Cite as: O'Connor, P., Hörmann, H.-J., Flin, R., Lodge, M., Goeters, K.M., & the JARTEL group. (2002). Developing a method for evaluating CRM skills: A European perspective. *International Journal of Aviation Psychology, 12(3)*, 263-286.

Developing a method for evaluating Crew Resource Management skills: A European perspective

ABSTRACT

The European Commission in conjunction with the European Joint Aviation Authorities (JAA, Human Factors Project Advisory Group) has been sponsoring a series of studies investigating a culturally-robust method for the evaluation of pilots' non-technical skills for multi-crew operations. This paper will outline the development of a European behavioural marker system for CRM evaluation called NOTECHS (NOn-TECHnical skills) and will present preliminary results from an ongoing test phase of this system (JAR TEL). The JAR TEL (Joint Aviation Requirements Translation and Elaboration of Legislation) project has involved 105 instructors from 14 European airlines who were given a short training session to use the NOTECHS system. Following the training phase, these instructor pilots used the system to evaluate the individual CRM skills of captains and first officers in eight different video scenarios filmed in a Boeing 757 simulator. Issues relating to rater training, reliability, accuracy as well as the instructors' opinions on the acceptability of method are discussed.

INTRODUCTION

The area of NTS evaluation in Europe has become increasingly important in the light of recent JAA (Joint Aviation Authorities) legislation which asks for the assessment of NTS (JAR-OPS; Joint Aviation Requirements for Flight Operations, 1999). However, although the regulations make recommendations as to what should be included in CRM training, they do not suggest how NTS should be evaluated, or which NTS should be included in the assessment framework. “The flight crew must be assessed on their CRM skills in accordance with a methodology acceptable to the Authority and published in the Operations Manual. The purpose of such assessment is to:

- Provide feedback to the individual and serve to identify retraining; and
- Be used to improve the CRM training system.” (JAA-OPS, NPA-16, 1999).

Therefore, the Joint Aviation Authorities-Project Advisory Group for Human Factors (JAA- PAG) tasked four research institutes (NLR, DLR, IMASSA, and the University of Aberdeen) to develop a NTS assessment framework that became known as NOTECHS (Non-Technical Skills). The JAR TEL (Joint Aviation Requirements and Translation Elaboration of Legislation) project was born out of the NOTECHS project, with the aim of assessing the usability and the validity of the set of behavioural markers established in NOTECHS through both experimental and operational evaluation.

This paper will provide some background on existing behavioural marker systems, outline the NOTECHS framework, describe the preliminary testing of the system in an experimental setting and discuss the implications of NOTECHS for training.

Behavioural Marker Systems

Only a limited amount of research has been conducted into the development of behavioural markers systems for the evaluation of pilots’ NTS (see Flin & Martin, 2001 for a review). The seminal research on behavioural markers comes from Helmreich’s group at the University of Texas/NASA/FAA Aerospace Crew research project. In the late 1980s they developed a data collection form called the LINE LOS Checklist (LLC) to gather information on flight crews’ CRM performance (Helmreich, Wilhelm, Kello, Taggart & Butler, 1990). The behaviours included in the LLC have their origins in pilot attitudes to cockpit management (Helmreich, 1984) and the analysis of accidents and incidents with identifiable human factors causation (Connelly, 1997). This checklist is widely cited and it has been used as the basis of many airlines’ behavioural marker systems (Flin & Martin, 2001). The LLC system has been refined over the years on the basis of ongoing observational research (Clothier, 1991; Helmreich, 2000), and was recently integrated into the current version of LOSA (Line Operations Safety Audit) Version 9.0 (Helmreich, Klinect & Wilhelm, 1999) elicits ratings in three broad categories (planning, execution, and review/modify plans) from four phases of flight (pre-departure, take-off and climb, cruise, and approach and landing). The system primarily concentrates on the performance of the crew, although overall performance contribution of the individual crew members can be made. As LLC and especially LOSA were designed to provide a wide range of safety related indications of the respective organisation, the entity of analysis in these systems is not the individual crew member per se but the organisation itself or parts of it (fleets, operational units etc.).

Another earlier marker system which developed by Fowlkes, Lane, Salas, Franz and Oser (1994) was a team performance measurement approach called TARGETs (Targeted Acceptable Responses to Generated Events or Tasks) for US military cargo helicopter teams. This was based on a set of critical aircrew co-operation behaviours, grouped into seven basic skill areas: mission analysis, adaptability/flexibility, leadership, decision making, assertiveness, situational awareness and communication. In this system, for each stimulus event in a scenario, there is a predefined set of acceptable behaviours, each is rated as present or absent. As with the LLC this is a measure of crew performance rather than individual performance. Fowlkes et al (1994) tested the TARGETs approach in a training and evaluation study of six military aircrew and found the measure to have sensitivity and an acceptable degree of inter-rater reliability.

