How Pilots Assess Their Non-Technical Performance – A Flight Simulator Study

Patrick Gontar[1], Hans-Juergen Hoermann[2], Juergen Deischl[1] and Andreas Haslbeck[1]

[1]Institute of Ergonomics
Technische Universität München
Boltzmannstrasse 15, 85747 Garching, Germany

[2]Institute of Aerospace Medicine
German Aerospace Center (DLR)
Sportallee 54A, 22335 Hamburg, Germany

ABSTRACT

Within a full flight simulator study which aims to quantify the appropriateness of checklists and procedures in abnormal situations (Haslbeck, Gontar, & Schubert, 2014), 60 randomly chosen crews of commercial pilots (Airbus A320 and A340) flew a challenging flight-simulator scenario. Workload as well as time pressure constantly increased throughout the scenario. Crew performance and especially different aspects of Crew Resource Management (CRM) were subsequently rated by both pilots, who were asked to assess themselves and the other crewmember regarding their CRM skills. To avoid direct interaction during this assessment pilots were separated after the simulator flight. This approach allowed a comparison between pilots' self and peer-rating of their CRM-skills in relation to their crew position and the specific CRM aspect being assessed; a comparable study was not found in literature. The results indicate that cognitive skills are rated to be less positive than social skills. Furthermore, pilots seem to rate their colleague to perform better than themselves regarding all skills. The study showed that different response sets are of concern when assessing own and colleagues' skills regarding CRM. These findings can be incorporated in pilots' training as well as in further research.

Keywords: self-rating, peer-rating, Crew Resource Management, NOTECHS, flight simulator study

INTRODUCTION

Besides the technical skills and the knowledge of flying a modern jetliner, non-technical skills nowadays display an important and legally required component in the education and training of airline pilots. The training and integration of such non-technical skills is realized in the concept of Crew Resource Management. CRM-training therefore can be seen as a special program that, among others, specially focuses on interpersonal and communicational skills with the goal of enhancing team co-operation among pilots in order to prevent and manage human errors (Dietrich, Grommes, & Neuper, 2004; Fischer & Orasanu, 1999). According to this, communication displays a key factor in the implementation of a successful CRM (Nevile, 2004). The importance of communication for the safe and effective operation of a flight has been known for a long time (Dietrich et al., 2004). After several communication-

related aircraft accidents in the late 1970s, NASA held a workshop regarding "Resource Management on the Flightdeck". After investigations of these accidents in 1979, the conclusion was that the majority of accidents are directly related to human errors in the fields of interpersonal communication, decision making and leadership within the cockpit (Helmreich, Merritt, & Wilhelm, 1999; Dietrich et al., 2004). The collision of two Boeing 747 at Los Rodeos in Tenerife in 1977 is often named as an example. Such accidents gave significant impetus for the NASA workshop and for the development of today's Crew Resource Management. Initially, CRM was referred to "*Cockpit* Resource Management", primarily focusing on the modification of personal styles and the correction of deficits concerning pilot's individual behaviors (Helmreich et al., 1999; Salas, Burke, Bowers, & Wilson, 2001). With the integration of group dynamic aspects, the further development from Cockpit Resource Management to Crew Resource Management took place; until that time however with only little reference to the operational environment (Helmreich et al., 1999). In the 1990's, the concept was expanded to other personnel like cabin crews or maintenance staff. Currently, it refers to essential characteristics of the aviation system, like the organizational culture's impact on flight safety or the dealing with problems in the field of human factors (Helmreich et al., 1999). The subsequent introduction of an "Advanced Qualification Program (AQP)" by the Federal Aviation Administration (FAA) resulted in a phase of integration and proceduralization of CRM (Salas et al., 2001). In connection with these voluntary programs, airlines in the US were able to adapt their trainings individually to their organizational needs and conversely committed themselves to offer a CRM and Line-Oriented Flight Training (LOFT) to all their flight crews and to integrate CRM into their technical training (Helmreich et al., 1999; Salas et. al, 2001). Nowadays, CRM is understood as a threat and error management approach according to an understanding that human errors are omnipresent, inevitable, and displaying a valuable source of information. In Europe, the training of CRM skills and the assessment of airline crews regarding those skills is prescribed by regulatory authorities like the European Aviation Safety Agency (EASA). The underlying training methodology must be accepted by the national authority and published in the Operations Manual (OM) of the respective Airline (European Aviation Safety Agency, 2012, S. 74-79).

