36 - Flight Crew Performance and CRM Ratings Based on Three Different Perceptions

Patrick Gontar[1] and Hans-Juergen Hoermann[2]

[1]Institute of Ergonomics, Technische Universität München, Munich, Germany

[2]Institute of Aerospace Medicine, German Aerospace Center (DLR), Hamburg, Germany

Abstract. Based on the foregoing work of Gontar, Hoermann, Deischl, and Haslbeck (2014), this paper presents different aspects of rater reliability in the context of a flight simulator study. For this purpose, 120 commercial airline pilots had to fly a challenging approach scenario in a full flight simulator. Afterwards, both the crew members and the instructor rated the non-technical skills separately for both pilots. Results indicate that social aspects of Crew Resource Management are subject to a broader rating variability than cognitive aspects are. Furthermore, it is indicated that a differentiation between the two crew members is not always possible for the raters.

Keywords: CRM, NOTECHS, Inter-Rater Reliability, Flight Simulator Study

Introduction

For decades, it has been seen that non-technical skills are a vital element for safe flight operations. With the Joint Aviation Authorities (JAA) asking for the assessment of those non-technical skills, the European project JARTEL (Joint Aviation Requirements Translation and Elaboration of Legislation) worked on a consolidated assessment method to evaluate pilots' non-technical performance (Flin et al., 2003).

The origin of such behavioral marker systems came up with the development of a checklist system called Line/LOS Checklist (LLC) (Helmreich, Wilhelm, Kello, Taggart, & Butler, 1990), which was developed from pilots' attitudes towards cockpit management systems (Helmreich, 1984) and an analysis of accidents and incidents (Connelly, 1997; O'Connor, Hoermann, Flin, Lodge, & Goeters, 2002). Those checklist-based analysis systems have been used to develop different behavioral marker systems over the years (Flin & Martin, 2001) and were finally incorporated within the *Line Operations Safety Audit (LOSA)* (Helmreich, Klinect, & Wilhelm, 1999). LOSA was set up in order to collect data on how flight crews manage threats and errors during normal line operations. The crews' performance is evaluated as a whole and not solely the individual pilots' skills (Klinect, Murray, Merritt, & Helmreich, 2003). In addition, LOSA considers contextual data about the respective flight operation and even more on the organization itself (Flight Safety Foundation, 2005). Over the years, LOSA became one of the most commonly used evaluating schemes for the cockpit crews' behavior and was validated in wide ranges (Flin & Martin, 2001); see Butler (1991) or Law and Wilhelm (1995) to mention only two important examples.

The goal of JARTEL's new assessment system to rate pilots' non-technical skills (NOTECHS) was to develop a feasible, efficient and Europe-wide marker system (van Avermaete & Kruijsen, 1998). Its objective was to assess pilots' non-technical skills in order to enhance team cooperation among pilots as well as to improve their interpersonal and communicational behavior (Dietrich, Grommes, & Neuper, 2004; Fischer & Orasanu, 1999). The NOTECHS-consortium separated the behavioral markers into four different categories called: *Cooperation*, *Leadership and Management*, *Situation Awareness*, and *Decision Making*. Furthermore, it was stated that communication is inherent in every category and that it is not defined as a separate category (O'Connor et al., 2002). Each category has different items that are rated on a five-point scale from *very poor* to *very good*. After validation, the NOTECHS scheme itself was the groundwork for a new rating system incorporated within an airline's performance rating system. For it, Burger, Neb, and Hoermann (2003) adapted the NOTECHS system to the airline's own organizational culture and Crew Resource Management philosophy. They came up with four categories of interpersonal competencies where *Communication* is seen as a discrete dimension in addition to *Leadership and Teamwork*, *Workload Management*, and *Situational Awareness and Decision Making*. Using the adapted NOTECHS system, Gontar, Hoermann, Deischl, and Haslbeck (2014) examined flight crew members who mutually assessed their own and others' CRM-skills after flying a simulator scenario. They found that self and peer ratings of pilots differ to a higher degree than one would expect. It was shown that pilots rate their own performance worse than their colleagues' performance in all four dimensions, which might be an effect of pilots not wanting to unmask their colleagues. Furthermore, it was found that social factors are rated higher than cognitive factors, which might be a result of the scenario that was used. While those results are based on ANOVA analyses and therefore on mean value comparisons of the two involved pilots, the question about inter-rater reliability is not yet answered. Running such analyses, low correlations might represent and reveal different perceptions of one's performance.

