International Journal of Industrial Ergonomics 44 (2014) 685–696

Contents lists available at ScienceDirect

International Journal of Industrial Ergonomics

journal homepage: www.elsevier.com/locate/ergon

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

Wolff-Michael Roth [a, c, *], Timothy J. Mavin [b, c], Ian Munro [d]

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

ARTICLE INFO

Article history:
Received 26 June 2013
Received in revised form
12 April 2014
Accepted 18 July 2014
Available online 28 August 2014

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

ABSTRACT

Regular performance assessment is an integral part of (high-) risk industries. Past research shows, however, that in many fields, inter-rater reliabilities tend to be moderate to low. This study was designed to investigate the variability of performance assessment in a naturalistic setting in aviation. A modified think-aloud protocol was used as research design to investigate the reasoning pairs of pilots use to assess the performance of an airline captain in a high-risk situation. Standard protocol analysis and interaction analysis methods were employed in the analysis of transcribed verbal protocols. The analyses confirm high variability in performance assessment and reveal the good, albeit fuzzy, justifications that assessor pairs use to ground their assessments. A fuzzy logic model exhibits a good approximation between predicted and actual ratings. Implications for the practice of performance assessment are provided.
Relevance to industry: Many industries aim at achieving consistency in identifying true performance levels. However, if the variability in performance assessment is a real phenomenon, as reported here, then practitioners and researchers might have to test whether it can be used positively, e.g., as opportunity for improving the resilience of crews.

1. Introduction

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

* Corresponding author. Applied Cognitive Science, Faculty of Education, MacLaurin Building A557, University of Victoria, Victoria, BC V8W 3N4, Canada. Tel.: +1 250 721 7764; fax: +1 250 721 7598.
E-mail address: mroth@uvic.ca (W.-M. Roth).

http://dx.doi.org/10.1016/j.ergon.2014.07.004
0169-8141/

2. Background: performance assessment and inter-rater reliability

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

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

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

In Europe, the NOTECHS (non-technical skills) system was developed for assessing pilots' crew resource management skills along the dimensions of cooperation, leadership and management, situational awareness, and decision making (Flin et al., 2003); in the southern hemisphere, the integrated MAPP system (Mavin and Roth, 2014)—which includes technical dimensions (aircraft flown within tolerances, knowledge/procedures) and non-technical dimensions (situational awareness, decision-making, management, communication)—is used by a number of airlines and the Australian national defence. However, a better understanding of the assessment process is required if it is to contribute to a decreased focus on threat and error management and an increase in crew resilience to surprises and anomalies in flight situations (Dekker and Lundström, 2007). This study was designed to better understand the sources of variability not only in pilots' performance assessment but also in the reasons provided for a particular assessment score. The study was to account for variability—that is, its nature as grounded in everyday reasoning of pilots—rather than treat it as a source of measurement error.

3. Methods

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

3.1. Research design

A standard method used by cognitive scientists to investigate what and how experts and non-experts think is the *think-aloud protocol* (Ericsson and Simon, 1993). Given some task, (non-) experts in the field of interest are invited to talk aloud while solving it. Although it has been suggested that the think-aloud protocol does not interfere with thinking, practitioners tend to find this method of eliciting data unnatural, providing them with difficulties of continuously saying what they think (Roth, 2007). However, a modified think-aloud protocol task asks participants to collaboratively solve tasks to arrive at a common solution (Suchman, 2007). This requires participants to articulate for each other everything needed to arrive at a shared solution; solving a problem in pairs and talking about one's reasons is experienced as more natural. This is particularly true in the airline industry where flight examiners and training captains externalize their assessments as part of their work; and the trainees and assessed pilots, as part of debriefing sessions, also externalize the reasons for making decisions while flying. The modified (pair-wise) design was therefore considered to provide higher ecological validity, to be experienced more naturally