The basis for the assessment of crews within those trainings is provided by so called "behavioral markers" that describe characteristic and observable behaviors, which serve as performance indicators (Burger, Neb, & Hoermann, 2003). With the aim of developing a practicable, efficient, and an Europe-wide marker system for the assessment of pilot's non-technical skills, the former Joint Aviation Authority (JAA) Project Advisory Group started the „Non-Technical Skills (NOTECHS)" project in 1996. The project finally ended up in the 2001 completed research project JARTEL (Joint Aviation Requirements – Translation and Elaboration of Legislation) that, among others, served as a validation of the NOTECHS system (Flin et al., 2003). This system is composed of four categories with a subdivision in elements and behavioral markers as displayed in Figure 1.

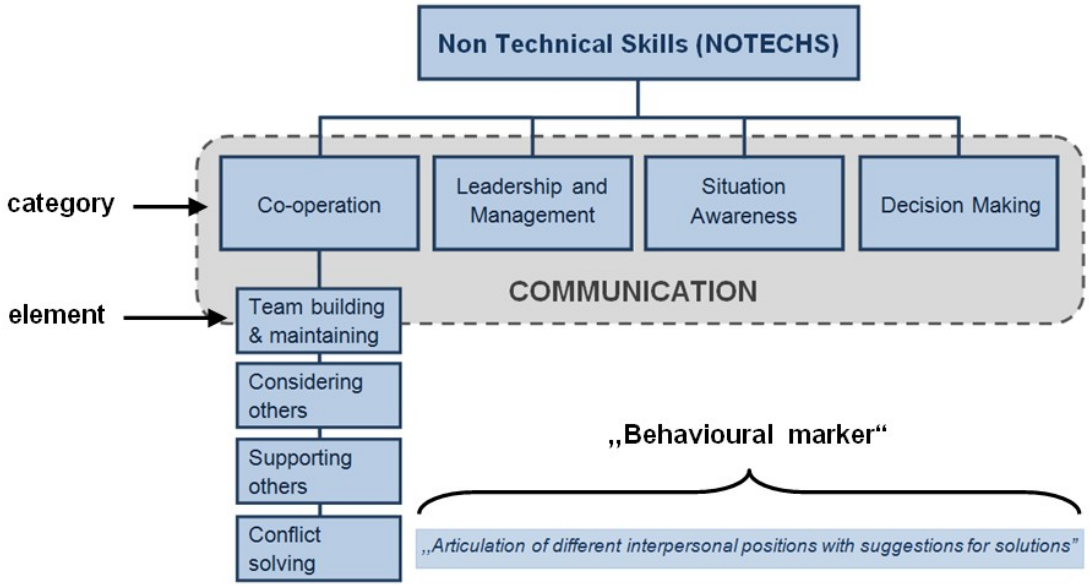

Figure 1: Non-Technical Skills "NOTECHS", based on Flin et al. (2003)

The four categories can further be divided into social skills (*Co-operation* and *Leadership and Management*) and cognitive skills (*Situation Awareness* and *Decision Making*). Communication thereby plays an essential role in all four categories because the described behaviors become observable, primarily in the form of verbal or non-verbal communication. Communication can thus be seen as a mediator for all four categories. Verbal utterances in form of discussions between pilots or in conjunction with decisions, for instance, allow conclusions about how a certain decision has been made (Flin et al., 2003).

A European airline, for example, has largely adopted the structure of this NOTECHS system and has developed its own marker system specially tailored to its organizational culture and CRM history (Burger et al., 2003). Especially the findings of a flight safety survey carried out between 1997 and 1999 within the airline and its subsidiaries with 2070 pilots reporting about 1897 safety-relevant occurrences played an integral part in the development of this system. The survey led to the conclusion that among the reported incidents factors related to the social relationship between crew members played a larger role than human error, operational or technical problems (Burger et al., 2003). According to the derived demand for a definition and for a training of specific communication behaviors, communication – in contrast to the NOTECHS system – had been included as a separate category within the non-technical skills referred to as *Interpersonal Competence*. Besides the other *Interpersonal Competence* categories (*Leadership and Teamwork*, *Workload Management*, and *Situational Awareness and Decision Making*) communication should be understood as a linkage between social and cognitive skills (Burger et al., 2003). In addition to technical and procedural competences those interpersonal competencies have been implemented in the airline's selection processes, training programs and recurring line-oriented and simulator trainings (Burger et al., 2003). Surveys of airline pilots like those of Beaubien and Baker (2002) have shown that CRM is currently perceived as a valuable training concept by most pilots.

