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.

Assessment of Non-technical Skills:

From Measurement to Categorization Modeled by Fuzzy Logic[1]

Wolff-Michael Roth

University of Victoria

Timothy J Mavin

Griffith University

Wolff-Michael Roth, Faculty of Education, University of Victoria, Victoria & Griffith Institute of Educational Research, Griffith University; Timothy J. Mavin, School of Biomolecular and Physical Sciences, Griffith University.

This research was funded, in part, by a grant from the Griffith University Industry Collaborative Scheme awarded to Wolff-Michael Roth (PI).

Correspondence concerning this article should be addressed to Wolff-Michael Roth, Applied Cognitive Science, Faculty of Education, MacLaurin Building A557, University of Victoria, Victoria, BC, V8W 3N4, Canada

E-mail: wolffmichael.roth@gmail.com (for correspondence with journal); mroth@uvic.ca (for publication)

[1] To appear in *Aviation Psychology and Applied Human Factors* DOI: 10.1027/2192-0923/a000045

Assessment of Non-technical Skills:
From Measurement to Categorization Modeled by Fuzzy Logic

Abstract

The assessment of pilots' performances and more specifically their non-technical skills turns out to be difficult because researchers and practitioners find that inter-rater reliability tends to be moderate to low. In this study, we propose, consistent with field observations, that assessment is a categorization rather than a measurement issue. Categorization, even though and especially when based on imprecise information and assessments, can be modeled mathematically using fuzzy logic. We present the structure of the approach and provide an example from a large database of pilots providing reasons for their assessments of other pilots' (simulator) flying performances. As implications, we discuss possibilities for using the variations in performance ratings positively, as a source for learning and resilience building.

Key words: assessment; measurement; inter-rater reliability; categorization; fuzzy logic

Assessment of Non-technical Skills: From Measurement to Categorization Modeled by Fuzzy Logic

Following initial predominance of technical skills in the assessment of pilot performances, the aviation industry has increasingly shifted to implementing programs to enhance so-called non-technical skills—e.g., situational awareness, management, and decision-making—because these have shown to be at the root cause of many accidents (Helmreich, Musson, & Sexton, 2004). Thus, for example, in Europe the NOTECHS (non-technical skills) system has been developed for assessing pilots' crew resource management skills (Flin et al., 2003). Non-technical skills training and assessment is tied to interrater reliability (IRR). Thus, for example, "only where organizations have developed good levels of standardization and inter-rater reliability does CASA support a move towards jeopardy based non-technical skills training" (Civil Aviation Safety Authority, 2011, p. 23). IRR is an important component to the drive for global standards in the aviation industry (ICAO, 2007). However, assessing these skills rigorously and consistently has shown to be more difficult, as shown in the problems of achieving high inter-rater reliability scores when multiple raters assess the same performance—though some authors appear to be satisfied with IRR scores of about .75 (e.g., Sharma, Orzech, Boet, & Grantcharov, 2013). In part, raters differ significantly not only between but also within levels of experience in terms of the severity with which they assess certain aspect of pilot performance (e.g., Mulqueen, Baker, & Dismukes, 2002). As a result, industry efforts in improving safety through non-technical skill (CRM) training have reached a performance plateau (Amalberti, 2001). Some authors go so far as to suggest the abandonment of (some) non-technical skills such as situational awareness because these correspond to folk models rather than to well-defined scientific concepts (Dekker & Hollnagel, 2004). There are even suggestions to remove the name *non-technical skills* with new approaches to resilience engineering (e.g., Saurin, Wachs & Henriqson, 2013). However, abandoning non-technical

skills wholesale—let alone the difficulty associated with changing twenty years of aviation language—may be throwing out the baby with the bathwater given recent ethnographic research that shows how useful practitioners find these concepts for talking about their everyday work generally and problems therein specifically (Mavin, Roth, & Dekker, 2013). That study, however, also shows the converse side of the problem: there is a rather low degree of consistency between raters, whether these work individually or in pairs. The lack of consistency reported was to the point that in some instances, the ratings spread from a 1 to a 5 on a 5-point scale. What other ways are there for understanding the spread of the performance ratings without abandoning the currently existing approaches to non-technical skills, which are of importance and usefulness to practitioners and researchers alike?

The present study offers a rigorous alternative. Based on our ethnographic work within the industry and on our modified think-aloud protocol with a small number of pilots of different rank (flight examiner, captain, first officer), we suggest that the issue of assessment is not a measurement problem but a categorization issue. This work, briefly described below, shows that the raters assign specific aspects of videotaped scenarios to categories of performance rather than doing the equivalent of measuring a physical object. Categorization that uses approximate descriptions (e.g., poor, satisfactory, and very good) can be meticulously modeled—as shown in mathematical approaches to medical diagnosis (e.g., Klir & Yuan, 1995; Innocent, John & Garibaldi, 2004) and medical expert systems (Adlassnig, 1986; Phuong, 1995)—using fuzzy logic. We thereby follow others, though using a differing approach, in choosing fuzzy logic to address the weaknesses of the traditional approach to inter-rater reliability (e.g., Ciavolino, Salvatore, & Calgagnì, 2013) and issues in personnel assessment (e.g., Capaldo & Zollo, 2001). We exemplify the approach with a case from our research in which pilots give their reasons for assessing the performances of peers videotaped in flight simulator scenarios, using as our example those parts of the assessment sessions that focused on situational awareness. The approach is

highly promising, as it affords understanding why and on what grounds different assessors arrive at quite different assessments of the same video clip. We conclude—differing here considerably from authors such as Dekker and Hollnagel (2004)—that there is usefulness in non-technical performance descriptions such as "situational awareness." Moreover, we have in our hands a rigorous approach for predicting how raters will assess a situation given how they understand specific parts of the assessment object. In this case, rather than abandoning existing the human factors related language used in the industry, we should be thinking about how to use the observed variations positively to (continuously) improve pilot performances.

