Gontar, P. & Hoermann, H.-J. (2015). Inter-rater reliability at the top end: Measures of pilots' non-technical performance. The International Journal of Aviation Psychology, 25(3/4), 171-190.

# Inter-Rater Reliability at the Top End – Measures of Pilots' Non-Technical Performance

Running head: Inter-Rater Reliability at the Top End

Patrick Gontar & Hans-Juergen Hoermann

Abstract

**Objective.** The aim of this study is to analyze influences on inter-rater reliability and within-group agreement within a highly experienced rater group when assessing pilots' non-technical skills.

**Background.** Non-technical skills of pilots are crucial for the conduct of safe flight operations. To train and assess these skills, reliable expert ratings are required. Literature shows to some degree that inter-rater reliability is influenced by factors related to the targets, scenarios, rating tools, or to the raters themselves.

**Method.** Thirty-seven type-rating examiners from a European airline assessed the performance of four flight crews based on video recordings using LOSA and adapted NOTECHS tools. We calculated $r_{wg}$ and *ICC(3)* to measure within-group agreement and inter-rater reliability.

**Results.** The findings indicated that within-group agreement and inter-rater reliability were not always acceptable. Both metrics showed that outstanding pilots' performance was rated with higher within-group agreement. For cognitive aspects of performance, inter-rater reliability was higher than for social aspects of performance. Agreement was lower on the pass/fail level than for the distinguished performance scales.

**Conclusion.** These results suggest to back pass/fail decisions not exclusively on non-technical skill ratings. We furthermore recommend that regulatory authorities more systematically address inter-rater reliability in airline instructor training. Airlines as well as training facilities should be encouraged to demonstrate sufficient inter-rater reliability when using their rating tools.





*This is an Accepted Manuscript of an article published by Taylor & Francis in International Journal of Aviation Psychology on 10 May 2016, available online: http://www.tandfonline.com/10.1080/10508414.2015.1162636.*

*Keywords:* inter-rater reliability, within-group agreement, non-technical skills, NOTECHS, LOSA

## Introduction

In-depth accident investigations in the 1970s highlighted the fact that the non-technical behaviors of pilots, like leadership, communication, teamwork and decision making, had clearly been neglected as significant factors for safe flight operations (Cooper, White, & Lauber, 1980). In succession, several approaches attempted to systematically include cockpit/crew resource management (CRM) in pilot training (Helmreich, Merrit, & Wilhelm, 1999). For the evaluation of training success, CRM-skills assessment became relevant. Goldsmith and Johnson (2002) name three major reasons why such an evaluation is important and how it can improve pilot performance: to judge if the *pilot is proficient* enough to fly in the respective airline, to give *sufficient and appropriate performance feedback* to the pilot, and to develop and modify the *airline's training program*. Regulatory authorities have provided standards and guidelines for the instruction and assessment of CRM by the airlines (cf. European Aviation Safety Agency, 2011, 2014; European Commission, 2011; Federal Aviation Administration, 2004; Joint Aviation Authorities, 2001). In this context, Robert Helmreich and his colleagues at the University of Texas were very influential, developing behavioral marker systems and other observation methods such as the Line/LOS Checklist for the aviation industry (Helmreich, Klinect, Wilhelm, & Jones, 1999). For European airlines, a rating system called NOTECHS became the standard (Flin et al., 2003; O'Connor, Hoermann, Flin, Lodge, & Goeters, 2002). Because behavioral marker and rating systems are subject to observation bias, aspects of inter-rater reliability (IRR) and inter-rater agreement (IRA) became important (Brannick & Prince, 1991; Brannick, Prince, & Salas, 2002). To ensure that pilots are trained to a required level of competence, reliability of the competence assessment is a vital precondition (cf. Nunnally & Bernstein, 1994).

While "…practical elements [of instructor training] should include the development of specific instructor skills, particularly in the area of teaching and assessing threat and error management and CRM" (European Aviation Safety Agency, 2011, FCL.920, p. 282), high IRR and IRA lead to transparent and traceable ratings and can therefore enhance the feedback during the debriefing, and thus the training quality (Gontar & Hoermann, 2014). Unreasonably harsh or unreasonably lenient ratings can not only lead to economic drawbacks, but also to critical safety consequences (Holt, Hansberger, & Boehm-Davis, 2002). For example, raters using overly harsh standards may give rise to needless additional training costs for the airline. An overly harsh rating in an examination flight could jeopardize the pilot's license and have a negative effect on his or her motivation without reasonable cause. The opposite is the case if the raters have pilots passing an examination although they performed below the required minimum. In the latter case, degrading standards will have safety implications.

Studies in aviation (Brannick et al., 2002; Holt et al., 2002; O'Connor et al., 2002; Williams, Holt, & Boehm-Davis, 1997), in air traffic control (Kontogiannis & Malakis, 2013), and in medical domains (Arora et al., 2011; Beard, Marriott, Purdie, & Crossley, 2011; Cooper, Endacott, & Cant, 2010; Dedy et al., 2015; Fletcher et al., 2003; Gale et al., 2010; Mitchell et al., 2012; Sevdalis et al., 2008; Yule et al., 2008; Yule et al., 2009) found that professional raters have different views when rating practitioners on their non-technical skills (NTS). The important research tasks in this context are obviously to identify the conditions under which the views of the raters tend to diverge or converge and to apply the outcomes to the improvement of inter-rater reliability. Based on the aforementioned studies, we divided the factors that influence inter-rater reliability into four major themes (comparable to Brannick et al., 2002). These themes are: *target*-related (e.g. target person's level of performance, target person's position in crew), *scenario & task*-related (e.g. taxiing, emergency procedures, cruise flight, approach), *measurement*-related (e.g. rating dimension, scale level, observable markers, anchors), and *rater*-related (e.g. experience, familiarity with rating tools, motivation). It is pointed out that these themes can also be interdependent.

## ***Target*-related Influences**

Regarding *target*-related influences, O'Connor et al. (2002) reported that captains (CPTs) were rated less accurately than first officers (FOs). Mishra, Catchpole, and McCulloch (2009) found slight differences between targets when analyzing IRA for nurses, surgeons, and anesthetists. Yule et al. (2009) found that it is easier to rate targets who perform very well or very poorly than crews whose performance is in the medium range, since extreme behaviors are normally more salient. As average-performing crews represent the majority of cases in reality, it is very important to train inter-rater reliability when rating those (Yule et al., 2009). In addition, these authors pointed out the problem that a target's performance may vary on the same dimension (e.g. communication) during the observation period. In this case, it is hard to decide how to weigh the different characteristics and to arrive at a final grade.

### *Scenario & Task*-related Influences

The second identified theme, *scenario & task,* is addressed by O'Connor et al. (2002). They were able to show that the content of flight scenarios and tasks influence inter-rater reliability and identified the crucial aspects in their specific scenarios. One major fact they asserted was the difficulty for the rater to "decide how to separate the behaviors and responsibilities of the two pilots" (O'Connor et al., 2002, p. 282). Yule et al. (2008) conducted a study with six different scenarios and found similar results, but attributed them to the special behaviors of the crews, which they stated were easier to rate. Mitchell at al. (2012) explained differences in inter-rater reliability between the scenarios as being affected by the short and variable duration of the scenarios. They furthermore suggested that the semi-scripted scenarios might have influenced inter-rater reliability due to the varying quality of the actors.