Many of the large airlines have also developed their own behavioural marker systems which are mostly used for training (see Avermaete, & Kruijsen, 1998; Flin & Martin, 2001). To aid in this, a research group at George Mason University have produced guidance to assist companies in the development of a CRM skills evaluation system, and to train instructors to use it (George Mason University, 1996). However, many of the smaller companies in Europe do not have the time, resources, or expertise to develop their own systems. Thus, with the recent change in the JAA regulations described above, it was recognised that there was a need for a basic, generic system which was not specific to any one company, country, or type of operation and allowed the pilots’ NTS skills to be assessed individually rather than as a crew.

The NOTECHS Framework

The NOTECHS project was sponsored by EC DGTREN (European Community Directorate for Transport and the Environment) and the Civil Aviation Authorities of France, Netherlands, Germany, and UK and it ran from March 1997 until March 1998 (see Avermaete & Kruijsen, 1998). The central goal of the NOTECHS project was to provide guidance for a feasible and efficient method for assessing pilots’ non-technical skills by instructor pilots and examiners during training and check events in multi-crew aircraft in countries across Europe. The method was to be based on common elements of NTS training and/or evaluation systems which were in use with European airlines such as Lufthansa and KLM. The first stage was to review existing systems of evaluating NTS. Flin and Martin (1998, 2001) surveyed 12 UK airlines and 14 large international carriers and found a wide range of marker systems in use. The NOTECHS group also looked in detail at the systems used by: Air France, British Airways, Lufthansa, KLM (Royal Dutch Airlines), and the Dutch CAA (Avermaete & Kruijsen, 1998). A number of conclusions were drawn from these surveys:

- No airline had simply adopted an off-the-shelf NTS assessment system, although a number of airlines had adapted their NTS system from the NASA/UT LLC system (Helmreich, Butler, Taggart & Wilhelm, 1995).
- Although there were differences in rating scales, all airlines attempted to define a distinction between acceptable and unacceptable NTS performance.
- Clear and unambiguous definitions of all terms in an NTS system are necessary for proper assessment and clear pilot debriefings, especially if the system is to be used by several different airlines in different countries.
- It would be advisable to set up a system of pilot NTS performance tracking, so that any NTS training and evaluation system could be adapted to changing operational procedures and expanding knowledge.
- Key categories of NTS across systems appeared to be related to decision making, situation awareness, leadership and teamwork.

Following the survey, a literature review of relevant research findings related to these key categories of NTS identified in the survey was carried out (Avermaete & Kruijsen, 1998), and extensive discussions were undertaken between the psychologists and pilots in the consortium. It was concluded that none of the existing systems could be adopted in their original form, nor did any single system provide a suitable basis for simple amendment. Particular attention was paid to two of the principal frameworks, namely the KLM SHAPE (Self, Human interaction, Aircraft, Procedures and Environment and situation) and NASA/UT LLC system. The following principles were used to guide the final choice of components and descriptor terms for the NOTECHS framework:

- The basic elements should be formulated with the maximum mutual exclusivity.
- A rule of parsimony was applied, in that the system should contain the minimum number of categories and elements in order to encompass the critical behaviours.
- The terminology used should reflect unambiguous everyday language for behaviour, rather than psychological jargon.
- The skills listed at the behaviour level should be directly observable in the case of social skills or could be inferred from communication, in the case of the cognitive skills.

The NOTECHS framework consists of a hierarchy of three levels: Elements, Categories, and Pass/fail (see Figure 1). Based on the individual behaviour ratings at Element level, the user formulates the ratings at the Category level, which finally leads to a pass or fail judgement (i.e. the recommendation of further training).

[insert figure 1 here]

The primary Category level can be divided into two social skills (Co-operation, Leadership & Management skills) and two cognitive skills (Situation awareness, Decision making). This elemental set was based on theoretical models identified from the literature review (see Avermaete & Kruijsen, 1998) and was compared against the KLM SHAPE system and the NASA UT LLC (version 4.4; Helmreich, Butler, Taggart & Wilhelm, 1997) to confirm that essential elements had been encompassed. Each Category is then further subdivided into three or four Elements (see Table 1). For each element a number of positive and negative exemplar behaviours were included, again devised from the literature review and existing systems (Flin & Martin, 1998). The exemplar behaviours were phrased as generic (e.g. closes loop for communication), rather than specific (e.g. reads back to air traffic control).

[insert table 1 here]

Two other possible categories taught as CRM modules which were considered and then rejected by the consortium were Communication and Personal limitations. Communication is included as a separate category in a number of systems. However, in the context of NOTECHS, communication is seen as a medium of observation, which is inherent in all four categories. A category of Personal limitations (e.g. stress and fatigue) was also rejected due to the difficulty in observing except in the most extreme of cases.

Once the framework had been developed, the aim of the JAR TEL project was to begin to evaluate the system by assessing the usability and the reliability of the method through experimental and operational testing. The JAR TEL project will be discussed below.

The JAR TEL Project

The JAR TEL project can be divided into two main stages: the experimental phase in which the usability and cultural robustness of the NOTECHS framework is tested by using video scenarios; and the operational phase in which the system is used in real-life by instructors to assess NTS. This paper will concentrate on the experimental phase of the project to assess the usability of the NOTECHS system in a controlled setting.