Background

Given the stakes involved in ascertaining the quality of pilot performance, it is not surprising that researchers and practitioners alike are interested guaranteeing the reliability and validity of regular and continued evaluations (Baker & Dismukes, 2002). However, there is some evidence that inter-rater reliability of the non-technical crew resource management items (e.g., situational awareness, decision making) may be difficult to achieve (Brannick, Prince, & Salas, 2002; Smith, Niemczyk, & McCurry, 2008), a finding that is consistent with more recent experiences of assessing non-technical skills in other fields such as medicine (e.g., Sharma et al., 2013). Even extensive, multiyear training efforts in improving inter-rater reliability have resulted in congruency of rating distributions, rater agreements, and consistency that are "generally acceptable" (Holt, Hansberger, & Boehm-Davis, 2002). The authors qualify the results of their study "encouraging but not definitive" (p. 328). The study also encourages investigations of inter-rater reliability benchmarks that may be achievable through training. It does not however suggest an alternative way of thinking about assessment, which would deal with the fact that examiners often have good reasons for varying in their assessments, as some recent work appears to suggest (Mavin et al., 2013).

The calculation of inter-rater reliability is based on frequentist statistics, which, as other frequentist statistics (e.g., correlations and analysis of variance tests), assumes that the assessments are distributed (usually Gaussian, sometimes bi-modal) and variations are explicitly treated as unexplained or as (rating) error variance.[1] Thus, reliability is defined as

$$\text{Re}\,liability = \frac{Var(true)}{Var(observed)} = \frac{Var(true)}{Var(true) + Var(error)}$$

(Hallgren, 2012). A problem of getting at inter-rater reliability arises from the fact that many early efforts at modeling assessment failed to differentiate between fuzziness and randomness of error variance (Esogbue & Elder, 1979). Thus, there tend to be good, albeit fuzzy reasons for the varying ways in which humans order their world and in how they account for its orderly nature (Garfinkel, 1988). A different approach to rater variance would be to consider categorization generally and assessment specifically as rational endeavors, even when the "inputs" into the decision-making process are inherently characterized by uncertainty and requiring local knowledge to be understood (Garfinkel, 1967). This is so because every present situation is inherently vague, a state that remains "invariant to the clarification furnished by [further] exchanges of questions and answers" (p. 92). In fact, even the most rigorously designed survey instruments contain categorizations, which weaken any assurances for achieving reliability and, therefore, the external validity of quantitative inference (Suchman & Jordan, 1990). Categorization during the assessment of flight deck performances was reported when flight examiners and other pilots (captains, first officers) tended to use major (e.g., situational awareness, communication, or management) and minor categories (e.g., perception, comprehension, projection) to make sense of flight-deck performances (e.g., Mavin & Dall'Alba, 2010; Mavin et al., 2013). Thus, for example, two flight examiners watching a scenario videotaped on a flight simulator talk about what they have seen in this way:

L: If we go back to the management of it [flight]. It started off quite well[1], I thought. But then it started to snowball a bit[2], near the end. But communication I thought was clear all the way through[3].

R: Yes.

L: I mean, they were talking, but what was coming out wasn't exactly what should be coming out[4].

R: No, it was sort of more "here it is," isn't it?

In this excerpt from their 40-minute assessment session, inherently imprecise and uncertain qualifications are made. For example, the flight is said to have "started off *quite well*" [1] where even what is meant by "the start off point" remains unarticulated. The performance then is described as "snowballing a bit" [2]. In the same way, communication is said to have been "clear all the way through" [3] or qualified by saying that "what was coming out wasn't exactly what should be coming out" [4]. Throughout our database of over 500 pages of transcription of similar conversations about performance, the language used is of this type rather than mathematically exact. We may therefore apply to assessment in aviation what Suchman and Jordan (1990) suggested for the quantitative information in survey situations: "In contrast to a thermometer or other instrument of measurement, an interview, no matter how standardized, remains *fundamentally a linguistic and interactional event*" (p. 240, emphasis added). Here we propose viewing pilot assessment as an interactional linguistic event—e.g., between flight examiner and pilot assessed; assessment reasons—and that it should be modeled as such.

To date, we have observed different situations—classroom-based assessments of videotapes; assessments of videotaped scenarios done by pairs; assessment during debriefing meetings with and without a debriefing tool; and self-assessment of pilots using a debriefing tool—where assessment first and foremost is a categorization effort: It describes performances as good, acceptable, "repeat," or "pretty good, pretty good, straight fives" (e.g., Mavin & Dall'Alba, 2010; Mavin et al., 2013). Ordering the world according to

categories, in fact, is fundamentally human and pervades our everyday lives (e.g., Lakoff, 1987; Rosch, 1978). However, even the most experienced and accomplished experts may be confronted with specimens that resist all efforts at classification, even when the efforts are collective (Bowker & Star, 1999). A pertinent expert think-aloud study in the medical field shows that although experienced family practitioners were much faster than advanced medical students in accessing their memory relevant information and in judging biomedical items related to previously presented brief case descriptions of cases, the former were not more accurate in their diagnoses than the latter (Rikers, Schmidt, & Moulaert, 2005). This result, in fact, replicates some of the findings in an aviation-related study (O'Connor et al., 2002). Variations arise because classification, as non-technical skills in aviation (Saurin et al., 2013), turns out to be physically and temporally situated and socially distributed. Even when highly trained scientific experts collaborate neither uncertainty nor inconsistency is eliminated. Instead, contradictions tend to be minimized rather than removed. Thus, "categories emerge in and through temporalized narratives, and through the body's gestures, making the practices highly local and difficult to communicate" (Roth, 2005, p. 611). Despite this situation, studies have shown that classification can be rigorously modeled using (a) fuzzy logic, which models the reasoning methods underlying approximate inference construction, and (b) fuzzy sets that define inexact entities (Adlassnig, 1986). In the following, we show how we can model pilot assessment rigorously (mathematically), even and especially when based on imprecise, inherently underdetermined language-based descriptions.