Motivated by the success in the aviation environment, efforts were undertaken to incorporate and validate an adjusted NOTECHS rating scale in surgical and resuscitational teams as well (Flin & Maran, 2004; Mishra, Catchpole, & McCulloch, 2009; Sevdalis et al., 2008; Steinemann et al.,2012).

MOTIVATION & RESEARCH QUESTION

This investigation is part of a flight simulator experiment observing pilots' behavior in abnormal situations (Haslbeck et al., 2014). Since the rating of NOTECHS is normally conducted by flight instructors, the objective of this study is to investigate how biased or unbiased the actual crew members can assess their CRM-skills and how this assessments might be influenced by different response sets. Pilots are trained to have high skills in assessing a situation and their own technical performance, but usually crewmembers do not rate each other in terms of NOTECHS performance. However, this mutual perception of CRM behaviors is an important factor for efficient crew performance and shared mental models. This study investigates how pilots rate themselves (internal assessment/self-rating) and their partners (external assessment/peer-rating) in the cockpit regarding their CRM performance as a function of their position within the flight crew. A realistic self-evaluation would stand for good self-reflection, which again is a central component of ones' interpersonal skills. Furthermore, a critical peer-rating could enhance constructive feedback after a training session because it is based on a team member who was directly involved in fulfilling the flight task in contrast to a rating from an instructor. To our knowledge, by now, no single study observed the differences in self and peer-rating within the scope of NOTECHS.

Harris and Schaubroeck (1988) summarized three different types of egocentric biases to be expected when looking onto self and peer-rating. Their findings are presented in Table 1. They furthermore found that the organizational level respectively the rank (in our case the cockpit position: captain or first officer) could influence the weighting of different performance dimension (Klimoski & London, 1974; Zammuto, London & Rowland, 1982). This would lead to a disagreement in an overall rating but probably not on a specific item (Harris & Schaubroeck, 1988). With this assumption met, NOTECHS seems to be appropriate to compare the ratings of captains (CPTs) and first officers (FOs).

Table 1: Effects and consequences of egocentric biases according to the meta-analysis of Harris and Schaubroeck (1988)

Effect of egocentric bias	Consequence on rating
Defensiveness	Self-rating would lead to overestimated rating (Holzbach, 1978; Steel & Ovalle, 1984) with restricted range.
Self-esteem	Self-rating done by high self-esteem raters might be overestimated; low self-esteem raters may not overestimate (Baird, 1977; Kay, Meyer, & French, 1965).
Attribution theory	Self-rating would assign good performance to one's own attributes, whereas in contrast poor performance would be assigned to environmental factors (DeVader, Beateson, & Lord, 1986; Jones & Nisbett, 1972).

With these potential biases in mind, the following research questions were formulated:

RQ 1: Are there differences in self and peer-rating within a cockpit crew? If so, which dimensions are primarily sensitive to this effect?

RQ 2: What is the influence of crew position on the level of self and peer-ratings?

METHOD

Test design

According to these research questions, the generated test design defines two independent variables: the crew position, where pilots are either captains or first officers and the source of rating, which can either be the self or the peer-rating. Both variables are permuted, so that each member of the crew conducts both ratings (compare Figure 2).

Figure 2: Visualization of the rating system

The rating scores are defined as dependent variable with four levels according to the adapted NOTECHS system*:* *Communication* (1), *Leadership and Teamwork* (2), *Work Organization* (3), *Situation Awareness and Decision Making* (4). Therein, (1) and (2) address the social factors, whereas in contrast (3) and (4) refer to cognitive factors. The four different levels respectively dimensions are assessed by a total of 40 items on a five-point Likert scale, where seven to 15 items are averaged to one level. The rating scale leads from the most positive '++', which was coded as '1', via a neutral state codes as '3' to the most negative '--', which was coded as '5'.
Regarding research question 1, the experimental scenario should lead to a very high workload with the chance of failure in order to observe the awaited effects formulated in the research questions. Furthermore, a strong emphasis on decision making is provided in order to make differences in social and cognitive factors sufficiently obvious. Additionally, both pilots must be participants and cannot be replaced by a member of the experimental team, since decision making and cooperation is always dependent on a whole team which itself depends on both individuals and their disposition.