### *Measurement*-related Influences

Dedy et al. (2015), Mishra et al. (2009), O'Connor et al. (2002), and Yule et al. (2008) found that within-group agreement and inter-rater reliability also depend on the rated dimension. O'Connor et al. (2002) and Yule et al. (2008) reported that interpersonal skills (e.g. communication) were rated in higher agreement than cognitive skills (e.g. decision making). Yule et al. (2008) attributed this effect to the raters, who only had 2.5 hours of training and were not educated in the underlying cognitive models. In contrast, Yule et al. (2009) found an opposite effect: Cognitive skills were rated in higher agreement than social skills. We found this same effect when analyzing pilots' peer and self-rating behaviors (Gontar & Hoermann, 2014). Social aspects such as communication, leadership, and teamwork were rated with lower inter-rater reliability than cognitive aspects, for example work organization, situation awareness, and decision making (Gontar & Hoermann, 2014). We concluded that this effect was due to the scenario, where the successful technical outcome was strongly related to good decision making. Brannick et al. (2002) analyzed the influence of item generality on reliability and found that interjudge agreement was higher for specific behaviors than for a general assessment of CRM in total.

### ***Rater*-related Influences**

Regarding the influence of the rater, Hamman and Holt (1997) found that factors such as personal interpretation and motivation influence and bias performance ratings (see Flin & Martin, 2001). Yule et al. (2008) argued that rating bias also depends on the expertise of the raters in their specific field. They analyzed the average reliability of different rater groups (in this case: general surgeons vs. orthopedic surgeons) and found variance in the agreement, suggesting that "…surgeons' ratings might be more homogeneous when they are rating scenarios based in their own specialty than when rating other specialties" (p. 552). In 2009, Yule et al. showed that prior rating experience can affect rating standards. They compared novice raters with expert raters and found that novices tend to rate more harshly than experts. As possible reasons for low reliability, Weber, Roth, Mavin, and Dekker (2013) suggested that raters might not recognize the same behaviors, or even if they do, they do not evaluate them equally. In a follow-up study, Weber, Mavin, Roth, Henriqson, and Dekker (2014) analyzed the degree to which raters gave different reasons (justifications) for their grading of pilots' behavior. They clustered similar justifications into *topics* and were able to show that raters use different *topics* to assess specific performance categories.

### **Research Needs**

Even though IRR of performance ratings is influenced by the above mentioned factors, check and training practices of the airlines have to rely on instructor pilots assessing the technical and non-technical skills of their trainees. This is primarily done through observation. Several studies have attempted to improve reliability and structural validity of the crew assessment by intensified instructor training (e.g. Holt et al., 2002) or by improving tools for rating non-technical skills (e.g. Sevdalis et al., 2008). Holt et al. (2002) looked at the development of IRR over a period of three years with rater training. They found generally acceptable results, but could not identify strong improvement over the years. These authors mentioned that due to the small number of raters in the beginning, turnover in the group of raters may have affected the group's rating performance. Furthermore, they noted that the rated scenarios differed from year to year. Sevdalis et al. (2008) analyzed the IRR of raters after revising their NTS rating tool. Even after these revisions, specific dimensions such as *cooperation and team skills* showed barely adequate reliability. However, they could not rule out that a lack of familiarity with the revised definitions led to lower IRR for that dimension.

In contrast to the studies cited here, this study kept the influence of the *raters* and the influence of the *scenario & task* constant, which allowed us to focus on the influence of *target* and *measurement*. Raters and scenarios were kept unchanged by selecting a homogeneous and very experienced group of type-rating examiners from the same airline and showing them videos with different flight crews performing the same flying tasks. We examined how reliably these raters, who worked for the same airline as the crews, used different rating tools to assess the pilots.

To our knowledge, no study has been published in which a large group of raters with homogeneous experience, education, and affiliation participated, in order to keep *rater*-induced effects constant. Furthermore, no study was found that kept the influence of *scenarios & tasks* constant across different crews. In this study we asked the instructor pilots to assess actual flight crews from the same airline in realistic simulator scenarios containing the same task for each crew. Such a situation is very common in reality: All pilots in an airline have to fly the same missions in training and examination flights. In addition, most of the previous studies had either volunteer raters (e.g. Mitchell et al., 2012; Yule et al., 2009) or did not specify how raters had been recruited (e.g. Fletcher et al., 2003; O'Connor et al., 2002). We suggest that using volunteers, and thus self-selected raters, would potentially bias the ratings and therefore would not reflect the daily practice. Normally, instructors and pilots are assigned to their specific training or check missions.

In our study, we expect similar results for within-group agreement as reported by O'Connor et al. (2002). These authors found an average $r_{wg}$ of .76 across all the different rating dimensions at the category level of NOTECHS, which is comparable to our $NTS_{dim}$ measurement (see dependent measures). O'Connor et al. (2002) showed that the agreement varied for different scenarios (from .64 to .87). In our study, the scenario remained unchanged. However, different crews exhibited the full range of performance, from *outstanding* to *poor*. We expect that raters show higher agreement for extreme performance than for average performance, because extreme performance is assumed to be more salient (Yule et al., 2009).

To summarize the above, the research questions addressed by this paper arise from the two major themes that influence inter-rater reliability: *target* and *measurement*. With respect to the *target*-related influences, we investigate differences in the ratings for the two crew members (CPT and FO) in relation to their level of performance. In terms of the *measurement*-related factor, we analyze the influence of the familiarity with the tools, the tools' dimensions, and the scale levels of the tools. We keep the influences from the *raters* and the *scenario & task* constant by choosing the best raters and using the same flight scenarios for all pilots.

# Method

## Participants

A sample of 37 type-rating examiners (TREs) from a major European airline, all holding valid licenses for the Airbus A320, took part as raters in this experiment. Their participation was not voluntary, since they were assigned to this rating experiment as part of a workshop. Due to their specific training and certification, the examiners are the most experienced instructors for this aircraft type within this company. They represent a homogeneous group with regard to their affiliation and experience. The mean age of the participants was 49.9 years (*SD* = 4.2 years). They had a mean experience of 11.5 (*SD* = 4.3) years as training and check pilots, and had a mean number of 13,604.2 (*SD* = 3,900.1) airline flight hours.

As part of their initial and recurrent instructor courses, all participants had received several days of theoretical and practical training for their rating skills. Rating exercises were done with video examples in classrooms as well as during real training sessions in the simulator. As part of the training, instructors received feedback on their individual rating tendencies. In addition, they participated in annual standardization meetings, which contain specific case study exercises. Since the simulator scenario was new and not previously included in routine recurrent trainings by the airline, none of the raters had specific experience with the presented simulator scenarios – neither as participating pilot nor as instructor pilot.



## Apparatus

The 37 raters assessed videotapes of the same flight scenario flown by four different crews in an A320 simulator; the different videotapes were presented on a screen using a projector in a classroom to all raters at the same time. For the purpose of de-identification, the pilots' voices in the videotapes were modified by changing the pitch; dialogs were still clearly understandable.

## Recorded Simulator Mission



*This is an Accepted Manuscript of an article published by Taylor & Francis in International Journal of Aviation Psychology on 10 May 2016, available online: http://www.tandfonline.com/10.1080/10508414.2015.1162636.*

The presented videotapes were extracted from a mission in a flight simulator study conducted with (*n* = 60) short-haul pilots on the Airbus A320 (see Gontar & Hoermann, 2014; Gontar, Hoermann, Deischl, & Haslbeck, 2014). The flight simulator mission aimed to analyze the pilots' behavior in unforeseen situations under high workload. Therefore, the pilots showed authentic non-scripted behaviors. This experiment was conducted in a full-flight simulator (*JAR STD 1A Level D*), but was not part of pilots' recurrent training within the airline.