Case Materials

This study was designed to propose an alternative approach to understanding assessment of pilot performance. Rather than viewing assessment as a measurement issue, where variation is treated as *unexplained* (error) variance, it is proposed to view it as a categorization issue, where variation is the result of imprecise and inexact entities and

reasoning. Using fuzzy logic, we can model variation mathematically, that is, understand and explain variation.

Background of the Data

The case materials presented here were collected in the context of a study that used a modified expert think-aloud protocol while pairs of pilots at different rank—flight examiner, captain, and first officers—assessed the performances of pilots shown in videotaped scenarios of 5–10-minute length. The sessions took place in the participating pilots' company headquarter-based training facility (see Mavin et al., 2013). In a normal think-aloud protocol, experts would be asked individually to solve a generally domain-related task; and, while solving the task, they would be asked to articulate all their thinking for the researcher and camera (Ericsson & Simon, 1993). When, however, two individuals solve tasks, they have to do together what one person could do on his/her own (Suchman, 2007). In this case, the collaborators articulate for one another their sense of what is going on, what a current problem is, why they do something, what remains to be done, and so forth. This information is then available to researchers as well. The transcription of such interaction "is a *naturally generated* protocol" (p. 123, emphasis added).

The pilots analyzed the scenarios according to the six dimensions that their company uses in its formal assessment process. These dimensions include two technical skills (flying the aircraft within tolerances, knowledge and procedures) and four non-technical skills (situational awareness, decision-making, management, communication). Also available are the scores that each pair attributed to the pilots in the scenarios on the six dimensions and the recommendations as to whether the pilot in the scenario should pass or fail.

The participating pilots had from 2,100 to 24,300 flight hours, ranging from 6 to 33 years as commercial pilots. The captains and flight examiners averaged about 25 years of experience and a total of 15,000 flight hours, whereas the first officers averaged 11 years as commercial pilots with a total of around 5,000 flight hours. Three cameras were used to

record the sessions, in which pilots used the grid to assess each pilot (captain and first officer) flying in a variety of scenarios.

In the scenario featured in the case materials below, the aircraft was approaching one of the regular airports served by the company; all pilots were familiar with the approach, which, in the scenario, would be a downwind circling approach. The aircraft descends below the cloud base. The actual glide path is consistently 60 feet above the indicated approach trajectory at the 12-, 9-, and 8-mile way points, and 140 feet above at the 5-mile point. Throughout, the speedometer needle is 20 knots above the white speed bug where it should have been. As the aircraft passes the runway, the captain (flying pilot) announces poor downwind weather conditions around the final approach. A little later, just as he takes the aircraft into the final turn, the aircraft flies into a rain shower. The captain announces "missed approach," where there is a quick discussion about turning right (captain) versus turning left (first officer) occurs. Following the discussion the captain maneuvers the aircraft into a correct left-turn for the missed approach. There is general agreement among the rating pilot pairs that the missed approach procedure was "a bit messy."

Assessments and Reasons

The data show that there is considerable variation in the assessment of the pilots, which often range over 3 and up to 5 points on the five-point scale (Mavin et al., 2013). In the particular scenario discussed here, the consistency of the agreements over the six performance categories reaches significance in 4 (out of 36) pairwise rater-pair comparisons ($.79 < r < .93$).[2] But these consistencies did not hold for the other pilots or across scenarios (e.g., only 1 significant correlation in the assessment of the first officer, for two different pairs than for the captain). On any performance level, and depending on the scenario assessed, the ratings included a considerable range of scores from 1 or 2, on the one hand, to 4 and 5, on the other hand. In the end, 5 rater pairs failed the captain based on company criteria—one 1 or three 2s. Asked for a global appreciation of the results of the ratings and the failed assessment, all rater pairs felt satisfied with their assessment results.

The variations are especially interesting given the fact that these were produced by consensus in pairs of raters (a practice some airlines use to reduce variability), which would have a regressive effect given that fewer important facts were missed and more extreme evaluations would be less likely. To understand these variations in the ratings, the transcripts of the assessment sessions were analyzed to reveal the reasons that each pair provided to ground its assessment decision. In the next section, we use the facts and reasons as the input for the modeling effort.

In this study, we focus on situational awareness in an exemplary fashion. The scores of the nine pairs for the flying pilot (captain) in the scenario described were $1 \leq SA \leq 3$, with a mean of $SA = 2.44$ ($SD = 0.73$) and a mode of 3. We extracted from the transcriptions all comments with respect to specific contexts of the flight generally linked to the discussions of situational awareness. The analysis shows that there were three contexts that were used to support statements about the captain's (flying pilot's) situational awareness: (a) the current weather situation in the area where the aircraft was to make its final turn to land during the circling approach; (b) the proposed or actual turn direction taken by the captain immediately following the go around call; and (c) flight parameters on the glide path during the approach prior to reaching minimum descend altitude.

