RELIABILITY OF INSTRUCTOR PILOTS' NON-TECHNICAL SKILLS RATINGS

Patrick Gontar
Institute of Ergonomics, Technische Universität München
Munich, Germany
Hans-Juergen Hoermann
Institute of Aerospace Medicine, German Aerospace Center (DLR)
Hamburg, Germany

This paper presents the results of different methods to assess reliability when instructor pilots rate pilots regarding their non-technical skills (NOTECHS). In preparation for a major inter-rater reliability study, this pretest analyzes the rating behavior of two instructor pilots during a full-flight simulator mission. Besides inter-rater reliability and test-retest reliability, the pilots' self-rating (n =12) and the instructors' point of view is analyzed. Results indicate a wide spread from poor to excellent reliabilities as a function of the different rating dimensions. Regarding inter-rater reliability, it is found that non-technical skills are rated more reliably under high workload conditions than under low workload conditions, and social aspects of non-technical skills are rated more reliably than cognitive aspects. Test-retest reliability is found to be .6 on average, whereas self-rating / instructor rating reliability is .5 on average. Based on these findings, implications for the major inter-rater reliability study will be derived and incorporated.

The importance of effective Crew Resource Management (CRM) has been known since the late 1970s, when NASA held their workshop on "Resource Management on the Flightdeck", and came to the conclusion that a majority of accidents are directly linked to interpersonal skills (Helmreich, Merritt, & Wilhelm, 1999; Dietrich, 2004; Gontar, Hoermann, Deischl, & Haslbeck, 2014). Consequently, adequate training methods and corresponding evaluation metrics were developed (O'Connor, Hoermann, Flin, Lodge, & Goeters, 2002). Although huge efforts were undertaken to train the raters, inter-rater reliability (IRR) is still an issue to be discussed. For example, Flin and Martin (2001), Law and Sherman (1995), Law and Wilhelm (1995), and Seamster, Edens, and Holt (1995) found different influencing factors that result in reduced inter-rater reliability in the aviation context. Sevdalis et al. (2008) and Yule et al. (2008) showed similar reduced reliabilities within the medical domain. In current airline practice, the trainer has to operate the simulator, simulate the air traffic controller, and assess the pilots during the mission – all at the same time. These circumstances make it worth analyzing the current evaluation practice in an airline to develop general recommendations to improve reliability of CRM ratings during training.

Background

The study presented here serves as a pretest in preparation for a study that aims to investigate the IRR of the most experienced instructor pilots (n = 45) when rating pilots' CRM skills within a major German airline. The goal of this pretest is to validate a flight scenario regarding general feasibility and its appropriateness for CRM ratings by instructor pilots. In addition to the instructor pilots' input, the self-assessment of the participating pilots will be taken into account as well. The findings will provide a rough estimation of the different reliabilities so the test design for the main study can be developed.

Research Questions

Based on the previous literature and the mentioned motivation for this study, the research questions (RQs) aiming for inter-rater reliability and test-retest reliability can be formulated as follows:

RQ 1 – Rating while operating: How reliable can two instructor pilots rate airline pilots' non-technical performance while operating a full flight simulator?

RQ 2.1 – Retest rating based on video recordings: How reliable can a rater asses pilots' performance based on a video recording?

RQ 2.2 – Self rating vs. instructor ratings: Which rating of the instructor (simulator or video-based) better reflects pilots' self-perception?

Method

Operationalization

In order to answer these questions, a full flight simulator scenario seems appropriate in order to have the same realistic environment as during normal simulator training missions. Further requirements are: captain and first officer as participants, no confederate pilot, realistic air traffic controller communication and noise (Schubert & Haslbeck, 2014), realistic unforeseen scenario (Casner, Geven, & Williams, 2013) with appropriate malfunctions to measure the influence of workload. Furthermore, the participants shall not be recruited on a volunteer basis but randomly selected to exclude any self-selection bias (Rosenthal & Rosnow, 2008).