The selected videotapes for the inter-rater reliability study show a sequence with high task load for the pilots. In the flight simulator mission, the crews began a visual approach (VOR B) to runway 22R at Nice Côte d'Azur Airport (LFMN), 15 miles east of the airport (D15 AZR) at an altitude of 3,000 feet with a speed of 170 knots, and a heading of 269 degrees; there was light rain, the runway was wet, visibility was 10,000 meters, wind was 10 knots from the south, and the temperature was 12° Celsius. The aircraft had 2,500 kg fuel on board (corresponding to a remaining flight time of approximately one hour) and was adequately set for the approach.

When the crew lowered the gear the green hydraulic system malfunctioned and prevented the nose gear from fully extending and locking; it could not be retracted. As a consequence, the crew had to go-around and follow several procedures. With the aerodynamic drag being doubled, flight endurance was halved (approximately 30 min). In their subsequent approach the crew was already in a mayday situation. Upon selecting the next flap level, due to the underlying failure of the green hydraulic system, the flaps or the slats (depending on the initial configuration) jammed. The high task load condition for the rating experiment began at this point. As the malfunction affected the landing performance of the aircraft, the crew was again requested to handle several procedures before they were able to land. For further details on the technical scenario, the reader is referred to Gontar and Hoermann (2014).

The scenario was very challenging and elicited the pilots' CRM skills on all the dimensions that are usually trained and rated during recurrent training. These dimensions include communication skills, leadership and teamwork, work organization as well as situation awareness and decision making. Since the malfunctions that occurred were unforeseen for the pilots, they were not able to prepare themselves beforehand, but had to make fast decisions, very efficient task assignments, and also handle the procedures, the automation and checklists with particular precision. Normally, the crews have enough time to work through their procedures step by step.

However, the fuel problem in our scenario forced them to work through the procedures more quickly and thus communicate more concisely and more effectively. Furthermore, the success of this mission was highly dependent on making the proper decisions in the right order (e.g. aborting or skipping procedures or declaring an emergency due to the very low fuel level). It was expected that only crews with high CRM skills would be able to complete this mission satisfactorily.

During the 30 flight simulator missions, we recorded audio data from the pilots and the ATC, flight simulator data, as well as video data showing the two participating pilots and the cockpit interior from behind. Pilot performance was rated by a flight instructor from the respective airline during the missions (benchmark rating). This benchmark rating was based on the evaluation form which is used in this airline (Burger, Neb, & Hoermann, 2003) and is explained below ($NTS_{item}$). We found high variance within the pilots' performance ratings. The ratings included crews that were able to manage the severe technical problems very quickly and very well, but also crews which were unable to deal with the problems. Based on the averaged performance grade of the benchmark ratings, we selected four videotapes that reflect the entire spectrum of CRM performance: *poor*, *medium-low*, *medium-high*, and *outstanding.* To validate the benchmark rating, the videotape selection was verified by an additional type-rating examiner.

**Dependent Measures**

The raters assessed the pilots' performance based on videotapes using three different rating tools: two NTS rating tools – one on a dimension basis, one on an item basis (Burger et al., 2003) – and the *Line Operations Safety Audit* (LOSA) rating tool (Klinect, Murray, Merritt, & Helmreich, 2003). Examples of the content for each tool are shown in Figure 1.

---

Insert Figure 1 around here

---





*This is an Accepted Manuscript of an article published by Taylor & Francis in International Journal of Aviation Psychology on 10 May 2016, available online: http://www.tandfonline.com/10.1080/10508414.2015.1162636.*

**NTS rating tool on dimension basis ($NTS_{dim}$).** The raters assessed each individual pilot's performance using a five-point scale ranging from *poor* (1) to *outstanding* (5), see Figure 1 at the top. The following dimensions were addressed: *communication, leadership & teamwork, work organization,* and *situation awareness & decision making* (Burger et al., 2003). *Communication* and *leadership & teamwork* are regarded as social aspects; *work organization* and *situation awareness & decision making* are regarded as cognitive aspects (Hoermann & Neb, 2004). This rating method requires the instructor to assess the pilots' performance globally across the whole videotape, but is not based on single items (Brannick et al., 2002). The raters themselves have to relate specific crew behaviors to the various $NTS_{dim}$ dimensions. Such rating methods require a higher degree of abstraction and are expected to be more subjective and thereby less reliable than directly observable behaviors (Brannick et al., 2002).

**NTS rating tool on item basis ($NTS_{item}$).** The raters assessed each individual pilot's performance on 40 items which reflect the same four dimensions as $NTS_{dim}$, but support the rater with more specific items (Burger et al., 2003), see the middle of Figure 1. The 40 items represent the dimensions *communication* (10 items), *leadership & teamwork* (15 items), *work organization* (8 items), and *situation awareness & decision making* (7 items), and were rated on the same five-point scale, ranging from *poor* to *outstanding*, as $NTS_{dim}$. The mean value of all items of a dimension was calculated to obtain a value comparable to $NTS_{dim}$ but based on items. The items of this rating tool were known to the raters and are regularly used in their airline's training. It is based on the NOTECHS method (Flin et al., 2003) and was adapted to the company's culture and CRM philosophy by a working group comprised of subject matter experts, such as training and check pilots, aviation psychologists, and human factors specialists (Burger et al., 2003). This work was influenced by the results of a safety survey that the airline conducted. The purpose of this study was to analyze safety-relevant events from the preceding five years. Based on this survey, Burger et al. (2003) were able to identify specific factors that contributed to the events and translated them to markers. A content analysis was performed to ensure that all NOTECHS markers were covered by the newly developed system.

**Line Operations Safety Audit.** The raters assessed the pilots' performance on four dimensions using 13 items that represent *planning behavioral markers* (4 items), *execution behavioral markers* (4 items), *review / modify behavioral markers* (3 items), and *overall behavioral*





*This is an Accepted Manuscript of an article published by Taylor & Francis in International Journal of Aviation Psychology on 10 May 2016, available online: http://www.tandfonline.com/10.1080/10508414.2015.1162636.*

*markers* (2 items). Ratings were obtained on a four-point scale from *poor* (1) to *outstanding* (4), see Figure 1 at the bottom. The rating of a dimension is given by the mean of all its item values. As Haeusler, Klampfer, Amacher, and Naef (2004) demonstrated, dimensions of LOSA strongly correlate with dimensions of the NOTECHS system, which was the basis of the $NTS_{item}$ system used here (Burger et al., 2003). In addition, LOSA incorporates aspects of technical skills as well. Since technical aspects such as *automation handling* are more overt and observable, we expect higher inter-rater reliability for the LOSA rating. The LOSA rating tool was sent to the raters two weeks before the rating experiment, but they were not familiar with it. We used the *LOSA Descent / Approach / Land* sheet (International Civil Aviation Organization, 2002, p. A-8; Klinect et al., 2003). In contrast to the NTS ratings, both crew members were rated together as a team.

In addition to the five-point scale of both NTS tools, we derived a dichotomous pass/fail scale by assigning the lower two scale points to *fail* and the upper three scale points to *pass.* Regarding the LOSA rating, we assigned the lowest scale point to *fail* and the upper three scale points to *pass*; if one item was rated as failed, the whole rating dimension was deemed unsatisfactory.

**Instructions and Procedure**

Two weeks before the rating experiment was conducted, all raters were informed about the upcoming assessment. They were briefed about the technical details of the scenario, such as the expected route, weather conditions, aircraft configuration, malfunctions, etc. Furthermore, the raters received a copy of the three rating forms they would have to use (see Figure 1 with examples of the content). While the $NTS_{item}$ rating tool was already known to the instructor pilots from their current training practice, the $NTS_{dim}$ and the LOSA tools had not been used by the airline before.