First, while flying at minimum descent altitude and about 1 minute after having become visual with the runway, the two pilots in the scenario are talking about the weather conditions, the wind speed and direction at the level of the aircraft and on the ground. The captain (flying pilot), looking out the window to his left notices: "the rest doesn't look too crash-hot so um out to the East there, we'll have to be careful?" The first officer acknowledges: "Yes." Just as the captain is turning the aircraft onto its final turn, 1:15 minutes after having talked about the weather in this part of the glide path, the aircraft loses visual reference to the runway when they fly through a rain shower (i.e., instrument meteorological condition, IMC). This leads the captain to initiate the missed approach procedure. The captain's actions and reactions to the weather situations supplied the raters

with the grounds for making an assessment of his situational awareness. The weather was an important aspect of the assessment discussions that mediated how the pairs assessed the captain's situational awareness (and decision-making). The comments concerning the weather show, however, that the same perceived facts led to different assessments as to the level of situational awareness (Table 1). All but one rater pair noticed that the captain had perceived the poor downwind weather situation in the area where the final turn would be effectuated. These pairs also tended to note—to different degrees—that the captain may have had difficulties comprehending the weather situation, translating the observation into actions "inside the cockpit." Especially, the captain failed to predict future events, that is, make a link between the observation of bad weather and the possibility of a missed approach. Given that situational awareness involves (a) perception of facts, (b) the comprehension thereof, and (c) the resulting prediction, it is perhaps not surprising that the assessments ranged from unsatisfactory (1) to satisfactory (4) (Table 1).

Second, assessors found evidence in the video that the captain had initially initiated a slight right turn, "rolling the wrong way" (G1, G7); one pair suggested that the captain "was initially going to turn the wrong way" (G1). Pairs described what was happening as pilot *intention* without actually stating that the aircraft had started rolling to the right: "going to turn the wrong way [right]" (G2, G5) and "want[ing] to go right" (G7, G9). Finally, some assessor pairs noted that there was "[a bit of] confusion" and that "rather than just do the wrong things," the captain *asked/confirmed* whether the turn was right before actually turning the wrong way (G3, G4, G6, G8). Thus, assessors suggested: "the confusion came as to whether it was a left or right turn" (G6) or "the confusion of whether it was a left or right hand" (G4); and one of the flight examiners made a hand gesture that that showed the aircraft level prior to doing the required left-turn following the go around call. In general, those pairs who saw evidence in the video that (a) the aircraft was actually rolling to the right or (b) the captain wanted to go right before the first officer "corrected him" tended to weigh this heavily in their assessment of situational awareness (downward). Assessor pair

G1, who did not comment on the weather at all, suggested that the captain appeared to "track out the radial [in] the opposite sense" and therefore engaged in a right rather than left turn.

Third, the assessors generally noted that the captain appeared to be unaware of the fact that the aircraft was above the glide path specified on the approach plate. The raters suggested that the deviations were between 60 feet and 200 feet higher and saw that the speed was generally 20 knots above reference speed. Some rater pairs also noted "big fluctuations" (G1). One group suggested that the captain failed to appreciate the wind from behind and the effect that this wind had on the glide path of the aircraft (G6.) One pair (G2) also stated that the captain was unaware of the minimum descent altitude. The transcript of the scenario provides evidence that the captain asked, "What's the minima again?" The pair G2 considered the minimum descent altitude as one of two things that a pilot absolutely should be aware of (the other being the turn direction on missed approach).

Finally, to a lesser extent, other reasons were used to ground assessments. For example, two assessor pairs (G2, G9) noted that the circling approach was dangerously close to mountains and that there had been a crash with loss of life in the past. The captain *should have been* aware of the danger and should have been prepared to fly the correct missed approach procedure given the weather conditions.

The preceding overview of the reasons provided for rating situational awareness of the captain shows that the raters mainly drew on three aspects of the scenario: (a) the way in which the captain dealt with the observed weather situation; (b) what the captain did following the initiation of the missed approach; and (c) the captain's awareness of the glide path (vertical profile during descent, speed, minimum descent altitude). The overview also shows that the pairs did not observe precisely the same facts, for example, in the case of the turn (to be confirmed, intended, actually enacted). Given the different facts noted and the different degrees to which observed facts entered the assessments, it may not surprise that the pairs came to different conclusions about how to rate the level of the captain's

situational awareness: from 1, an "immediate fail" in the company's performance language, to a 3, the "performance of an average line crew." Finally, the overview shows that the facts and reasons were stated in imprecise mundane language. But this did not prevent the assessor pairs to assign scores to the performances they rated. In the following, we show how ratings can be modeled and thus explained despite the imprecise language.

Modeling Assessment of Pilot Performance: A Fuzzy Logic Approach

In the preceding section, we show that raters observe different facts and weight the observed facts differently. In this study, it is therefore proposed to view assessment of pilot performance—at least when conducted in those several airlines in our research using this or similar assessment grids—as a categorization issue where multiple, *inherently vaguely defined and ambiguous categories and terms* (e.g., Suchman & Jordan, 1990) are in play. That is, the purpose here is to provide a way of arriving at a categorization of performance given the facts and reasons as well as their weights that enter the assessment. The approach is premised on the understanding that flight examiners and others in pilot performance assessment *have good albeit imprecise reasons* for grounding their assessment-related decisions.

Modeling the Assessment of Situational Awareness

Here we propose modeling assessment in the same way that medical diagnosis has been modeled when alternate diseases are possible given the symptoms observed: using a fuzzy logic approach (e.g., Esogbue & Elder, 1979, 1980, 1983). The fundamental idea is to model how the everyday, uncertain, imprecise, and underdetermined descriptions assessors (initially) make lead to specific categories of performance that they might express with a score from 1 to 5. Rather than using a score, the 5 categories of performance we use are *unsatisfactory, minimum standard, satisfactory, good*, and *very good*. We compare the results of the modeling with the scores (on a five-point scale) that assessors

assigned in our study. We model assessment drawing on the method of fuzzy cluster analysis.