To rate the pilots' CRM skills, it is important that the raters are already familiar with the evaluation tool. In this case, we use the company-adopted evaluation form, which uses the NOTECHS method (O'Connor et al., 2002). It was adapted to the airline's philosophy (Burger, Neb, & Hoermann, 2003) and is known to the two instructor pilots as well as to all participating pilots (important in terms of self-evaluation). The four dimensions measured on a five-point scale are defined as: *Communication, Leadership & Teamwork, Work Organization,* and *Situation Awareness & Decision Making.* In order to evaluate pilots' procedural and more technical skills, the Line Operations Safety Audit (LOSA) *Descent / Approach / Land* sheet (Klinect, Murray, Merritt, & Helmreich, 2003) is appropriate and measures *Planning, Execution, Review & Modify,* and *Overall Behavioral Markers* on a four-point scale. In contrast to the internal company evaluation form, the instructor pilots did not work with this LOSA sheet before.

When it comes to reliability measurements, one will find a lot of different metrics that can be computed. In the domain of evaluating non-technical skills ratings, intraclass correlation coefficients (ICCs) are commonly used as a measurement of reliability (Shrout & Fleiss, 1979). To assess systematic differences in the mean values, the ICC model can be adjusted to take those differences into account, called *absolute agreement.* Since both raters will rate every participant, a two-way random model can be applied (Wirtz & Caspar, 2002). The values of ICC can range between 0 and 1, where 1 represents fully explained variance; Landis and Koch (1977) postulated that values between .41 and .60 are moderate, values greater than .61 are substantial and values above .8 are almost perfect for kappa statistics. Fleiss, Levin, and Paik (2003, p. 604) stated that values "greater than 0.75 or so may be taken to represent excellent agreement". Current practice sets the cut-off value of reliability values for CRM ratings to .7 (Yule et al., 2008), as does this paper. To enhance reliability, it is possible to calculate the mean of two or more raters. With the help of the Spearman-Brown prophecy formula, the reliability of the average of *m* different raters can be calculated when the reliability of a single rater is known (Lienert & Raatz, 1994).

Test Design

Based on the research questions stated above, the test design can be visualized as shown in Figure 1. Both raters (rater 1 and rater 2), in this case the flight instructors, rate *n* subjects (pilots) during a full-flight simulator mission, while simultaneously operating the simulator and acting as air traffic controllers. In addition, all subjects rate themselves regarding their CRM performance (Gontar & Hoermann, 2014).

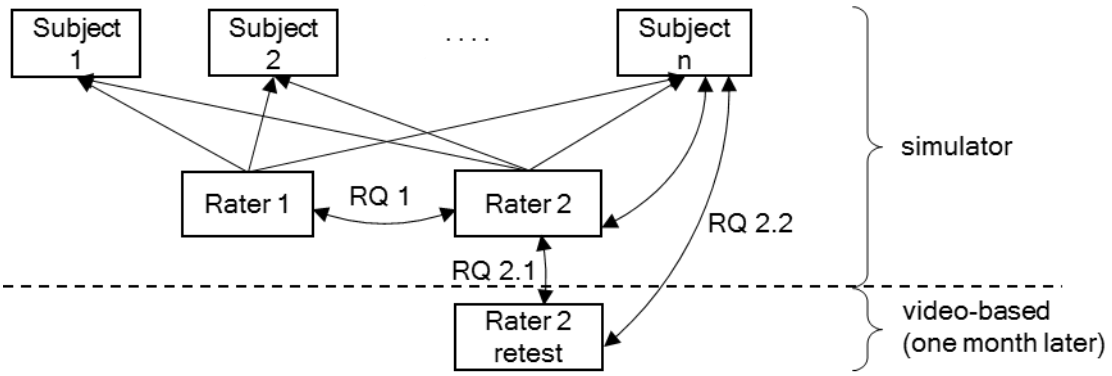

Figure 1. Test-design visualization.