Method

Although Gontar et al. (2014) have shown that the source of a rating (self or peer rating) has a huge influence on the average level of the rating, the perceptions of the instructors were not taken into account. With this rating provided, it might be possible to see whose perceptions of the performance are more in accordance with those of a third non-participating instructor. To get an idea of the actual rater reliability, the first approach asks for differences of the inter-rater reliability between the four rating dimensions, assuming that the reliability might depend on the specific dimension which is rated. This possible influence was already indicated by O'Connor et al. (2002) showing the dependencies of the inter-rater agreement on the flight scenarios. So this can be formulated as a global research question (RQ):

> RQ1: How good is the inter-rater reliability of the three raters based on the four dimensions of the adapted NOTECHS?

Assuming only low inter-rater reliability as seen in Flin and Martin (2001), a coherent approach asks for the origin of those effects, which leads to more specific research questions:

> RQ2: Which raters agree on which dimension and to what degree? Are there particular raters that do not agree with each other?

To answer the stated research questions, a test design according to Figure 1 was established. According to Gontar et al. (2014), both pilots rate themselves (self rating) and their colleagues (peer rating). In addition, an instructor pilot rates the two pilots differently as well (supervisor rating).

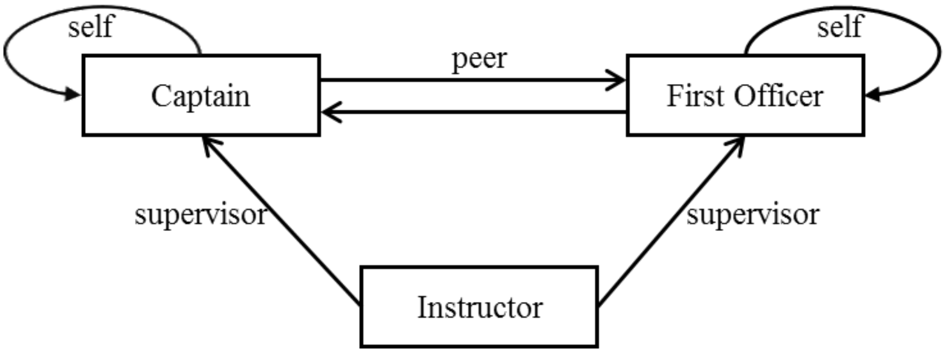

Figure 1. Ratings conducted by the three raters based on Gontar et al. (2014)

The CRM skills that are assessed are based on the internal company assessment guidelines (Burger, Neb, & Hoermann, 2003), which refer to the mentioned and adapted NOTECHS scheme (O'Connor et al., 2002), and are known to all raters (pilots and instructor) in advance. The rating is based on four dimensions which are defined as a dependent variable with four levels: *Communication* (1), *Leadership and Teamwork* (2), *Work Organization* (3), and *Situation Awareness and Decision Making* (4). Therein, (1) and (2) address the social factors, whereas in contrast (3) and (4) refer to cognitive factors (Gontar et al., 2014). Those dimensions are assessed with the aid of a total of 40 items. The underlying five-point scale leads from the most negative rating "--" to the most positive rating "++". To avoid bias when filling out the rating forms, the three raters were separated into different rooms.