Immediately before the rating experiment began, we once again explained the rating tools and the whole scenario to the raters. The raters were explicitly advised to leave items blank if they did not observe a corresponding behavior. In addition, the raters were instructed not to talk to each other and were told that they were not allowed to page back in the rating sheets – neither during the rating process itself, nor between rating the different crews. The sheets were then handed out.

Following these instructions, the four videotapes with a duration of $M$ = 5.98 ($SD$ = 1.42) minutes each were presented one by one. Rewinding or repeating was not an option; however, raters could take notes. The videotapes were presented in the following order of crew performance:

(1) *medium-high*, (2) *medium-low*, (3) *outstanding*, and (4) *poor*. In order to minimize sequence effects, the medium-performing crews (1; 2) were presented first. Between the presentations, the instructors rated the pilots' skills; all raters had as much time as they wanted to complete their ratings. This took approximately 20 minutes after each scenario. The videotape began with a map showing the current location of the aircraft, speed, heading, and the remaining fuel on board for a duration of 30 seconds. The actual flight scenario was then presented, starting exactly 30 seconds before the second malfunction occurred (flaps or slats jammed) and the high task load condition began.

**Analysis**

For each of the dependent measures, we calculated $r_{wg}$ (cf. James, Demaree, & Wolf, 1984) to assess within-group agreement for the five-point and four-point scales, and on the pass/fail level. "The technique [of $r_{wg}$] was cast as a heuristic form of interrater reliability…" (James, Demaree, & Wolf, 1993, p. 306) and sees total variance (in contrast to classical test theory) as being related to the rater (cf. Liao, Hunt, & Chen, 2010). Values of $r_{wg}$ were calculated for each rating dimension of the respective tool and for each performance level of the crew. This allowed us to identify the *measurement*-related and *target*-related influences on inter-rater reliability. In addition, for $NTS_{dim}$ and $NTS_{item}$, $r_{wg}$ was calculated separately for the CPTs' and the FOs' performance ratings. As the crew was assessed as one team by LOSA, a comparison between the crew members was not possible. The threshold value for acceptable within-group agreement was set to .70 (Nunnally & Bernstein, 1994), so that agreement equal to or above .70 is interpreted as *acceptable*, and values below .70 are interpreted as *not acceptable* agreement; for an in-depth discussion about this commonly used criterion, see Harvey and Hollander (2004).

Intraclass correlation coefficients *ICC(3)* for single measures were calculated using a two-way mixed model for each of the dependent measures to assess inter-rater reliability (cf. Shrout & Fleiss, 1979). The factor *rater* was determined as fixed since the raters were preselected for the workshop. In contrast to most of the studies in the medical domain, average measures of *ICC(3)* do not seem appropriate here, since the reliability of one single rater, and thus the single measure, is decisive. This is due to the fact that only one instructor pilot rates the crew's performance in real training and examination flights. *ICC(3)* analyses were conducted at the five-point and four-point scale level only (not on the derived dichotomous level). ICCs were calculated for all the different rating dimensions of the two NTS measurements and the LOSA measurement.

Although ICCs can be calculated for true dichotomous data, a subsequently derived dichotomous level from a higher level scale (as shown here) would require tetrachoric correlation coefficients (Wirtz & Caspar, 2002). Based on our data set, it was not possible to calculate tetrachoric correlation coefficients due to missing data and the resultant singularities. According to Cicchetti (1994) who subdivided the recommendation of Landis and Koch (1977), the *ICC(3)* values are interpreted as follows: Values below .40 represent *poor* clinical significance, values between .40 and .59 represent *fair*, and values between .60 and .75 *good* clinical significance. Values greater than .75 are considered *excellent* clinical significance (Cicchetti, 1994); see also Fleiss, Levin, and Paik (2003). When calculating mean values of ICCs, *Fisher z' transformation* (Fisher, 1925) was used. With respect to the ICCs, the Spearman-Brown prophecy formula (cf. Lienert & Raatz, 1998) was used to calculate the minimum number of raters that is required to achieve a specific level of reliability (e.g., .60 for good clinical significance according to Cicchetti, 1994).

According to the model assumptions for *ICC(3)* made by Shrout and Fleiss (1979), we used Shapiro-Wilk tests to analyze for normal distribution as suggested by Thode (2002) and Razali and Wah (2011). Analyses showed that the sample data were non-normally distributed ($p < .05$), except for the LOSA dimensions *execution behavioral markers,* $W(143) = .99$, $p = .39$, and *overall behavioral markers,* $W(143) = .98$, $p = .06$. Based on the visual inspection of the plots, we concluded that the significant results in the tests were rather due to the large sample size than due to meaningful deviations from the normal distribution (Field, 2009). The residuals showed non-normal distributions ($p < .05$) as well, except for the $NTS_{item}$ dimensions *communication,* $W(288) = .99$, $p = .45$, *leadership & teamwork* $D(288) = .99$, $p = .20$*, situation awareness & decision making* $W(288) = .99$, $p = .72$, and for the LOSA dimension *review / modify,* $W(136) = .99$, $p = .24$. Since the *analysis of variance,* which corresponds to the *ICC(3)* model, is robust against violations of normal distribution (Schmider, Ziegler, Danay, Beyer, & Buehner, 2010), we did not anticipate problems in using ICCs for these data.

## Results

### Results Regarding the $NTS_{dim}$ Tool


*This is an Accepted Manuscript of an article published by Taylor & Francis in International Journal of Aviation Psychology on 10 May 2016, available online: http://www.tandfonline.com/10.1080/10508414.2015.1162636.*

Table 1 illustrates the results with respect to the analysis of $NTS_{dim}$. Regarding the within-group agreement $r_{wg}$ on the five-point scale, the results showed acceptable agreement for the CPT (.79 on average) and for the FO (.74 on average) in the videotape that showed crew members with *outstanding* performance. The performance of all other pilots was rated with an agreement lower than .70 and was therefore not acceptable. When looking at the average of the different rating dimensions, it can be seen that the agreement across all videotapes was below the .70 threshold for every dimension. In three out of four videotapes, the FOs' performance was rated in higher agreement than the CPTs'; the average agreement across all videotapes (.57) was not acceptable. It can be concluded that the agreement of raters depended on the level of performance that was exhibited by the pilots.

*ICC(3)* for inter-rater reliability was found to be poor for the dimensions *communication* (.12), *leadership & teamwork* (.28), and *work organization* (.34) of $NTS_{dim}$. Only ratings for *situation awareness & decision making* (.45) represented fair reliability. Inter-rater reliability of social aspects (*communication* and *leadership & teamwork*) was lower than for cognitive aspects (*work organization* and *situation awareness & decision making*). The average inter-rater reliability was poor (.30). In order to reach a good level of reliability with respect to the ICCs (.60), the Spearman-Brown prophecy formula revealed that on average, four raters would be required to assess pilots' non-technical skills on the dimensional level for such scenarios.

---

Insert Table 1 around here

---

Looking at the within-group agreement results from the derived pass/fail scale (see Table 1, bottom), once again the *outstanding* performing crew was the only one which was represented by acceptable ratings for the CPT (.97) and for the FO (.92). While the ratings for this crew include six ratings that were in perfect agreement (1.0), the *medium-high* and *poor* performing crews included ratings that showed no agreement (0.0). The mean $r_{wg}$ of the rating dimensions were all below the




*This is an Accepted Manuscript of an article published by Taylor & Francis in International Journal of Aviation Psychology on 10 May 2016, available online: http://www.tandfonline.com/10.1080/10508414.2015.1162636.*

required minimum of .70. Comparing the pass/fail scale with the five-point scale, the *outstanding* performing crew was rated in higher agreement on the pass/fail scale (.97 / .92 vs. .79 / .74). The opposite was true for the *medium-high* and *poor* performing crews. The average agreement was lower on the dichotomous pass/fail scale (.46) than on the five-point scale (.57).