Experimental scenario

Within a full-flight simulator environment, 60 randomly chosen crews (60 CPTs and 60 FOs) performed a critical high workload scenario (Haslbeck et al., 2014). The CPTs including one female were M = 47, SD = 6 years old and had M = 13,380, SD = 3,626 hours of flight experience, whereas the FOs including five females were M = 33, SD = 5 years old and had M = 5,325, SD = 2,723 hours of flight experience. According to their aircraft types, either John

F. Kennedy Airport in New York (Airbus A340) or Nice Côte d'Azur Airport (Airbus A320) was approached to land. When lowering the gear, a technical malfunction was evoked, which led to a damaged hydraulic system affecting the nose gear such that it was neither positioned down and locked nor able to retract. With this misconfigured flight state, the crew had to perform a go-around, implement the compulsory procedures and prepare for a new approach while their fuel consumption – affected by the high aerodynamic drag – was doubled. At the time when the crew extended their flaps, the remaining hydraulic pressure, which led to slow moving flaps, could not retain symmetric movement so that the wing tip brake irrevocably stopped the flaps extending. According to the regulations, another go-around had to be performed and further procedures had to be completed. But as the gear was still down and the low fuel led to a mayday situation, it was expected that some crews might have been forced to abort the procedures and directly perform a landing to ensure a higher level of safety.
An approach briefing was conducted by the crews before entering the simulator. At this time, the participants did not know that they have to rate CRM related items later on. The following scenario lasted about 30 minutes – depending on the crews' performance. After the flight being completed, both pilots were asked to fill out the CRM rating questionnaire for both – self and peer – in two different rooms to avoid prior exchange of views. An instructor operating the flight simulator rated both pilots according to their CRM performance and the crew in total regarding their overall behavior using the LOSA rating procedure (Flight Safety Foundation, 2005). A debriefing concerning the crews' decision making completed the experiment.

RESULTS

In total, 120 pilots assessed themselves and their colleague, which leads to a total amount of 240 ratings, whereof eight are incomplete and therewith excluded from further analysis. A descriptive analysis of the results shows that both, the self and the peer-rating are better rated (M = 2.1, SD = 0.5) than the neutral score, which would be represented by a value of '3'. The mode of all 40 items rated, comes to a value of '2', which indicates – although the used Likert scale with five steps does not lead to a force choice format – an *error of central tendency* is rather not observed.

Regarding statistical significance, results are stated as significant at a .05 level in all research questions and assumptions following.

Research question 1 (NOTECHS dimensions)

A two factorial ANOVA was conducted to compare the differences in the four CRM rating dimensions being affected by the source of judgment. Mauchly's test shows that sphericity condition could not have been met for the main effect of CRM-dimension, $^{2}(5) = 75.37$, $p < .001$, as well as for interaction effect CRM-dimension*self/peer, $^{2}(5) = 39.19$, $p < .001$. Consequently, Greenhouse-Geisser corrected degrees of freedom were used in these cases.