The reliability between the two raters can be seen as inter-rater reliability (RQ 1). Four weeks after all simulator missions were completed, one of the raters (rater 2) assessed the performance of n = 8 pilots based on video and audio recordings again (retest). The comparison between the two ratings of rater 2 (simulation vs. video-based) is considered as test-retest reliability (RQ 2.1). The accordance between the self-ratings on the one side and the two time-delimited ratings of rater 2 are interpreted as the degree of accordance between self and external perception (RQ 2.2).

Experiment

In order to assess instructors' performance when assessing pilots' non-technical skills, it is important that a broad performance variance is shown by the participants. Within a full-flight simulator experiment, this means to induce a rather high workload with the help of an adequate scenario and malfunctions. The experimental scenario has to be sufficiently difficult, so that a proportion of pilots will not succeed and the corresponding wide range in performance can be observed.

Scenario. Before the experiment began, the pilots were informed about the aircraft's current state, including fuel on board, remaining flight time, navigation issues (e.g. approach details, maps), position and altitude via email two weeks in advance. After arriving at the simulator facilities, the pilot flying conducted his approach briefing. When the pilots entered the simulator, the aircraft was established on a visual approach under good weather conditions; fuel on board would suffice for 60 minutes at that time. Upon lowering the gear for the final approach, a malfunction was evoked which represented the leakage of the hydraulic system so that the nose gear was not able to fully extend and remained unlocked and unable to retract (malfunction 1). With this failure, the crew was forced to perform a go-around and work through the mandatory checklists and procedures. Due to the doubled aerodynamic drag, the fuel shortage meanwhile led to a mayday situation. With about 20 minutes of remaining flight time, the crews were now on their second approach. As a consequence of the damaged hydraulic system, the flaps and slats did not only extend slowly, but also jammed in their current position (malfunction 2); at that time, the high workload condition began. Again, the crew had to abort the approach and handle the procedures. At that point it was expected that about one half of the crews had to abort their trouble shooting process and force a landing with the current flight configuration.

Participants & Experimental Conduction. For this part of the experiment, 12 randomly selected pilots (6 Airbus A320, 6 Airbus A340) and therewith 6 crews flew the scenario one by one, while two instructor pilots operated the full flight simulator and acted as air traffic controllers in parallel. The pilots (Captains / First Officers) of the A320 fleet were M = 47/29, SD = 1.7/2.7 years old and had a total amount of M = 14,277/2,900, SD = 525/848 flight hours. The pilots of the A340 fleet were M = 52/37, SD = 1.7/3.3 years old and had experience of M = 17,667/9,354, SD = 1,699/2,248 flight hours; all the pilots hold valid ATP licenses with appropriate type ratings. The two flight instructors were both recently retired, M = 60 years old and had M = 21,000 hours of flight experience and served M = 20 years as instructor pilots within the same airline as the participating pilots.

The experimental conduction took place during two nights at a training facility, where three crews flew the scenario in a full flight simulator (*JAR STD 1A Level D*) each night. After the scenario was completed, the pilots left the simulator and both instructors independently rated the pilots' performance using the CRM evaluation form and the LOSA sheet mentioned above; the instructors were not allowed to talk to each other and the two participating pilots were separated into two rooms to conduct their ratings. One month later, one of the two instructors received the edited video and audio recordings of the scenario for retesting. At that point, the rater did not have any copies of his initial ratings. Since he rated another 72 pilots during the remaining experiment, it can be assumed that he did not remember particular ratings, but in the case of outliers, it is assumed he would probably recognize the pilot's behavioral patterns.

Results & Discussion

The results are presented in the order of the research questions stated; a short discussion of the particular aspects directly follows. Reliabilities are analyzed using intraclass correlation coefficients based on a two-way random model *ICC(2)* under the requirement of absolute agreement.