**Results Regarding the $NTS_{item}$ Tool**

Within-group agreement $r_{wg}$ showed acceptable ratings for the *poor* (.71 / .71) and *outstanding* (.77 / .80) performing crews as well as for the FO of the *medium-low* (.78) performing crew (see Table 2, top). All dimensions of $NTS_{item}$ showed acceptable agreement, although they were close to or at the threshold of .70. The trend indicated in the results of $NTS_{dim}$, i.e. that the FOs' performance was rated as slightly more in agreement than the CPTs' was seen here as well. A comparison of the mean of the dimensions shows that $NTS_{item}$ ratings were in higher agreement than $NTS_{dim}$ ratings on the five-point scale.

As already measured for $NTS_{dim}$, *ICC(3)* reliability was fair for *situation awareness & decision making* (.48), but poor for all the other dimensions; the trend that social aspects are rated less reliably than cognitive aspects was reflected in these results as well. Comparing the $NTS_{item}$ and the $NTS_{dim}$, it can be seen that the inter-rater reliability for *communication* was higher for $NTS_{item}$ than for $NTS_{dim}$. With respect to the ICCs, the Spearman-Brown prophecy formula found that on average, three raters would be sufficient to assess pilots' non-technical skills on this five-point scale with good (.60) reliability.

---

Insert Table 2 around here

---

The pass/fail scale showed unacceptable agreement and thus lower agreement than the five-point scale in every single value (see Table 2, bottom). In contrast to the five-point scale, the average agreement for the FOs' performance ratings was lower than for the CPTs' performance ratings. The average agreement using the $NTS_{dim}$ tool (.46) was higher than the average $NTS_{item}$ agreement (.19) on the pass/fail scale.

### Results Regarding the LOSA Tool

Results regarding the LOSA rating showed acceptable within-group agreement for the *planning* (.74) and *execution* (.76) dimensions, but agreement below the defined .70 threshold for the *review / modify* (.63) and *overall* (.61) dimensions (see Table 3, top). Rating for the *outstanding* (.74) and *medium-low* (.71) performing crews showed acceptable agreement on average. Agreement for *poor* (.68) and *medium-high* (.61) performing crews was slightly below the threshold. Inter-rater reliability was fair for the *planning* (.47) and *execution* (.43) dimension, but was poor for the *review / modify* (.25), and *overall* (.30) dimensions. Although the raters had not used this rating tool in their training before, the average reliability of LOSA (.37) was slightly higher than for $NTS_{dim}$ (.30) and $NTS_{item}$ (.35). With respect to the ICCs, the Spearman-Brown prophecy formula found that on average, three raters would be required to assess pilots' skills on the LOSA scale with good (.60) reliability.

---

Insert Table 3 around here

---

Agreement on the pass/fail scale (see Table 3, bottom) was high (.92) for the *outstanding* performing crew. Agreement for the lower performing crews was below the acceptable threshold. On average, the rating dimensions did not exceed the acceptable threshold. As for $NTS_{dim}$ and $NTS_{item}$, the average agreement for *medium-high*, *medium-low*, and *poor* performing crews was lower for the pass/fail scale than for the four-point scale. Both rating tools, which had not been used by the raters before ($NTS_{dim}$ and LOSA), showed higher agreement for the pass/fail scale than for the five-point/four-point scale when rating the *outstanding* performing crew.

## Discussion




*This is an Accepted Manuscript of an article published by Taylor & Francis in International Journal of Aviation Psychology on 10 May 2016, available online: http://www.tandfonline.com/10.1080/10508414.2015.1162636.*

The discussion is divided into several parts that correspond to the *measurement* and *target* themes introduced in this paper. It concludes by pointing out some limitations of the study.

### ***Measurement*: Different Rating Tools ($NTS_{dim}$, $NTS_{item}$, LOSA) and Familiarity**

The results showed that neither of the rating tools used here achieved the necessary standard for sufficient inter-rater reliability on average across all dimensions and performance levels. Although the raters were not trained with the LOSA sheet, the reliability was roughly the same for this rating tool at the four-point scale as compared to the other rating tools. The average agreement with LOSA on the pass/fail scale was even better than agreement with the rating tool known from training ($NTS_{item}$). This means on one hand that familiarity with the rating tool alone does not necessarily lead to higher inter-rater reliability. On the other hand these results indicate that more precisely formulated items (LOSA) can outweigh the potential familiarity advantages ($NTS_{item}$). In this context it has to be mentioned that NOTECHS (which was the basis of $NTS_{item}$) was not intended to be used on a pass/fail level unless a rating could be linked to technical consequences (Flin et al., 2003).

Another reason for the similar inter-rater reliability of the LOSA tool could be that the crew is rated as one team (LOSA) and not as two single pilots (as for $NTS_{dim}$ and $NTS_{item}$). Perhaps it is more difficult for raters to assign separate performance contributions to the two crew members, since interaction, which is the basis for most of the NTS dimensions introduced here, is the result of a collaboration of at least two persons. O'Connor et al. (2002) came to a similar conclusion when addressing inter-rater reliability differences between scenarios. Another aspect could be that LOSA also incorporates technical performance aspects such as *automation handling* that are easier to observe. Finally, LOSA only uses 13 items in contrast to $NTS_{item}$, which uses 40 items; raters could have lost interest in thoroughly considering the item definitions before assigning a score. In our study, they had to go through all ratings eight times, which could have led to a *checking-boxes* response style. Although the agreement for $NTS_{item}$ on the five-point scale is acceptable on average, the assessment on the pass/fail scale is what an examination ultimately depends on.

### ***Measurement*: Scale Levels and Degree of Differentiation**

When comparing the two scales, which represent different levels of differentiation, all the rating tools showed lower agreement on the derived pass/fail scale than on the five-point/four-point scale ($NTS_{dim}$: .46 vs. .57; $NTS_{item}$: .19 vs. .72; LOSA: .43 vs. .68) on average. It seems that examiners




*This is an Accepted Manuscript of an article published by Taylor & Francis in International Journal of Aviation Psychology on 10 May 2016, available online: [http://www.tandfonline.com/10.1080/10508414.2015.1162636](http://www.tandfonline.com/10.1080/10508414.2015.1162636).*

can give reliable feedback in general (e.g. pilot A was better than pilot B), but are in less agreement with respect to the level of pass and fail (e.g. pilot A passed, but pilot B failed). O'Connor et al. (2002) found similar results for two of their eight scenarios. This issue is relevant in particular because this threshold between pass and fail is what counts most for the individual pilot. This finding confirms earlier concerns that NTS ratings alone should not be used to pass or fail a crew member unless safety consequences are directly involved.

**_Measurement_: Differences Between the Rating Dimensions**

The results showed that inter-rater reliability was dependent on the dimension being rated. Comparing the different dimensions, both $NTS_{dim}$ and $NTS_{item}$ showed lower inter-rater reliability for the social aspects than for the cognitive aspects. When we assessed the reliability of the pilots' self, peer, and supervisor ratings, we also found less agreement for the social aspects than for the cognitive aspects using the $NTS_{item}$ tool (cf. Gontar & Hoermann, 2014). In this earlier study, the entire mission was rated by all 60 pilots who took part in the simulator experiment. When the whole mission was rated, inter-rater reliability for the social dimensions was even lower than in the present study. The opposite was true for the cognitive aspects, which led to higher inter-rater reliabilities when rating the entire mission and when the rater was directly involved and operating the simulator. In contrast to the findings presented here, Yule et al. (2008) found that for surgeons, aspects of communication, leadership and teamwork (which correspond to our social skills) were rated more reliably than aspects of task management, decision making and situation awareness (which correspond to our cognitive skills). One reason for this different finding could be that the videotapes we selected for this rater study featured a strong emphasis on aspects of decision making and situation awareness. This is because the success in this scenario mainly depended on the appropriate decision making by the crew. This aspect might also be the reason why the LOSA dimension of *Planning* is rated with slightly higher inter-rater reliability than the other LOSA dimensions.