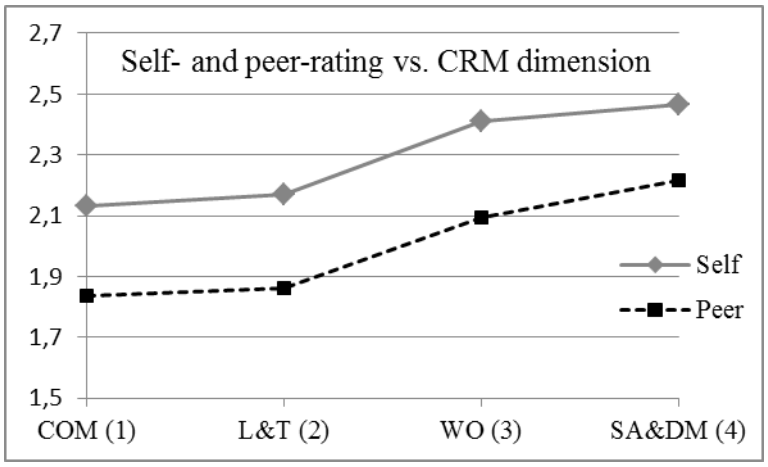

Figure 3: Self and peer-rating as a function of the CRM rating dimensions

Figure 3 shows the results regarding the first research question. It shows a significant main effect of the factor CRM-dimension, $F(2,218.05) = 49.19$, $p < .001$, $\eta_p^2 = .31$. Pairwise comparisons for the main effect CRM-dimension using Bonferroni corrected adjustments show a non-significant difference between level 1 (*Communication*) and level 2 (*Leadership and Teamwork*). All other comparisons show significant differences on the specified .05 level. Especially level 2 ($M = 2.02$, $SD = .47$) leads to significantly better values than level 3 (*Work Organization*, $M = 2.25$, $SD = .56$), which is also higher rated than level 1. Level 4 (*Situation Awareness and Decision Making*, $M = 2.34$, $SD = .61$) is rated significantly least positive compared to all other levels. Furthermore these findings indicate that the social aspects (level 1 and level 2) are better rated than the cognitive dimensions (level 3 and level 4). A significant main effect was also found for the influence of the source of judgment, $F(1,109) = 79.97$, $p < .001$, $\eta_p^2 = .42$. Pairwise comparisons for the second main effect using Bonferroni corrected adjustments show differences between the both characteristics self-judgment and peer-judgment. Judgments rating oneself ($M = 2.29$, $SD = .53$) are assessed significantly higher (less positive) than ratings for the respective colleague ($M = 2.00$, $SD = .54$). Furthermore, the interaction effect between the category of CRM-dimension and the self/peer-judgment was not found to be significant, $F(2.44,266.25) = 1.16$, $p = .32$. This means that both main effects can be interpreted separately.

In sum, the results show a large effect regarding the source of rating. Pilots in general rate themselves lower performing than their colleagues; this seems to be independent from the aspect they are rating. Furthermore, the four CRM-rating dimensions are assessed differently. An interaction between the CRM-dimension and the source of judgment would mean that self-ratings and peer-ratings would differ in some dimensions, which is not the case.

Research question 2 (effect of crew position)

Again, a two factorial ANOVA was conducted to compare the effect being induced by the factor of crew position and the already contemplated main effect of the factor source of judgment on the overall rating (see Figure 4).

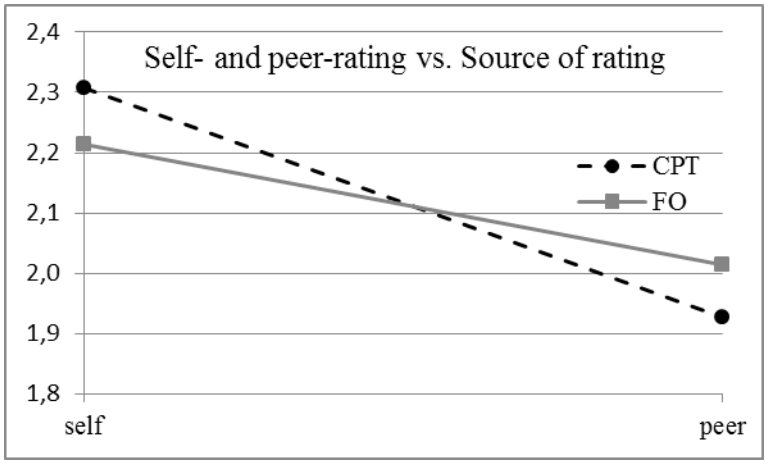

Figure 4: Self and peer-rating as a function of crew position

Within this analysis, the awaited main effect of crew position could not be identified, $F(1,53) = .002$, $p = .97$. The analysis shows in this context again a main effect of source of judgment, $F(1,53) = 90.14$, $p < .001$, $\eta_p^2 = .63$. The difference between the judgment of CPTs ($M = 2.12$, $SD = .52$) and FOs ($M = 2.12$, $SD = .59$) is obviously not significant. This means, CPTs and FOs in general do not rate differently, but seem to rate differently when they rate themselves or their colleague.