To cover a broad variety of performance in the different aspects, a challenging mission has to be flown by several crews of commercial pilots holding valid licenses. For this 120 pilots (60 Captains, 60 First Officers) were randomly chosen from a commercial airline conducting a landing mission with a double malfunction, which itself is based on one single underlying failure. The Captains (CPTs), including one female, were M = 47, SD = 6 years and had M = 13,380, SD = 3,626 hours of flight experience, the First Officers (FOs), including five females, were M = 33, SD = 5 years and had M = 5,325, SD = 2,723 hours of flight experience (Gontar et al., 2014). Ratings and simulator handling were alternately conducted by two retired instructor pilots. The flight mission using two *JAR STD 1A Level D* full flight simulators began with the aircraft being fully configured and established for a visual approach either to John F. Kennedy Airport (Airbus A340 fleet) or Nice Côte d'Azur Airport (Airbus A320 fleet) with fuel on board for approximately one hour of remaining flight time. When lowering the landing gear for the final configuration, the green hydraulic system leaked and consequently lost pressure and the total amount of hydraulic fluid. While only the main gear was locked, the nose gear could not have been locked and remained in an uncertain position and not able to retract. As the landing gear is extended, the aerodynamic drag is almost doubled and so is the fuel consumption, leading to a remaining flight time of approximately 30 minutes. The crew now had to abort their approach and proceed in the standard missed approach, while handling the procedures and deciding which runway to approach next.

With the second approach and meanwhile the gear fully down and locked (using landing gear gravity extension), the slow movement of the flaps (only driven by one hydraulic system), led to their jamming. With 15 minutes of remaining fuel, the crew again had to complete further procedures and had to work through the malfunction of the flight controls.

At that time, the crew will have to decide whether to proceed with their checklists or to force a landing without the procedures being completed. It can be expected that only a part of the flight crews will manage to complete all the checklists and procedures within the mission; other crews are expected to fail at some point due to the enormous time pressure.

The whole simulation experiment took place at a flight training facility for a duration of 20 nights, where three crews participated every night. After a demographic questionnaire being completed by the participants, the crews were randomly paired and one after the other went through the scenario, which lasted about 30 minutes. Afterwards, the two pilots, as well as the instructor, rated their CRM performance independently from each other based on the introduced rating forms. A debriefing concerning decision making completed the experimental process.

Results

Regarding the global approach, the intraclass correlation analysis compares the three different judgments of the three raters (self, peer, supervisor) for the two subjects (CPT and FO) using a two-way random model. Table 1 shows the results based on consistency agreements for both subjects.

Table 1. Intraclass correlation coefficient of the ratings for the respective subjects

ICC	COM	L&T	WO	SA&DM
CPT is rated	.23	.10	.52	.48
FO is rated	.08	.03	.45	.47

Both ratings show relatively low intraclass correlation coefficients (ICCs) for the social aspects but relatively high ICCs for the cognitive aspects. Those results are in contrast to what was expected when assuming that social aspects are easier to observe and to evaluate. It seems that the pilots have either different perceptions or diverging expectations of good *communication* and *leadership & teamwork*. However, the cognitive aspects, which have to be concluded from observed behaviors show substantially higher consensus among the raters.

To get a more detailed idea of where those different ratings have their origin, the Pearson correlation coefficient between each pair of ratings is calculated. For better and easier interpretation, the aspects of *Communication* and *Leadership & Teamwork* are merged to social aspects as well as *Work Organization* and *Situation Awareness & Decision Making* are merged to cognitive aspects, using Fisher-Z transformations. Figure 2 shows those correlation coefficients using the notation .social/.cognitive aspects for the coefficients and *Rater → Subject* to identify the source and the target of the ratings. For example, the rating of the instructor for the Captain (INST→ CPT) correlates with the rating of the instructor for the First Officer (INST → FO) with .65 for the social skills and with .82 for the cognitive skills.

All the coefficients of social aspect ratings represent a lower correlation than for the cognitive aspects in every combination. Although the cognitive aspects were expected to be more difficult to observe (O'Connor et al., 2002), it seems that the raters are more in agreement when delivering their judgments. A possible explanation could be that the mental model of what is a good or adequate behavior regarding *communication* and *leadership & teamwork* varies greatly between the raters, at least in this scenario. Moreover, the evaluation of good social performance might imply more interpretation producing this higher degree of variability. As the success of this scenario is mainly subject to the right decisions at the right time as well as to good work organization, which implies good cognitive skills, the performance of those aspects needs less interpretation, because they can simply reflect the outcome of the flight. The overall performance of the entire mission might also be assigned to those cognitive aspects as they are considered more critical.