**_Target_: Crews Representing Different Levels of Performance**

Based on the work from Yule et al. (2009), we expected that the most extreme performance characteristics, such as the *outstanding* and *poor* performing crews, would be rated with higher agreement than the *medium* performing crews. What we found was that only the *outstanding* performing crew was rated with acceptable agreement on average. That the *poor* performing crew was rated with lower agreement on the pass/fail scale than the *outstanding* performance is even more surprising, because it was rated directly after the latter. This may indicate that the raters were not subject to sequence effects; had that been the case, they would have consistently rated the *poor* performing crew as very poor. An explanation could be that the raters are seldom faced with *poor* performing crews so that they struggle to assess them in high agreement.

### *Target*: Differences Between the Ratings for Captains and First Officers

The agreement using the two NTS rating tools showed that the FOs are rated slightly more in agreement on the five-point scale than the CPTs, except using the $NTS_{dim}$ tool for the *outstanding* performing crew. O'Connor et al. (2002) compared the deviations of ratings for CPTs and FOs to a reference rating. They found similar results with the ratings for the FOs' performance showing less deviation from the reference ratings than the ratings for the CPTs' performance. It seems that the raters can assess FOs' performance more accurately than the CPTs'. As all the raters were CPTs themselves, they were used to flying together with FOs more often than with CPTs. This daily experience could have provided them with better framing conditions to compare the behaviors of the FOs. Another aspect is that in examination flights for example, a good pilot can disproportionately influence the team's performance and thus compensate for the effects of a more poorly performing crew member. If the good pilot continuously supports his crew member, the performance by the poorly performing pilot might be hidden. So even when both crew members are rated independently ($NTS_{dim}$ and $NTS_{item}$), it might be hard to identify the differences between two pilots' non-technical skills.

### Strengths and Weaknesses of this Study

One weakness of this study might be that the raters did not have enough time to actually observe every behavior, because they only observed a 6-minute sequence of the whole mission. Moreover, they may have assessed aspects which they did not observe instead of leaving the respective items blank or marking it as *not observable* (as they were instructed to do). Such behavior has already been reported by O'Connor et al. (2002). Another effect of the rather short duration and the content of the videotape might have been that the raters did not have the opportunity to observe the pilots' behavior during normal operation flight phases. This would have been the case in examination flights, which usually start with normal operations before the crews are exposed to critical situations. Nevertheless, it is especially the performance in unforeseen and abnormal events that determines the success of a mission.

In contrast, a strength of this study is the large sample of non-volunteer raters which represents the most experienced instructors in the entire airline fleet. We used a clean environment, meaning the raters did not have any parallel tasks such as operating the simulator or acting as air traffic controllers, which could have led to high rater workload during the assessment (Deaton et al., 2007; Seamster, Hamman, & Edens, 1995). Furthermore, all videotapes contained the same task with different performance levels, as would be expected from daily practice.

## Conclusion

The results of this inter-rater reliability study show that the *measurement* as well as the *target* influence inter-rater reliability. We were able to show these effects while keeping other influences by the *rater* and the *scenario & task* constant. On the other hand, we demonstrated that inter-rater reliability is still an unsolved issue even within a group of highly experienced instructors when assessing the non-technical skills of pilots. In Europe, current regulatory material by the European Aviation Safety Agency (2011) states that the practical training of instructors should "include the development of specific instructor skills, particularly in the area of teaching and assessing threat and error management and CRM" (European Aviation Safety Agency, 2011, FCL.920, p. 282). In particular, instructors are required to observe and assess CRM behaviors in order to provide constructive feedback to both the pilots and to the training department (European Aviation Safety Agency, 2011). All these requirements assume that such ratings are based on reliable observations. According to our findings, we strongly recommend incorporating specific inter-rater reliability exercises into trainer standardization and assessment trainings. Therefore, it would be beneficial to describe more precise anchors for desired and undesired behaviors on all observed CRM dimensions. Based on our findings we recommend caution when using NTS-ratings on a pass/fail level. In line with the second NOTECHS principle it should be emphasized that in order to fail a pilot in an examination flight due to non-technical skills, "flight safety must be actually (or potentially) comprised [, which] requires a related objective technical consequence" (Flin et al., 2003, p. 109).

We agree with the opinion that mission-specific CRM evaluation tools and other objectifying resources, such as shown by Brannick et al. (2002), would lead to higher inter-rater reliability in training. Deaton et al. (2007) for example developed a tool which supports the instructors when rating pilots' performance in specific scenarios. This tool will alert the instructor when it detects events in training scenarios that are important for the rating. These authors could show that such a supporting tool leads to more differentiated and more accurate ratings (Deaton et al., 2007). Such an approach seems to be very promising. A goal of future research should be to further elaborate and extend the usage of such techniques with the goal of providing the instructors with more reliable information for their assessments. In parallel, regulatory authorities should explicitly advise airlines and flight training organizations to address and demonstrate sufficient inter-rater reliability among their instructor pilots when utilizing their performance evaluation tools.

Acknowledgements

The authors greatly acknowledge the support of all the examiners, and thank Frederik Niedner for his support and help during the experiments. We also like to thank Maresa Biermann and Fabian Fischer for their help in data handling and Jurek Breuninger as well as Markus Zimmermann for their support on proofreading the manuscript. Special thanks go to Lynne Martin and Rhona Flin for their valuable comments and remarks on this paper.





*This is an Accepted Manuscript of an article published by Taylor & Francis in International Journal of Aviation Psychology on 10 May 2016, available online: http://www.tandfonline.com/10.1080/10508414.2015.1162636.*

References

Arora, S., Miskovic, D., Hull, L., Moorthy, K., Aggarwal, R., Johannsson, H., . . . Sevdalis, N. (2011). Self vs. expert assessment of technical and non-technical skills in high fidelity simulation. *The American Journal of Surgery, 202*(4), 500–506. doi:10.1016/j.amjsurg.2011.01.024

Beard, J. D., Marriott, J., Purdie, H., & Crossley, J. (2011). Assessing the surgical skills of trainees in the operating theatre: A prospective observational study of the methodology. *Health Technology Assessment, 15*(1), 1–162. doi:10.3310/hta15010

Brannick, M. T., & Prince, C. (1991). *Assessment of aircrew rating from within and between scenarios (Tech. Rep. No. DAAL0–3–86–D–001).* Orlando, FL.

Brannick, M. T., Prince, C., & Salas, E. (2002). The reliability of instructor evaluations of crew performance: Good news and not so good news. *The International Journal of Aviation Psychology, 12*(3), 241–261. doi:10.1207/S15327108IJAP1203_4

Burger, K.-H. (Ed.). (1999). *Basic competence for optimum performance*. Frankfurt.

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. In R. S. Jensen (Ed.), *Proceedings of the 12th International Symposium on Aviation Psychology* (pp. 172–175). Dayton, OH: Wright State University Press.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. doi:10.1037/1040-3590.6.4.284

Cooper, S. J., Endacott, R., & Cant, R. P. (2010). Measuring non-technical skills in medical emergency care: A review of assessment measures. *Open Access Emergency Medicine, 2,* 7–16. doi:10.2147/OAEM.S6693

Cooper, G. E., White, M. D., & Lauber, J. K. (Eds.) (1980). *Resource management on the flightdeck: Proceedings of a NASA/Industry Workshop.* NASA Conference Publication 2120, Moffett Field, CA.