Results regarding the first research question (RQ 1), which refers to a classical inter-rater reliability problem, show the dependency of inter-rater reliability on different rating dimensions (compare Figure 2). When

looking at the CRM skills, only *Communication* and *Leadership & Teamwork*, which can be defined as social competencies, reach the cut-off level of .7, whereas, in contrast, the cognitive skills (*Work Organization* and *Situation Awareness & Decision Making*) do not. This could be explained by the fact that those social aspects can be observed directly and no further interpretation is necessary. *Work Organization* and *Situation Awareness* may require of the instructors more assumptions about observable behaviors, which could differ. These results are in accordance with Sevdalis et al. (2008), where *Communication* and *Teamwork* as well as *Leadership* achieve the highest reliabilities (.63 and .66) compared to the other categories. Results from the LOSA analyses show that only the *Planning Behavioral Markers* under the high workload condition reach the required reliability of .7; all other dimension are rated with a lower reliability. The reliabilities of the LOSA rating under the low workload conditions are significantly smaller than under the high workload conditions. It can be assumed that the spread of performance under high workload is more developed and therefore easier to rate. Both workload conditions reflect medium to high reliability regarding the *Planning Behavioral Markers* and very low reliabilities regarding *Review & Modify Behavioral Markers*.

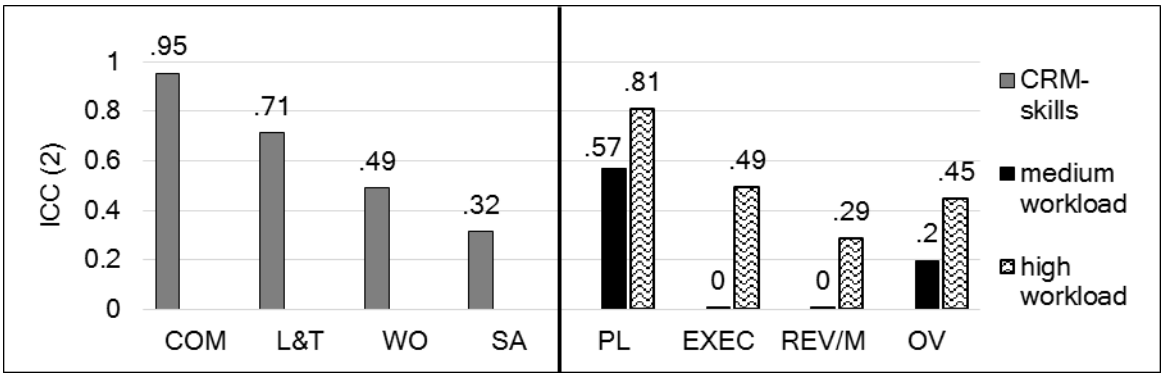

Figure 2. Inter-rater reliability between rater 1 and rater 2 as a function of different rating dimensions. COM = Communication, L&T = Leadership & Teamwork, WO = Work Organization, SA = Situation Awareness & Decision Making, PL = Planning, EXEC = Execution, REV/M = Review and Modify, OV = Overall.

In order to reach an acceptable level of reliability (.7), the Spearman-Brown prophecy formula (see Lienert & Raatz, 1994) was used to calculate the minimum number of raters required for a reliable rating of pilots during simulator missions. For this calculation, the single dimensions of the respective categories were averaged using *Fisher z' transformation* (Fisher, 1925). It is confirmed that for average non-technical skills ratings, one rater is sufficient. In comparison, for averaged LOSA ratings under the low workload condition, nine raters would be needed; under high workload conditions, two raters are sufficient. Applying this data to a required .9 reliability level, CRM ratings would need eight pilots. These findings are in accordance with Brannick, Prince, and Salas (2002), who postulated a need for nine raters for their comparable data on the .9 reliability level. Regarding research question 2.1, which aims for the test-retest reliability, results indicate, in contrast to the previous mentioned results, that this kind of reliability is higher for the LOSA categories than for the non-technical skills (compare Figure 3).