In their assessment, the raters attribute an observed performance to one of five possible performance categories. Each performance category is described by a matrix in which the upper and lower bounds of the normal range of the performance components form one dimension and the severities of the components form the other dimension. These boundaries represent a shared value system such as it might be specified in a performance-describing document. We begin by defining a set of lower and upper boundaries of performance in terms of two *fuzzy sets* $\mathbf{B}_l$ and $\mathbf{B}_u$, associated with the five stated categories of performance (Esogbue & Elder, 1980). This in fact makes an association between a continuum and a set of categories, much as it has been proposed for turning World Health Organization categories of function, disability, and health into a continuum (Bharadwaj, 2007).

$$\mathbf{B}_l = \begin{matrix} W & T & GP \end{matrix} \begin{bmatrix} 0 & 0 & 0 \\ .2 & .2 & .2 \\ .4 & .4 & .3 \\ .7 & .7 & .8 \\ .9 & .9 & .8 \end{bmatrix}, \quad \mathbf{B}_u = \begin{bmatrix} .2 & .2 & .2 \\ .4 & .4 & .3 \\ .7 & .7 & .8 \\ .9 & .9 & .8 \\ 1 & 1 & 1 \end{bmatrix} \begin{matrix} \text{unsatisfactory} \\ \text{minimum} \\ \text{satisfactory} \\ \text{good} \\ \text{very good} \end{matrix} . \tag{1}$$

Each matrix **B** defines a fuzzy set, where, in the example used here, the first columns define the lower (l) and upper boundary (u) of the weather-related performance component for a given level of performance ($\mathbf{B}_i$). The lowest lower boundary values are set to $\mathbf{B}_{l,1,j} = 0$, which indicates "complete failure"; the uppermost upper boundary values are set to $\mathbf{B}_{u,5,j} = 1$, which corresponds to a "flawless performance." That is, the two fuzzy sets allow a translation from language-based formulation (e.g., "unsatisfactory") to a range of values (i.e., $0 \leq \mathbf{B}_i \leq 0.2$).

We also define a *fuzzy relation* **W**, which specifies the pertinence of a specific behavior in the assessment of the matrix of each component. This matrix, too, represents a norm,

which we specified based on the general observations of the importance of weather, turn, and glide path to the scenario. The weight matrix might be given by

$$\mathbf{W} = \begin{array}{c} \begin{matrix} u & ms & s & g & vg \end{matrix} \\ \left[\begin{matrix} .9 & .9 & .8 & .9 & 1 \\ 1 & .9 & .9 & .9 & 1 \\ 0 & .2 & .5 & .8 & 1 \end{matrix}\right] \end{array} \begin{matrix} W \\ T \\ GP \end{matrix} \tag{2}$$

where each number signifies the weight of a given performance component {*W, T, GP*} for a particular level of performance {*u, ms, s, g, vg*}. Thus, for example, for an unsatisfactory (u) rating, how the rated pilot dealt with the weather has a weight of .9, how he dealt with the turn has a weight of 1, and his awareness of the glide path parameters do no figure at all (weight = 0). For receiving a perfect score on situational awareness, all aspects would have to be perfect (i.e., *W, T, GP* = 1).

Let us now assume that a particular rater pair *k* perceives pilot *x* display a range of performance components. Each consists of a specific behavior ("symptom") *W* (appreciation of downwind weather situation), *T* (turn following the missed approach call), and *GP* (glide path parameters) with a particular performance level $\mathbf{A}_{k,x}$ = [*W, T, GP*]. This *fuzzy set* encodes the degree to which the assessor appreciates each context by translating fuzzy categories into mathematical values. To keep our description from becoming too abstract, we exemplify the approach by modeling the assessment that the pair G7 (two first officers) produced. Rater pair G7 noted that the flying pilot perceived the weather but questioned whether he actually understood it. The pair noted that the pilot *probably* had *some* difficulty predicting what could happen given the weather situation and suggested that a pilot "got to be careful of [the weather] and see it coming." The pair noted that he did identify cloud and risk but still ended up entering the cloud and going IMC. The pair agreed, however, that "although there were a few errors," the pilot was "still quite good overall" and only his "comprehension a little bit down"; for this reason, they "must not be too harsh." The weather variable was therefore set to *W* = .5 (i.e., satisfactory, see eq. 1). The

assessors did note that there "was a brief turn to the right," which was mitigated by the fact that "they corrected it." Although the action was not good, there were mitigating factors: the turn parameter was therefore set to $T = .4$ (at the upper end of minimum standard and lower end of satisfactory (see eq. 1), which reflects the fact that the turn was wrong but corrected. Finally, the pair suggested that the glide path parameters were "pretty reasonable," leading us to set the parameter to $GP = .75$, which is at the upper boundary of satisfactory and almost good (eq. 1). The fuzzy set of observations used by G7 therefore is

$$\mathbf{A}_{G7,x} = \begin{matrix} W & T & GP \\ \left[\; .5 \quad .4 \quad .75 \; \right] \end{matrix}. \tag{3}$$