Deaton, J. E., Bell, B., Fowlkes, J., Bowers, C., Jentsch, F., & Bell, M. A. (2007). Enhancing team training and performance with automated performance assessment tools. *The International Journal of Aviation Psychology, 17*(4), 317–331. doi:10.1080/10508410701527662

Dedy, N. J., Szasz, P., Louridas, M., Bonrath, E. M., Husslein, H., & Grantcharov, T. P. (2015). Objective structured assessment of nontechnical skills: Reliability of a global rating scale for the in-training assessment in the operating room. *Surgery, 157*(6), 1002–1013. doi:10.1016/j.surg.2014.12.023

European Aviation Safety Agency. (2011). *Annex to ED decision 2011/016/R: Acceptable means of compliance and guidance material to part-FCL.*

European Aviation Safety Agency. (2014). *Annex to ED Decision 2014/022/R*.

European Commission. (2011). *Commission regulation (EU) No 1178/2011 of 3 November 2011: Laying down technical requirements and administrative procedures related to civil aviation aircrew pursuant to regulation (EC) No 216/2008 of the European parliament and of the council.*

Federal Aviation Administration (2004). *Advisory Circular No 120-51E: Crew resource management training*. Washington, DC: U.S. Department of Transportation.

Field, A. P. (2009). *Discovering statistics using SPSS*. Los Angeles, CA: SAGE Publications.

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*. *Wiley Series in probability and statistics*. Hoboken, NJ: J. Wiley.

Fletcher, G., Flin, R., McGeorge, P., Glavin, R., Maran, N., & Patey, R. (2003). Anaesthetists' Non-Technical Skills (ANTS): Evaluation of a behavioural marker system. *British Journal of Anaesthesia, 90*(5), 580–588. doi:10.1093/bja/aeg112

Flin, R., & Martin, L. (2001). Behavioral markers for crew resource management: A review of current practice. *The International Journal of Aviation Psychology, 11*(1), 95–118. doi:10.1207/S15327108IJAP1101_6

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.

Gale, T. C. E., Roberts, M. J., Sice, P. J., Langton, J. A., Patterson, F. C., Carr, A. S., . . . Davies, P. R. F. (2010). Predictive validity of a selection centre testing non-technical skills for recruitment to training in anaesthesia. *British Journal of Anaesthesia, 105*(5), 603–609. doi:10.1093/bja/aeq228

Goldsmith, T. E., & Johnson, P. J. (2002). Assessing and improving evaluation of aircrew performance. *The International Journal of Aviation Psychology, 12*(3), 223–240. doi:10.1207/S15327108IJAP1203_3

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). Malta.

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. Part I* (pp. 119–128). Krakow.

Hamman, W., & Holt, R. (1997). Line operational evaluation (LOE): Air carrier scenario based evaluation. In E. Smith (Ed.), *Proceedings of the Human Factors and Ergonomics Society 41st annual Meeting* (pp. 907–911). Albuquerque, NM: Human Factors and Ergonomics Society.

Harvey, R. J., & Hollander, E. (2004). Benchmarking $r_{wg}$ interrater agreement indices: Let's drop the .70 rule-of-thumb. In *Annual Conference of the Society for Industrial and Organizational Psychology*. Chicago, IL.

Haeusler, R., Klampfer, B., Amacher, A., & Naef, W. (2004). Behavioral markers in analyzing team performance of cockpit crews. In R. Dietrich & T. M. Childress (Eds.), *Group interaction in high risk environments*. Aldershot: Ashgate.

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. doi:10.1207/s15327108ijap0901_2

Helmreich, R. L., Klinect, J.R., Wilhelm, J. A., & Jones, S.G. (1999). *The Line/LOS Error Checklist, Version 6.0: A checklist for human factors skills assessment, a log for off-normal events, and a worksheet for cockpit crew error management*. Austin, TX.

Hoermann, H.-J. & Neb, H. (2004). From NOTECHS to LH behavior markers: An implementation case study. Paper presented to the Royal Aeronautical Society - Human Factors Group Seminar on Assessment & Accreditation, London April 30, 2004.

Holt, R. W., Hansberger, J. T., & Boehm-Davis, D. A. (2002). Improving rater calibration in aviation: A case study. *The International Journal of Aviation Psychology, 12*(3), 305–330. doi:10.1207/S15327108IJAP1203_7

International Civil Aviation Organization. (2002). Line Operations Safety Audit (LOSA): DOC 9803. AN/761. Montreal.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*(1), 85–98. doi:10.1037/0021-9010.69.1.85

James, L. R., Demaree, R. G., & Wolf, G. (1993). $r_{wg}$: An assessment of within-group interrater agreement. *Journal of Applied Psychology, 78*(2), 306–309. doi:10.1037/0021-9010.78.2.306

Joint Aviation Authorities. (2001*). Joint aviation regulations: JAR OPS 1.940, 1.945, 1.955, and 1.965.* Hoofddorp, Netherlands.

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* (pp. 663–668). Dayton, OH.

Kontogiannis, T., & Malakis, S. (2013). Strategies in coping with complexity: Development of a behavioural marker system for air traffic controllers. *Safety Science, 57,* 27–34. doi:10.1016/j.ssci.2013.01.014

Landis, R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Liao, S. C., Hunt, E. A., & Chen, W. (2010). Comparison between inter-rater reliability and inter-rater agreement in performance assessment. *Annals Academy of Medicine Singapore, 39*(8), 613–618.

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse* [Test construction and test analysis]. Weinheim: Beltz, Psychologie Verl.-Union.

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. doi:10.1136/qshc.2007.024760

Mitchell, L., Flin, R., Yule, S., Mitchell, J., Coutts, K., & Youngson, G. (2012). Evaluation of the scrub practitioners' list of intraoperative non-technical skills (SPLINTS) system. *International Journal of Nursing Studies, 49*(2), 201–211. doi:10.1016/j.ijnurstu.2011.08.012

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. *McGraw-Hill series in psychology*. New York: McGraw-Hill.

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

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics, 2*(1), 21–33.

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Buehner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology, 6*(4), 147–151. doi:10.1027/1614-2241/a000016

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.

Thode, H. C. (2002). *Testing for normality*. New York: CRC Press.

Weber, D. E., Roth, W.-M., Mavin, T. J., & Dekker, S. W. (2013). Should we pursue inter-rater reliability or diversity? An empirical study of pilot performance assessment. *Aviation in Focus – Journal of Aeronautical Sciences*, *4*(2), 34–58.