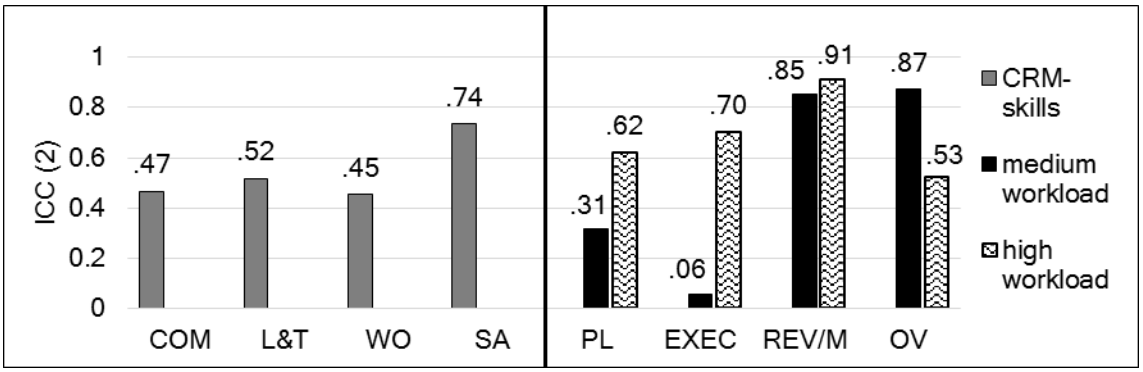

Figure 3. Test-retest reliability within rater 2.

Especially the test-retest reliability for the high workload condition leads to very high reliabilities in comparison to the inter-rater reliability. This means that the rating is consistent for one rater (high test-retest reliability), but nevertheless strongly differs between the raters (medium reliability). The personal interpretation of the instructor seems to induce more variance for the rating than the actual performance of the subject does.

In terms of agreement regarding pilots' self-estimation compared to an external point of view, RQ 2.2 delivers the following results (compare Figure 4):

1) The agreement highly depends on the dimension that is rated; aspects of *Situational Awareness & Decision Making* are rated with good reliability.
2) In three out of four dimensions (L&T, WO, and SA), the test rating (simulator) fits better with the self-evaluation than the retest rating (video-based). Only the category *Communication* is rated with slightly higher agreement during the retest rating.

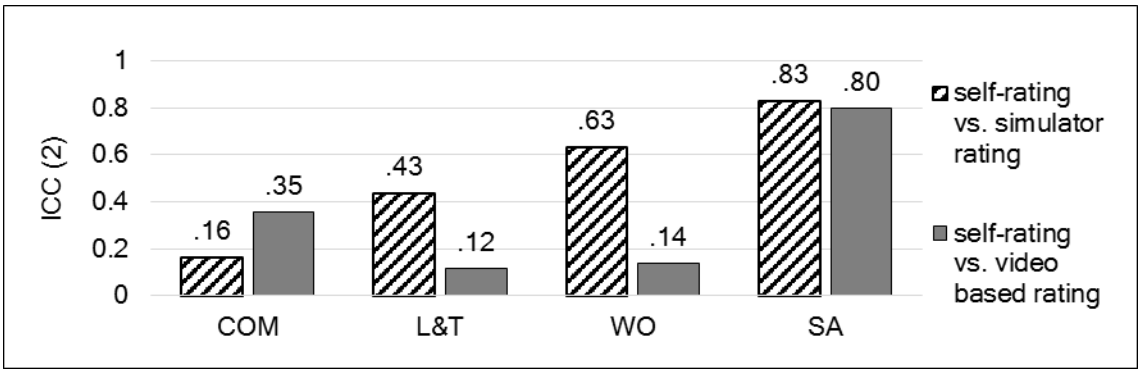

Figure 4. Self-rating (pilot) vs. simulator respectively video-based rating (rater 2).