To model the assessment, a clustering technique is used. It determines that assessment category with which the assessment is most similar. We calculate the distance between the raters' fuzzy set of observations and the performance categories on each of the observed context (i.e., the diagnostic cluster specified in eq. 3) (Klir & Yuan, 1995). Thus, for example, for the unsatisfactory category given by assessor pair k, we get

$$\mathrm{D}_{k,uns} = \left[\sum_{j=1}^{3} [\mathbf{W}_{uns,j}(\mathbf{B}_{j,l} - \mathbf{A}_{k,j})]^2 + \sum_{j=1}^{3} [\mathbf{W}_{uns,j}(\mathbf{B}_{j,l} - \mathbf{A}_{k,j})]^2 \right]^{1/2} \tag{4}$$

where, we use a Euclidean metric commonly employed in clustering techniques (Esogbue & Elder, 1980).[3] Concretely, then, in the case of G7 we get

$$\mathrm{D}_{G7,\mathrm{uns}} = \left[\left[.9(0-.5)^2 + 1(0-.4)^2 + 0(0-.75)^2 \right] + .9(.2-.5)^2 + 1(.2-.4)^2 + 0(.2-.75)^2 \right]^{1/2}$$

as distance between actual observation and performance specifications for an unsatisfactory grade. This yields $\mathrm{D}_{G7,\mathrm{uns}} = 0.784$. Doing the same for the satisfactory category

$$\mathrm{D}_{G7,\mathrm{s}} = \left[\left[.8(.4-.5)^2 + .9(.4-.4)^2 + .5(.3-.75)^2 \right] + .8(.7-.5)^2 + .9(.7-.4)^2 + .5(.7-.75)^2 \right]^{1/2}$$

yields $\mathrm{D}_{G7,\mathrm{s}} = 0.352$. A similar calculation for the minimum standard category yields $\mathrm{D}_{G7,\mathrm{m}} = 0.447$. Thus, because $\mathrm{D}_{G7,\mathrm{s}} < \mathrm{D}_{G7,\mathrm{m}} < \mathrm{D}_{G7,\mathrm{uns}}$, pair G7 would more likely rate the pilot performance as satisfactory than as minimum standard or unsatisfactory. Moreover,

calculating the same for the *good* category yields $D_{G7,g} = 0.572$. Again, because for G7 $D_{G7,m} < D_{G7,g}$, we would anticipate its rating more likely to be satisfactory than good. On their company's rating sheet, this pair of assessors provided a score of 3 (on the five-point rating from 1 to 5) for the flying pilot's (captain's) situational awareness, which corresponds to the satisfactory rating that our fuzzy logic model arrives at.

Assessor Group 2 (captains) was quite adamant about the glide path parameters and the turn: the assessors emphasized that the "two things a pilot should be aware of" are (a) minimum descend altitude that the aircraft was descending to (i.e., it was to remain set during the circling approach) and (b) the way in which to turn, especially given the high terrain to the right of the aircraft. These two issues were characterized as "two sort of gaping holes in situational awareness." However, the pair did note that the captain "was going to turn right" rather than that he was actually turning. Glide path parameter and turn are therefore are set at $T = .2$ and $GP = .1$ in this illustration. The pair also listed weather as another issue that the flying pilot was not aware off, especially in anticipating the possibility of a go around. The weather parameter therefore is set, in this modeling example, at $W = .1$. Conducting the same calculations as before for the unsatisfactory, minimum standard, and satisfactory categories, we obtain $D_{G2,uns} = 0.237$, $D_{G2,m} = 0.350$, and $D_{G2,s} = 0.810$. In the case of assessor Group 2, therefore, $D_{G2,uns} < D_{G2,m} < D_{G2,s}$. Thus, the unsatisfactory category would be the most likely one that this group would be opting for. This corresponds well to the pair's overall assessment that situational awareness "was barely acceptable for a simulator check situation." This assessor pair actually provided a score of 1 on the rating scale from 1 to 5, which corresponds to our "unsatisfactory" rating.

Extensions of the Model

A case such as the weather and its implications for the flight safety, here set by holistically choosing a parameter value, can be modeled as a composite using a fuzzy logic approach. Thus, the question would be whether the raters focus on the perception of the weather in the area of the final turn, its comprehension, and associated projections for the

flight. Each of these dimensions is associated with fuzzy descriptions such as low, acceptable, and high; these values then would be translated into parameter values such as .3, .6, and .9. Similarly, the awareness of glide path parameters might be a composite of different components, each of which associated with fuzzy evaluations such as unacceptable, acceptable, and accurate. The turn parameter is the result of assessor perceptions that there was a confirmation of a left turn (medium, medium-high awareness), an intention to turn (minimum standard), and an actual turn (unacceptable). In the other direction towards integration of all technical and non-technical skills into a summary of a pass/fail decision, we can extend the present model so that it integrates the totality of fuzzy observations into one overall assessment. In our study, the company policy was to fail a pilot who had one score = 1 or three scores = 2.

The fuzzy logic approach can be extended to situations when assessors (a) use a reason (component) that is little or not used by other raters or (b) do not use a particular dimension in their assessment even when most or all other assessors do use it (Esogbue & Elder, 1979). In fact, Esogbue and Elder (1983) include a large number of fuzzy information pieces that enter a medical diagnosis. Here we present an example of the second case. For example, in our modified think-aloud study, G1 (flight examiners) did not address the weather situation at all. At no point in its 40-minute session assessing the performance of the pilots in the flight scenario did they comment on the conversation related to the weather in the circling area or about the wind in the air and on the ground (Table 1).[4] The pair agreed that the pilot was "tracking out the radial in the opposite sense," which ended in a situation where "indications were that he was turning right" and the "turn was incorrect." The pair concluded that the pilot "was lost," especially "in the latter part, anyway." We therefore may set the turn parameter to $T = .1$. With respect to the glide path, the pair repeatedly noted that the pilot was unaware of the deviations both in terms of height and speed (e.g., "not once did he mention the variations from where he was to where he should be and it was about roughly 200 feet high most the way in"). The rater pair also

noted that the pilot "confused" the current circling and the regular "straight in" approaches when selecting the missed approach altitude on the ADU (advisory display unit) when the minimum descend altitude should have remained selected; and when the missed approach altitude was to be selected, the pilot had to be prompted to do so. However, this problem was mitigated by the fact that "it was resolved straight away" when the pilot noted the error. We therefore set *GP* = .25; a further mitigating factor was that the pilot's "overall performance was fine," it was just at the end that problems appeared ("later in the piece, things sort of started shrinking for him. (Without the mitigating factors, it would have been set to something like *GP* = 0 or 0.1.) Using the same limit and weight matrices as before, we obtain $D_{G1,uns} = 0.255$, $D_{G1,m} = 0.171$, and $D_{G1,s} = 0.561$. The predicted category for this assessor pair would be "minimum standard." This corresponds to the rating of 2 that G1 actually provided. The predicted category ("minimum standard") is also consistent with the actions these examiners said would take: ask the pilot to "try it again [now]" so "you can see improvement," that is, that this was a one-off rather than a consistent problem of the pilot.