Indeed, analyzing interactions, a significant interaction effect crew position * source of judgment was found, $F(1,53) = 9.14$, $p = .004$, $\eta_p^2 = .15$. CPTs seem to rate themselves more skeptical than FOs rate themselves, but FOs rate CPTs more skeptical than CPTs rate FOs. Whether this interaction reflects objectively worse performance of the CPTs has to be clarified by further analysis.

https://openaccess.cms-conferences.org/#/publications/book/978-1-4951-2097-8

CONCLUSION & DISCUSSION

The results regarding the first research question show differences in the rating regarding the NOTECHS-dimensions as well as between self and peer-rating. One possible explanation for the observed differences between the CRM dimensions might be based on objective differences between social and cognitive performance, and on the other hand on subjective perception. The mentioned objective differences can be analyzed when having a look onto the instructor ratings, which will be conducted in a next step and reported separately. The results indicate that crews may have a good and open social atmosphere, but nevertheless perform less well in work organization and decision making. Subjective differences might either be based on different strictness regarding the different dimensions or being influenced by the success of the completed scenario. The subjective differences might be connected to a hindsight bias, a known response set influencing the rating when the success of the action (in this case the whole scenario) is known. Especially in this scenario, where the success highly depends on the right decisions at the right time during high workload, it may lead to less positive rating scores in the cognitive aspects of work organization and situation awareness/decision making. Other aspects could be an influence of the recency effect. Since the overall success in the entire scenario does not become obvious before the safe landing, it seems coherent that success or non-success during this final actions influence the overall rating to a larger degree. This aspect will be further discussed after analyzing the LOSA performance data.

Results regarding the second research question additionally show the effects of the factor crew position towards the self and peer-rating reflected in a significant interaction effect. Airline pilots are trained to high cooperation and team building which may lead to the observed difference in self and peer-rating. Pilots in general seem to be rather self-critical when rating CRM, which is in agreement with the training goals and the strict psychological selection at the beginning of their careers (Goeters, 2004). Another possible explanation would be that no one wants to unmask his colleague. It can be assumed that both, the self-criticism and the unmasking aspect, have an influence that cannot be assessed separately within this experimental design. Egocentric biases as introduced in Table 1 were not observed. Rather the opposite seems to be the case. A positive peer-bias is found because self-ratings are less favorable than peer-ratings. The fact that CPTs' CRM-skills are assessed more skeptical than FOs' in both ratings (self and peer) might be based on the CPTs objectively performing less well than their colleagues. On the other side, CPTs are using a wider range of the evaluation spectrum. This might be the result of the CPTs having more flight experience, working for a significantly longer amount of time at the company and being evaluated more often during training. This might result in higher evaluation proficiency and potentially diverging response sets.

Although the results reflect significant differences including strong effects, it has to be kept in mind that the pilots themselves are not explicitly trained to assess one's CRM behavior. However, as said before, the used items and scales are familiar to all pilots within the company. The degree of correspondence between self and peer-ratings can also be regarded as an important element of a shared mental model of the crew.

Taken together, it was clear to see that pilots' assessment of their own performance was less positive than that given for their colleagues. Possible explanations were given above and are subject to further analyses as well. These will compare this data to the external training-instructors' rating. This may give further details to which extend CPTs are actually performing worse. Furthermore, interclass correlations coefficients (ICCs) can be calculated and the inter rater reliability estimated. One step further, the relationship between the CRM and the additionally evaluated LOSA rating (Flight Safety Foundation, 2005) will be assessed. A first look onto the data shows that the worst performing crews do not show significant difference in their overall subjective CRM rating compared to the best crews, which in fact might be an indication of a self-serving bias.

ACKNOWLEDGEMENTS

The authors highly acknowledge the support and help by Cpt. Manfred Mueller, Cpt. Manfred Binder, Cpt. Peter Cröniger, Verena Porstner and Cpt. Carsten Schmidt-Moll as well as their contribution and commitment to the project.

REFERENCES

References marked with an asterisk (*) indicate studies included in the meta-analysis (Harris & Schaubroeck,1988).

*Baird, L. S. (1977). Self and Superior Ratings of Performance: As Related to Self-esteem and Satisfaction with Supervision. *Academy of Management Journal*, *20*(2), 291–300.

Beaubien, J. M., & Baker, D. P. (2002). *Airline Pilots' Perceptions of and Experiences in Crew Resource Management (CRM) Training*. Warrendale, PA: SAE International.

Burger, K.-H., Neb, H. and Hoermann, H.-J. (2003). Lufthansa's new basic performance of flight crew concept - A competence based marker system for defining pilots performance profile. *Proceedings of The 12th International Symposium on Aviation Psychology*, *1*, 172–175.

