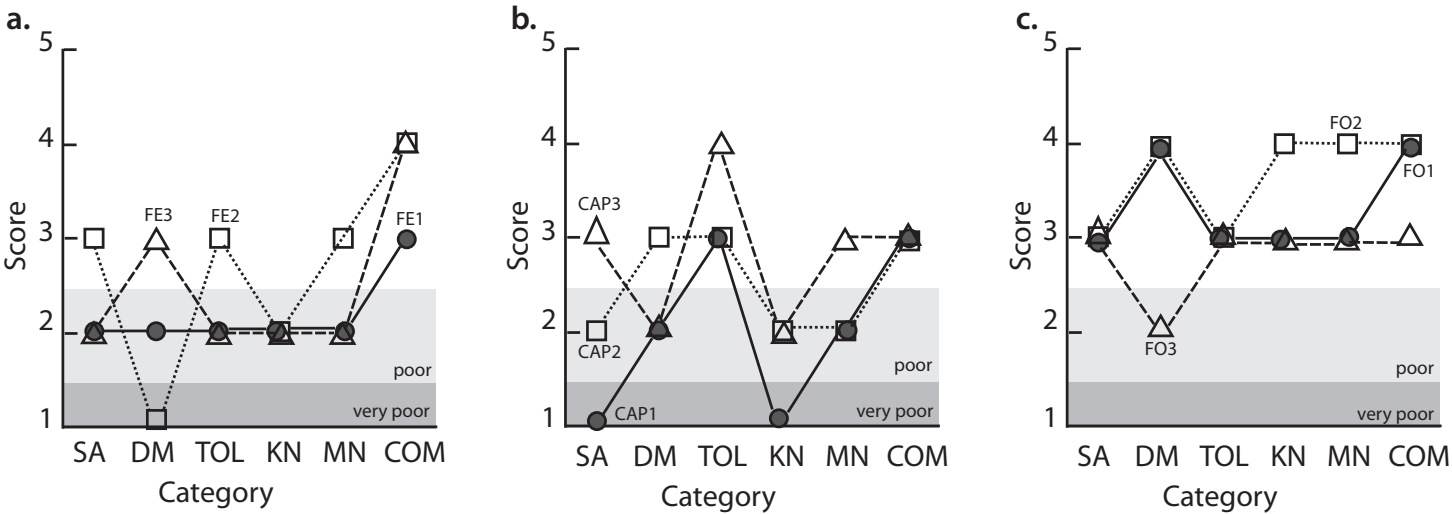
a.
b.
c.
Score
Category
SA
DM
TOL
KN
MN
COM
poor
very poor
FE1
FE2
FE3
CAP1
CAP2
CAP3
FO1
FO2
FO3

Figure
Click here to download high resolution image

Table

Table 1. Background information on participant pilots (means, standard deviation)

Rank	Age	Flight Hours (hrs)	Years as Commercial Pilot	Years as Training Captain	Years as Examiner
Flight examiner	49.2 (6.2)	14,250 (4,910)	25.3 (5.9)	4.9 (3.1)	8.5 (2.9)
Captain[1]	45.3 (6.7)	15,420 (6110)	24.7 (7.0)	0.6 (0.9)	0.0 (0.0)
First officer[2]	30.5 (3.3)	4,900 (1,960)	11.3 (3.9)	0.3 (0.8)	0.1 (0.0)

[1]Two captains had served as training captains
[2]One first officer had 2 years as training captain and 6.5 years as flight examiner while serving as a fighter pilot

Table

Table 2. Evidence in support of the assessment of knowledge/procedure category (KN)

	Failed					Passed			
Facts	**FE1** (2)[1]	**FE2** (2)	**FE3** (2)	**CAP1** (1)	**CAP2** (2)	**CAP3** (2)	**FO1** (3)	**FO2** (4)	**FO3** (3)
Go-around / Missed approach procedure (general, e.g., confusion, lack of familiarity, minor errors)		√	√		√	√	√		√
• Procedure slow		√						√	
• Missing go-around call				√				√	
• Fails to set go-around button	√	√		√				√	
• Going left or right	√		√	√				√	√
• Fails to cancel A/P alarm (warning light)			√			√			
• Fails to recall debriefing			√						
• Missing Flap 15 call				√		√		√	
• Checks late								√	
• Wind				√					
• Hills, mountains				√					√
Approach									
• Forgotten MDA				√			√	√	
• Profile high					√		√		
• Speed, too fast		√						√	
Positive facts									
• Go around procedure correct									√
• Gear up, check power call								√	
• Acceleration altitude								√	
• Heading, low bank, IAS						√			

Note 1: Assessment scores for the knowledge/procedure category

Table

Table 3. Coding of fuzzy observations used by a flight examiner pair (FE1) to characterize decision-making *and support the rating thereof*

Category	**Protocol Excerpt**	**Coding of Fuzzy Observation**
Option generation	"A little bit dodgy"	.30
	"Not really developing a plan"	.20
	"[Planning] wasn't what it should be"	.25
	"Had to be prompted"	.20
Contingency planning	"No contingency, really"	.20
	"Had difficulty planning"	.20
	"No planning at all"	.00
Overall comments	"[Decision-making] blew out after [entering cloud]"	.10
	"He didn't decide"	.15

Table

Table 4. Actual ratings of decision-making *and most likely ratings based on a fuzzy logic model of assessment (boxes)*

Pair	**Actual Rating**	**Distances**[1]				
		Unsatisfactory	Minimal	Satisfactory	Good	Very Good
FE1	2	0.29	0.29	0.73	1.18	1.51
FE2	1	0.76	0.82	1.33	1.93	2.40
FE3	3	1.39	0.98	0.55	0.56	0.92
CAP1	2	1.08	0.86	1.04	1.51	1.95
CAP2	3	0.86	0.54	0.30	0.54	0.86
CAP3	2	0.67	0.50	0.76	1.19	1.56
FO1	4	1.45	1.09	0.62	0.21	0.41
FO2	4	1.75	1.36	0.85	0.33	0.32
FO3	2	1.07	0.88	1.10	1.59	2.03

[1] Grey shading indicates agreement