Discussion

The reliable and valid assessment of pilot performance, including crew resource management (non-technical) skills, despite good efforts to address variation through training, continues to be a difficult area in the aviation industry in the face of efforts by the International Civil Aviation Organization to make it an integral aspect of global quality assurance. Traditional approaches to inter-rater reliability are associated with considerable problems (e.g., Ciavolino et al., 2013; Hallgren, 2012). Much of the effort conducted so far has been devoted to the development of methods that decrease variation attributed to rater error. The underlying assumption is that assessment is a measurement issue and decreasing the error through training can reduce related error variance. Few studies have investigated different ways of thinking about assessment: that raters have good, even if somewhat imprecise, language-mediated reasons (heuristics) for making their

assessments. Such an approach acknowledges and builds on the rationality of human endeavors (Garfinkel, 1967). It models these endeavors even in the face of uncertainty, ambiguity, and variations with which assessment objects can be perceived, described, interpreted, or appreciated (Esogbue & Elder, 1983). In this study, we present such an approach that models how the imprecise descriptions provided while assessors talk to and for each other about the performances of pilots lead to specific scores that they attribute to these performances. We show how good agreement can be achieved between the predicted and actual assessment scores. However, the present is a conceptual study with exemplary materials. Future studies are required in which the degree of deviation between predicted and actual categorization are systematically investigated.

In the fuzzy logic approach proposed here, even such ephemeral descriptions as the relation between captain and first officer can be modeled. Thus, for example, in the context of communication in a second scenario, assessors often talked about the degree to which the captain "was leaning on the first officer" to get his point of view accepted, that "there was a certainly some subtle pressure," about a first officer "who stuck to his guns" and "did not succumb to the captain's pressure," and about "the very subtlety thing" of using body language to pressure a first officer all the while saying "I don't want to pressure you." Degrees of "leaning on" could be defined as a fuzzy set and used to model a non-technical skill (human factor) such as communication.

If assessors of pilot performance have good, if inherently imprecise reasons for varying in their assessments—i.e., justifying the variations—then there may be little hope to push up inter-rater reliability as if the phenomenon was arising from error variance. If, on the other hand, we can understand and model these variations as arising from good reasons, then researchers and practitioners (industry) might want to begin thinking about assessment in different ways. For example, we might begin to think using the diverse and divergent descriptions of performance as starting points for conversations that might lead participating pilots and examiners to say things like "What an interesting way to think

about what happened there?" or "I have never thought about doing this in such a situation!" We would then be going from an assessment to a learning approach to maintaining and improving pilot performance levels. This is consistent with a resilience engineering approach, which is based on the assumption that error and minimum breakdown is required to enhance safety (Amalberti, 2001; Saurin et al., 2013). In our ongoing observations of debriefing meetings, where examinees and examiners have access to a debriefing tool (which records the examination session), the latter are in some instances incorrect and learn from the assessed pilots. Thus, an examiner's observations, descriptions, and explications in terms of human factor concepts (e.g., situational awareness, communication, or management) might serve as the starting point for discussing performance for the purpose of triggering learning events. Our research in progress concerning different time allocations to SIM sessions and debriefing suggests that talking about and reflecting on performance may be more beneficial for improving practice than a situation of extensive assessment (e.g., 4-hr SIM sessions) and with little reflection (20–30 minute debriefing).

We begin this study by referring to observations that there is often considerable variance in the way raters assess pilot non-technical skills (CRM) performance. We also refer to the suggestions of some authors that current human factors concepts such as situational awareness should be abandoned. The present study suggests a different approach, for even though there is considerable variance in assessment, it is not due to error but can be explained. The question for researchers and industry alike ought then to be: How do we use this apparently natural variation to improve pilot performance specifically and industry performance more generally? In other words, how can we make the industry more resilient to failure by making use of the language that we already have for talking about performance, including the human factors (CRM) concepts of situational awareness, management, decision-making, or communication?

References

Adlassnig, K.-P. (1986). Fuzzy set theory in medical diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics, 16*, 260–265.

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

Baker, D. P., & Dismukes, R. K. (2002). A framework for understanding crew performance assessment issues. *International Journal of Aviation Psychology, 12*, 205–222.

Bharadwaj, B. (2007). Development of a fuzzy Likert scale for the WHO ICF to include categorical definitions on the basis of a continuum. *ETD Collection for Wayne State University.* (http://digitalcommons.wayne.edu/dissertations/AAI1442894)

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. Cambridge, MA: MIT Press.

Capaldo, G., & Zollo, G. (2001). Applying fuzzy logic to personnel assessment: a case study. *Omega: The International Journal of Management Science, 29,* 585–597.

Ciavolino, E., Salvatore, S., & Calgagnì, 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

Civil Aviation Safety Authority (2011, April). Non-technical skills training and assessment for refgular pubic transport operations. Accessed July 24, 2013 at www.casa.gov.au/download/caaps/ops/sms-3-1.pdf

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* (revised edition). 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 (non-technical skills) system for assessing pilots' skills. *Human Factors and Aerospace Safety, 3*, 97–119.

Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.