Weber, D. E., Mavin, T. J., Roth, W. M., Henriqson, E., & Dekker, S. W. A. (2014). Exploring the use of categories in the assessment of airline pilots' performance as a potential source of examiners' disagreement. *Journal of Cognitive Engineering and Decision Making, 8*(3), 248–264. doi:10.1177/1555343414532813

Williams, D., Holt, R., & Boehm-Davis, D. (1997). Training for inter-rater reliability: Baselines and benchmarks. In R. S. Jensen (Ed.), *Proceedings of the 9th Symposium on Aviation Psychology* (pp. 514–519). Columbus, OH.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen* [Rater agreement and rater reliability: Methods to measure and to improve the reliability of assessments using categorical data and rating scales]. Goettingen: 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. doi:10.1007/s00268-007-9320-z

Yule, S., Rowley, D., Flin, R., Maran, N., Youngson, G., Duncan, J., & Paterson-Brown, S. (2009). Experience matters: Comparing novice and expert ratings of non-technical skills using the

NOTSS system. *ANZ Journal of Surgery, 79*(3), 154–160. doi:10.1111/j.1445-2197.2008.04833.x




*This is an Accepted Manuscript of an article published by Taylor & Francis in International Journal of Aviation Psychology on 10 May 2016, available online: http://www.tandfonline.com/10.1080/10508414.2015.1162636.*

$NTS_{dim}$

| CPT | | | | | | FO | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ++ | + | o | - | -- | CRM-rating dimension | ++ | + | o | - | - - |
| | | | | | Communication | | | | | |

*"Communication includes information transfer and social aspects. The crew members share their information, and assure reception and understanding. Suggestions of other crew members are considered, even if one does not agree. Ambiguities and uncertainties are announced."* (Burger, 1999, p. 14)

$NTS_{item}$

| CPT | | | | | | FO | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ++ | + | o | - | -- | Communication | ++ | + | o | - | - - |
| | | | | | announce ambiguities | | | | | |

LOSA (from International Civil Aviation Organization, 2002, p. A-8; Klinect et al., 2003)

| 1 = Poor<br>Observed performance had safety implications | 2 = Marginal<br>Observed performance was barely adequate | 3 = Good<br>Observed performance was effective | 4 = Outstanding<br>Observed performance was truly noteworthy |
|---|---|---|---|
| Overall behavioral markers | | | Crew Rating |
| Communication environment | Environment for open communication was established and maintained | Good cross talk – flow of information was fluid, clear, and direct | |



Figure 1. Examples of content for the different rating tools that were used for the inter-rater reliability study. From top to bottom: $NTS_{dim}$ (see Burger et al., 1999), $NTS_{item}$ (see Burger et al., 2003), LOSA (see International Civil Aviation Organization, 2002, p. A-8; Klinect et al., 2003). For $NTS_{dim}$, the definitions of the items were not specified on the rating tool, but were included in the airline's training material the raters had.

Table 1. $r_{wg}$ and *ICC(3)* of the $NTS_{dim}$ ratings as a function of the performance level shown in the scenario and crew position.

| $NTS_{dim}$ (CPT/FO) | Performance Level | | | | $r_{wg}$ mean | *ICC(3)* |
|---|---|---|---|---|---|---|
| | Out-standing | Medium-high | Medium-low | Poor | | |
| **Five-point scale** | | | | | | |
| Communication | .77 / .62 | .33 / .62 | .53 / .67 | .42 / .53 | .56 | .12 |
| Leadership & Teamwork | .84 / .83 | .42 / .62 | .25 / .66 | .62 / .63 | .61 | .28 |
| Work Organization | .75 / .77 | .37 / .57 | .56 / .66 | .36 / .48 | .57 | .34 |
| Situation Awareness & Decision Making | .81 / .73 | .47 / .29 | .50 / .53 | .44 / .55 | .54 | .45 |
| Mean | .79 / .74 | .40 / .53 | .46 / .63 | .46 / .55 | .57 | .30 |
| **Pass/fail scale** | | | | | | |
| Communication | .89 / .69 | .10 / .60 | .51 / .51 | .29 / .51 | .51 | - |
| Leadership & Teamwork | 1 / 1 | .00 / .59 | .50 / .59 | .29 / .58 | .57 | - |
| Work Organization | 1 / 1 | .00 / .14 | .59 / .43 | .00 / .13 | .41 | - |
| Situation Awareness & Decision Making | 1 / 1 | .10 / .00 | .59 / .17 | .00 / .00 | .36 | - |
| Mean | .97 / .92 | .05 / .33 | .55 / .43 | .14 / .30 | .46 | - |

*Note*. In addition to the five-point scale (top), we derived a dichotomous pass/fail scale by assigning the lower two scale-points to *fail* and the upper three scale-points to *pass* (bottom). The mean ICC value was calculated using Fisher z' transformation.





*This is an Accepted Manuscript of an article published by Taylor & Francis in International Journal of Aviation Psychology on 10 May 2016, available online: http://www.tandfonline.com/10.1080/10508414.2015.1162636.*

Table 2. $r_{wg}$ and *ICC(3)* of the $NTS_{item}$ ratings as a function of the performance level shown in the scenario and crew position.

| $NTS_{item}$ (CPT/FO) | Performance Level | | | | $r_{wg}$ mean | *ICC(3)* |
|---|---|---|---|---|---|---|
| | Out-standing | Medium-high | Medium-low | Poor | | |
| **Five-point scale** | | | | | | |
| Communication | .76 / .77 | .54 / .68 | .69 / .76 | .74 / .75 | .71 | .22 |
| Leadership & Teamwork | .79 / .84 | .73 / .71 | .75 / .79 | .76 / .71 | .76 | .32 |
| Work Organization | .76 / .79 | .52 / .64 | .69 / .79 | .64 / .75 | .70 | .37 |
| Situation Awareness & Decision Making | .78 / .81 | .60 / .63 | .64 / .79 | .69 / .62 | .70 | .48 |
| Mean | .77 / .80 | .60 / .67 | .69 / .78 | .71 / .71 | .72 | .35 |
| **Pass/fail scale** | | | | | | |
| Communication | .43 / .13 | .24 / .10 | .00 / .00 | .13 / .05 | .13 | - |
| Leadership & Teamwork | .17 / .43 | .30 / .10 | .00 / .00 | .36 / .00 | .17 | - |
| Work Organization | .59 / .59 | .30 / .09 | .02 / .00 | .13 / .02 | .22 | - |
| Situation Awareness & Decision Making | .36 / .23 | .44 / .19 | .09 / .00 | .51 / .23 | .25 | - |
| Mean | .39 / .34 | .32 / .12 | .03 / .00 | .28 / .08 | .19 | - |

*Note*. In addition to the five-point scale (top), we derived a dichotomous pass/fail scale by assigning the lower two scale-points to *fail* and the upper three scale-points to *pass* (bottom). The mean ICC value was calculated using Fisher z' transformation.





*This is an Accepted Manuscript of an article published by Taylor & Francis in International Journal of Aviation Psychology on 10 May 2016, available online: http://www.tandfonline.com/10.1080/10508414.2015.1162636.*

Table 3. $r_{wg}$ and *ICC(3)* of the LOSA ratings as a function of the performance level shown in the scenario.

| LOSA | Performance Level | | | | $r_{wg}$ mean | *ICC(3)* |
|---|---|---|---|---|---|---|
| | Out-standing | Medium-high | Medium-low | Poor | | |
| **Four-point scale** | | | | | | |
| Planning | .76 | .71 | .78 | .68 | .74 | .47 |
| Execution | .78 | .72 | .80 | .74 | .76 | .43 |
| Review / Modify | .70 | .53 | .65 | .65 | .63 | .25 |
| Overall | .70 | .48 | .60 | .65 | .61 | .30 |
| Mean | .74 | .61 | .71 | .68 | .68 | .37 |
| **Pass/fail scale** | | | | | | |
| Planning | .89 | .00 | .59 | .01 | .37 | - |
| Execution | 1 | .03 | .59 | .01 | .41 | - |
| Review / Modify | .89 | .09 | .51 | .07 | .39 | - |
| Overall | .89 | .19 | .69 | .42 | .54 | - |
| Mean | .92 | .07 | .60 | .13 | .43 | - |

*Note*. In addition to the four-point scale (top), we derived a dichotomous pass/fail scale by assigning the lowest scale point to *fail* and the upper three scale-points to *pass* (bottom). The mean ICC value was calculated using Fisher z' transformation.



*This is an Accepted Manuscript of an article published by Taylor & Francis in International Journal of Aviation Psychology on 10 May 2016, available online: http://www.tandfonline.com/10.1080/10508414.2015.1162636.*