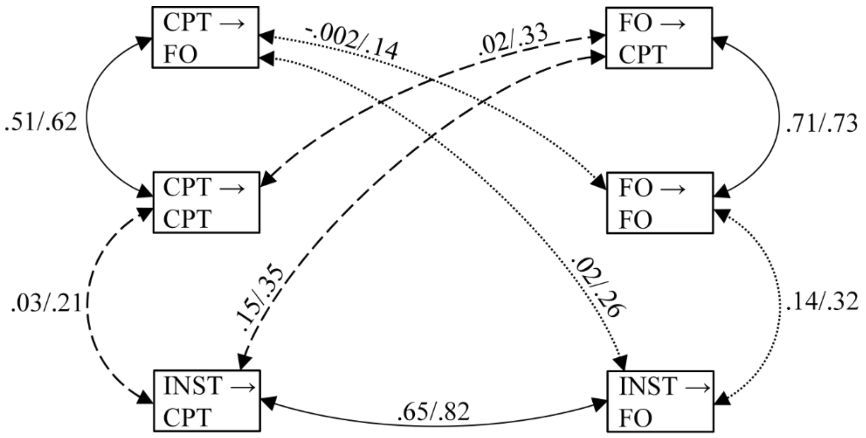

Figure 2. Correlation coefficients between the particular ratings (Rater → Subject), where the coefficients represent the social and cognitive aspects of the adopted NOTECHS scheme. Those coefficients are arranged as following: .social/.cognitive. The solid lines show the correlations between the ratings of one rater; the dashed and dotted line show the correlations for the same subject, which is rated.

When looking at the ratings from one rater, the correlations are expected to be rather low, since the subjects rated are different (solid lines in Figure 2). However, both dimensions of the instructor's rating for the two pilots (INST → CPT and INST → FO) have a high correlation (.65 and .82). It seems that the instructor does not distinguish between the two crew members and rates the crew as one entity. The same finding holds true for the two other raters (CPT and FO) with their related ratings (CPT: .51/.62, FO: .71/.73). While the instructor sees the crew from an outside point of view, the pilots involved in the mission may also see themselves as one crew and perceive the crews' CRM as a team result. Single contributions cannot be assigned directly to one of the pilots. Those results may further indicate that a differentiated rating of two pilots forming a crew is not always possible, perhaps also not useful.

The ratings for one target were assumed to show high correlations as they evaluate the same person and therefore one performance (dashed and dotted lines in Figure 2). In contrast, to the correlations of the ratings conducted by one person, the ratings for one subject correlate rather low. This means that every rater has a different perception and understanding of the subjects' performance. Gontar et al. (2014) already showed that the self rating is consistently less positive than the peer rating. A comparable bias might be observed here as well. Thereby, it is interesting to see that the ratings of the instructor correlate more with those of the FO than with those of the CPT. Although the CPTs are assumed to have more experience and expertise, this cannot be confirmed by the accuracy of their NOTECHS-ratings. The higher correlation of the ratings between INST and FOs may be based on better subjective perception by the FO when defining the instructors' rating as the gold standard. A contrary approach might follow the line of thought that the instructor, who is also a CPT with special trainings, does not have the insight of the real scenario and so the participating CPT's expected performance on the team is higher. Consequently, the CPT might evaluate a situation where he is part of the crew, differently and more differentiated than a situation, where he serves as an instructor. Analyses of further correlation coefficients are not conducted since respective research questions have not been formulated.