Garfinkel, H. (1988). Evidence for locally produced, naturally accountable phenomena of order*, logic, reason, meaning, method, etc. In and as of the essential quiddity of immprtal ordinary society, (I of IV): An announcement of studies. *Sociological Theory, 6,* 103–109.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutor Quantitative Methods in Psychology, 8,* 23–34.

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

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.

Innocent, P. R., John, R. I., & Garibaldi, J. M. (2004). Fuzzy methods for medical diagnosis. *Applied Artificial Intelligence, 19*, 69–98.

International Civil Aviation Organization (ICAO) (2007). The level 4 language proficiency deadline: Issues and challenges. *The ICAO Journal, 63* (1), 5–25.

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

Lakoff, G. (1987). *Women, fire and dangerous things: What categories reveal abou the mind*. Chicago, IL: University of Chicago Press.

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.*

Mulqueen, C., Baker, D. P., & Dismukes, R. K. (2002). Pilot instructor rater training: The utility of the multifacet item response theory model. *International Journal of Aviation Psychology, 12*, 287–303.

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.

Phuong, N. H. (1995). Fuzzy set theory and medical expert systems: Survey and model. *SOFSEM '95: Theory and Prctice fo Informatics. Lecture Notes in Computer Science vol. 1012,* 431–436.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 15–35). Hillsdale, NJ: Lawrence Erlbaum Associates.

Roth, W.-M. (2005). Making classifications (at) work: Ordering practices of science. *Social Studies of Science, 35,* 581–621.

Rikers, R. M. J., Schmidt, H. G., & Moulaert, V. (2005). Biomedical knowledge: Encapsulated or two worlds apart. *Applied Cognitive Psychology, 19,* 223–231.

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.

Sharma, B., Orzech, N., Boet, S., & Grantcharov, T. (2013). Non-technical skills assessment in the post-operative setting. *Journal of the American College of Surgeons, 213 (supplement),* p. S122.

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.

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

Suchman, L. A., & Jordan, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association, 85*, 232–244.

Zaiontz, C. (2013). Real statistics using Excel. Accessed July 24, 2013 at http://www.real-statistics.com/reliability/cohens-kappa/

Footnotes

[1] For example, the Pearson product moment correlation, *t*-test, or *F*-test (ANOVA) involve (weighted) ratios of the explained to unexplained variance.

[2] This is not a true measure of IRR (Hallgren, 2012) but a ratio indicating agreement. Though often used as an indicator of IRR, overlapping arising by *chance* has not been factored out. In our situation, based on observations both within and across scenarios, Cohen's kappa values that Hallgren recommends as estimates of IRR lie in the poor ($0 < \kappa <$ .20) to fair range ($0.2 \leq \kappa <$ .4) (Zaiontz, 2013).

[3] In two-dimensional space, the Euclidean distance of any point from the origin, for example, can be calculated using the Pythagorean theorem $d = \sqrt{x^2 + y^2} = \left(x^2 + y^2\right)^{1/2}$.

[4] It may be argued that the pair simply missed this conversation. However, the protocols of the study allowed for the assessors to replay the scene as many times as they wished. This pair repeatedly replayed the scenario or sections thereof. It was concluded that this pair did not view this event as significant to comment on.

Table 1

The Captain's Handling of the Weather Situation by Level of SA

Pair	Summary of assessor pair comments	SA Level[1]
G2	Bad weather, but he is not predicting go around. Has difficulty predicting future events. There was a mention about the wind and that situation with the circling, but he never discussed the fact that they were down at minima and the chances are they might have to go around. Maybe didn't stay low enough to get under cloud base.	1
G1	(none)	2
G8	Captain made comment that weather is not flash out east. He had looked outside, saw the weather, but did not prepare the inside of the cockpit for dealing with it. Talked about weather in the East is not flash, is aware of deterioration. But did not link deterioration to having to fly missed approach. He perceived the environmental factors but didn't work out how this would affect them; he didn't project, didn't get missed approach into his headspace.	2
G6	He was aware of cloud in the circling area, that there [might be] IMC downwind and situation is worse. [We may] assume that cloud suddenly appeared.	2
G3	Recognizes that area ahead was terrible, says "it's a bit murky," as opposed to someone suddenly going IMC. He talks about weather, about effect might have, sees the fact that they were probably going into cloud again, acknowledged risk, weather looked horrible, looks dangerous, but carried on, did not do anything about it. He should have said "go-around He predicted future events, but carried on. He noted that there is a huge cloud of turbulence over there, gee looks dangerous, and fly straight into it. They were VMC, they could see that they were going IMC again." Should have reviewed go-around procedure after seeing weather crap downwind, too dangerous.	3
G4	He was able to see, was aware of weather, was able to see rain, but not about what was actually going to happen. Did not predict future events impacting flight safety fully. If had been briefing solid, if comment about weather, this could have been backup with recap: this is going to happen.	3
G5	Knew there was tailwind, did nothing about it. Went IMC without implications on flight safety.	3
G7	Did mention weather, was aware; perceived, but did he comprehend? Didn't Mentioned it, still flew into it, didn't do anything about it. He identified cloud and risk going IMC, but ended up entering cloud. Didn't comprehend implications of going IMC. Whilst perceived cloud, didn't comprehend what cloud meant . . . in terms of flight path. Could have been more proactive about the weather. Talked about it, refreshed it.	3
G9	Said at downwind that it didn't look good in circling area and continued to turn on left base. He just carried on. He was aware conditions were not so good [but] went into cloud. He had a gut feeling that this was going to happen, he had a feeling weather might not be good but went into cloud and was confused. He was surprised when got into left base and went IMC.	3

Note: According to the assessing pilots, based on their company rating scheme, which goes from 1 to 5.