Conclusion

When interpreting the results, it has to be kept in mind that the LOSA rating forms were new to the instructor pilots. In general, the results showed that reliability highly depends on the dimension that is rated and even a retest does not lead to higher congruency between pilots' self-estimation and instructors' ratings. This could mean that in video-based debriefing situations, where the instructor and the pilot have time to reflect the mission more often, the subjective perception can differ more between pilot and instructor than directly after the mission being completed. Furthermore, it seems that it can make sense to incorporate more instructor pilots in one assessment when it comes to specific rating dimensions.

Acknowledgements

The authors acknowledge the support of Cpt. Manfred Binder, Cpt. Peter Croeniger and Tanja Kammann, B.Sc. during data collection and handling. This work was funded by the German Federal Ministry of Economics and Technology via the Project Management Agency for Aeronautics Research within the Federal Aeronautical Research Program (LuFo IV-2).

References

Brannick, M. T., Prince, C., & Salas, E. (2002). The Reliability of Instructor Evaluations of Crew Performance: Good News and Not So Good News. *International Journal of Aviation Psychology, 12*(3), 241–261.

Burger, K.-H., Neb, H. & 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.

Casner, S. M., Geven, R. W., & Williams, K. T. (2013). The effectiveness of airline pilot training for Abnormal events. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 55*(3), 477–485.

Dietrich, R. (2004). Determinants of effective communication. In T. M. Childress & R. Dietrich (Eds.), *Group interaction in high risk environments*. Aldershot: Ashgate.

Fisher, R. A. (1925). *Statistical Methods For Research Workers*. Edinburgh: Oliver and Boyd.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). *Wiley series in probability and statistics*. Hoboken, N.J.: J. Wiley.

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

Gontar, P., & Hoermann, H.-J. (2014). Flight Crew Performance and CRM Ratings Based on Three Different Perceptions. In A. Droog (Ed.), *Aviation Psychology: facilitating change(s): Proceedings of the 31st EAAP Conference* (pp. 310–316).

Gontar, P., Hoermann, H.-J., Deischl, J., & Haslbeck, A. (2014). How Pilots Assess Their Non-Technical Performance - A Flight Simulator Study. In N. A. Stanton, S. J. Landry, G. Di Bucchianico, & A. Vallicelli (Eds.), *Advances in Human Aspects of Transportation* . AHFE International.

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.

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.

Landis, R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*(1), 159–174.

Law, J., & Sherman, P. (1995). Do raters agree? Assessing inter-rater agreement in the evaluation of aircrew resource management skills. In R. S. Jensen (Ed.), *Proceedings of the 8th Symposium of Aviation Psychology* (pp. 608–612). Columbus, OH: Ohio State University.

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.

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse* (6th ed.). Weinheim: Beltz, Psychologie Verl.-Union.

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.

Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis* (3rd ed). Boston: McGraw-Hill.

Schubert, E., & Haslbeck, A. (2014). Gestaltungskriterien für Szenarien in Flugsimulatoren zur Untersuchung von Verhalten und Leistung von Verkehrspiloten. In Deutsche Gesellschaft für Luft- und Raumfahrt - Lilienthal-Oberth e.V. (Ed.), *DGLR-Bericht, 2014-01. Der Mensch zwischen Automatisierung, Kompetenz und Verantwortung* (pp. 125–137). Bonn.

Seamster, T., Hamman, W., & Edens, E. (1995). Specification of observable behaviors within LOE/LOFT event sets. In R. S. Jensen (Ed.), *Proceedings of the 8th Symposium of Aviation Psychology* (pp. 663–668). Columbus, OH: Ohio State University.

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. doi:10.1016/j.amjsurg.2007.08.070

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin, 86*(2), 420–428.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.

Yule, S., Flin, R., Maran, N., Rowley, D., Youngson, G., & Paterson-Brown, S. (2008). Surgeons' Non-technical Skills in the Operating Room: Reliability Testing of the NOTSS Behavior Rating System. *World Journal of Surgery, 32*(4), 548–556.