Discussion & Conclusion

The presentation of the results shows the broad variation between different pilots rating the same subject. Answering RQ1, this means that only a rather low inter-rater reliability was measured differing between social and cognitive aspects. Possible explanations are given above and are subject to further investigations and discussions. The findings seem to partly coincide with those of O’Connor et al. (2002), which show that complex scenarios, where more interpretation is needed, might lead to lower inter-rater reliability when assessing NOTECHS. In contrast to their findings, where the inter-rater reliability was nearly .8, the results in this paper are way off. However, it must be kept in mind that O’Connor et al. (2002) used instructor pilots that rated the crew and calculates their inter-rater reliability. A different approach was used here. Our goal was to examine where different perceptions become evident. Nevertheless, the low degree of agreement between the three raters might have an influence on operational safety and flight training, which both depend on a common understanding of the crew status and on adequate feedback. The analyses concerning RQ2 showed that the perceptions, and therefore the ratings, highly depend on the rater himself and less on the rated subject. If the perception of the crew members differ much from the perception of the instructors, it will be difficult for the trainees to accept possible suggestions for improvement. An integrated approach to solve this problem could be to train the pilots from the beginning on how to notice and how to judge CRM related aspects during flight operations and simulator missions. When arguing in this direction, one has always to be aware that the crew was not explicitly trained to use standardized rating methods, but is trained on what those aspects mean.

Another important factor is that it seems not all the raters are able to or willing to distinguish between the two crew members’ performance. On the one hand side this could be due to the crew establishing a good common ground and mental model, but on the other hand, it leads to less powerful feedbacks. There may be a notion that if one crew member is especially good in one aspect, he biases the rating for his colleague; or in other words: it can be assumed that specific items and performance aspects are not only rated better, but are actually better, if only one crew member performs very well.

References

References marked with an asterisk (*) indicate secondary literature from O’Connor et al. (2002)

Burger, K.-H., Neb, H. and Hörmann, 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.

Butler, R. (1991). Lessons from cross-fleet/cross-airline observations: Evaluating the impact of CRM/LOFT training. In R. S. Jensen (Ed.), *Proceedings of the 6th Symposium of Aviation Psychology* (pp. 326–331). Columbus, OH: Ohio State University.

*Connelly, P. (1997). *A resource package for CRM developers: Behavioral markers of CRM skills from real world case studies and accidents (Tech. Rep. No. 97-3).* Austin, TX.

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.

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, Ohio.

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

Flin, R., & Martin, L. (2001). Behavioral markers for crew resource management: A review of current practice. *International Journal of Aviation Psychology*, *11*, 95–118.

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

Gontar, P., Hoermann, H.-J., Deischl, J., & Haslbeck, A. (2014). How Pilots Assess Their Non-Technical Performance – A Flight Simulator Study. In T. Ahram, W. Karwowski, & T. Marek (Eds.), *Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics* . Krakow.

*Helmreich, R. L. (1984). Cockpit management attitudes. *Human Factors*, *26*, 583–589.

*Helmreich, R. L., Klinect, J. R., & Wilhelm, J. A. (1999). *The line operations safety audit (LOSA) observer's manual, version 7.0 (Tech. Rep. 99-0)*. Austin, TX.

*Helmreich, R. L., Wilhelm, J. A., Kello, J. E., Taggart, W. R., & Butler, R. (1990). *Reinforcing and evaluationg crew resource management: Evaluator/LOS instructor reference manual: Technical Manual 90-2.* Austin, TX.

Klinect, J. R., Murray, P., Merritt, A. C., & Helmreich, R. L. (2003). Line Operations Safety Audit (LOSA) - Definition and operating characteristics. In *Proceedings of the 12th International Symposium on Aviation Psychology* . Dayton, OH.

Law, J., & Wilhelm, J. (1995). Ratings of CRM skill markers in domestic and international operations. In R. S. Jensen (Ed.), *Proceedings of the 8th Symposium of Aviation Psychology* (pp. 669–675). Columbus, OH: Ohio State University.

O'Connor, P., Hoermann, H.-J., Flin, R., Lodge, M., & Goeters, K.-M. (2002). Developing a Method for Evaluating Crew Resource Management Skills: A European Perspective. *The International Journal of Aviation Psychology*, *12*(3), 263–285. doi:10.1207/S15327108IJAP1203_5

Van Avermaete, J. A. G. & Kruijsen, E. (1998). *NOTECHS - The evaluation of non-technical skills of multipilot aircrew in relation to the JAR-FCL requirements - Final Report NLR-CR-98443*. Amsterdam, Netherlands.

Contact Information

Patrick Gontar
Institute of Ergonomics
Technische Universität München
Boltzmannstr. 15
85747 Garching, Germany
Tel. +49.89.28915428
E-mail. gontar@tum.de