*DeVader, C. L., Bateson, A. G., & Lord, R. G. (1986). Attribution theory: A meta-analysis of attributional hypotheses. In E. Locke (Ed.), *Generalizing from laboratory to field settings* (pp. 63–81). Lexington, MA: Lexington Books.

Dietrich, R., Grommes, P., & Neuper, S. (2004). Language processing. In T. M. Childress & R. Dietrich (Eds.), *Group interaction in high risk environments* (pp. 87–101). Aldershot: Ashgate.

European Aviation Safety Agency. (2012). *Acceptable Means of Compliance (AMC) and Guidance Material (GM) to Part-ORO*.

Fischer U., & Orasanu, J. (1999). Say it again, Sam! Effective Communciation strategies to mitigate pilot error. In R. Jensen & L. Rakovan (Eds.), *Proceedings of the 10th International Symposium on Aviation Psychology* (pp. 362–366). Columbus, OH.

Flight Safety Foundation. (2005). Line operations safety audit (LOSA) provides data on threats and errors. *Flight Safety Digest*, *24*(2), 1–18.

Flin, R., & Maran, N. (2004). Identifying and training non-technical skills for teams in acute medicine. *Quality and Safety in Health Care*, *13*(Suppl 1), i80.

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

Goeters, K.-M. (2004). *Aviation psychology*. Aldershot, Burlington, VT: Ashgate.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self– supervisor, self–peer, and peer–supervisor ratings. *Personnel Psychology*, *41*, 43–62.

Haslbeck, A., Gontar, P., & Schubert, E. (2014, in press). How Can Procedures and Checklists Help Pilots in Abnormal Flight Situations? In *Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics*.

Helmreich, R. L., Merritt, A. C., & Wilhelm, J. A. (1999). The evolution of crew resource management training in commercial aviation. *The International Journal of Aviation Psychology*, *9*(1), 19–32.

*Holzbach, R. L. (1978). Rater bias in performance ratings: Superior, self-, and peer ratings. *Journal of Applied Psychology*, *63*(5), 579–588.

Jones, E. E., Kanouse, D. A., Kelly, H. H., Nisbett, R. E., Valines, S., & Weiner, B. (Eds.). (1972). *Attribution: Perceiving the causes of behavior*. Morristown, NJ: General Learning Press.

*Jones, E. E., & Nisbett, R. E. (1972). The actor and the observer: Divergent perceptions of the causes of behavior. In E. E. Jones, D. A. Kanouse, H. H. Kelly, R. E. Nisbett, S. Valines, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 79–94). Morristown, NJ: General Learning Press.

*Kay, E., & Meyer, H. H. (1965). Effects of threat in a performance appraisal interview. *Journal of Applied Psychology*, *49*(5), 311–317.

*Klimoski, R. J., & London, M. (1974). Role of the rater in performance appraisal. *Journal of Applied Psychology*, *59*(4), 445–451.

Mishra, A., Catchpole, K., & McCulloch, P. (2009). The Oxford NOTECHS System: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. *Quality and Safety in Health Care*, *18*(2), 104–108.

Nevile, M. (2004). *Beyond the black box: Talk-in-interaction in the airline cockpit*. *Directions in ethnomethodology and conversation analysis*. Aldershot, Hants, England: Ashgate.

Salas, E., Burke, C. S., Bowers, C. A., & Wilson, K. A. (2001). Team training in the skies: Does crew resource management (CRM) training work? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(4), 641–674.

Sevdalis, N., Davis, R., Koutantji, M., Undre, S., Darzi, A., & Vincent, C. A. (2008). Reliability of a revised NOTECHS scale for use in surgical teams. *The American Journal of Surgery*, *196*(2), 184–190.

*Steel, R. P., & Ovalle, N. K. (1984). Self-appraisal based upon supervisory feedback. *Personnel Psychology*, *37*(4), 667–685.

Steinemann, S., Berg, B., DiTullio, A., Skinner, A., Terada, K., Anzelon, K., & Ho, H. C. (2012). Assessing teamwork in the trauma bay: introduction of a modified "NOTECHS" scale for trauma. *The American Journal of Surgery*, *203*(1), 69–75.

*Zammuto, R. F., London, M., & Rowland, K. M. (1982). Organization and rater differences in performance appraisals. *Personnel Psychology*, *35*(3), 643–658.