To test the NOTECHS framework it was decided to use a five-point rating scale at the Element and Category levels. Criteria were developed for deciding which rating should be awarded at the Element and Category levels (see Table 2).

[insert table 2 here]

If a behaviour was not observed, then a Not observed response should be recorded. At the Pass/fail level, a rating of fail should be given if the pilot displays overall behaviour related to the NTS, that endangers, or could endanger, flight safety; and a rating of pass should be given if the pilot’s overall behaviour does not endanger flight safety. Specific exemplar behaviours were not specified at each of the 5 points for each exemplar Elements, as only the basic usability and cultural sensitivity of the scale was being tested at this stage.

A number of methods were used to test the robustness of the NOTECHS framework:

- An assessment of the internal consistency of the NOTECHS system was carried out by determining the extent to which ratings at the Category level were in line with those at the Element and Pass/fail levels.
- The accuracy was assessed by measuring the extent to which the participants’ ratings matched those of the reference ratings (‘expert benchmark’ formulated by scenario designers plus two groups of experienced NTS evaluators, see method for details). This is calculated as a consensus at the Category and Pass/fail levels. The Element level is not included because this is not the focus of the system at this stage of development.
- The inter-rater agreement was measured using an index developed by James, Demaree and Wolf (1984, 1993) called the within group inter-rater reliability measure (r_{wg}; see James et al, 1984, 1993 for more detail). Values of r_{wg} can vary from 0 to 1. When the variance of the obtained ratings is random, then r_{wg} =0, reflecting no agreement

among raters. However when there is total agreement between the raters, then r_{wg} =1.This measure was selected as it has been used to assess inter-rater reliability of behavioural marker systems in aviation by Law and Sherman (1995) and Hamman, Beaubien and Holt (1999).

- The acceptability of the system was assessed using the feedback from the raters on the Evaluation Questionnaire.

METHOD

Prior to carrying out the experiment, it was necessary to develop a set of training and test videos to be used in the experiment, and to establish a method for calculating an expert benchmark or reference rating.

Design of video scenarios

The scenarios to be used in the experiment were filmed in a Boeing 757 simulator, with the Captain and the First Officer (F/O) played by male pilots from two major European airlines with experience in CRM training and video production. In order to minimise the risk of the raters “type-casting” the actors after viewing a particular scenario, every effort was made to distribute the various performances as widely as possible. This factor was also covered during the pre-experiment briefings given to the participants.

Eight scenarios were used in the main experiment (average length 7 minutes, range 3 to 15 minutes), chosen from a total of fifteen which were filmed. The scenarios were designed by a training captain and a psychologist from the consortium to demonstrate range of realistic situations, and although the scenarios were not scripted to the level of prescribing exactly what should be said, each scenario has a set of design references which were levels of NTS for each behaviour Category (on the five point scale) which the pilot actors were supposed to illustrate.

1. Descent- the F/O is the pilot flying. A passenger problem is reported by the cabin crew. The action centres around the Captain allowing himself to be distracted by secondary events and not monitoring the F/O's actions. This developed into an altitude violation.

2. In cruise over Brussels- 170 miles to destination London Heathrow. After suffering an engine fire, the Captain decides to continue to destination against the good advice of the F/O.

3. Crew carrying out pre-departure checks. The F/O is unfamiliar with the airfield and receives little or no support from the Captain.

4. Top of descent- an electrical failure occurs. Problem well handled by both pilots working as a team.

5. Approach in very gusty conditions. The Captain is very supportive of the under-confident F/O and achieves a very positive result after good training input.

6. A night approach in the mountains. Captain decides to carry out a visual approach through high terrain and triggers a GPWS warning. F/O takes control and prevents an accident.

7. An automatic approach in CAT III conditions. Very good standard operation. An example of a typical everyday flight deck activity with both pilots contributing to a safe outcome.

8. Joining the holding-pattern awaiting snow-clearance. The Captain persuades the F/O that they should carry out a visual approach with an illegally excessive tail-wind for commercial reasons.

Reference rating

A set of benchmark or reference data was required for the analysis process in order to examine rater accuracy. Two independent groups of experienced Training Captains (three from British Airways and five from Lufthansa) were used to establish the reference ratings. The criteria for being a member of these groups was that they had to hold valid licenses for instruction and examination, be experienced in CRM training, and to be actively carrying out both Line-Oriented Flight Training and NTS evaluation. The two independent groups were briefed using the JARTEL training material and were then asked to assess the NTS shown in the eight test scenarios. Each group member rated the scenarios individually and then, following a group discussion, arrived at a consensus rating for each of the Categories and Pass/fail judgements.

The consensus ratings from the two groups of experts did not agree exactly at the Category level in 46% of the total Category evaluations (2 pilots x4 Categories x8 scenarios= 64). However, in only 6% of the total evaluations made was this across the acceptable/poor divide. Thus, the remaining discrepancy was either at the poor/very poor level, or the acceptable/good/very good level which are fine grained judgements to make. At the Pass/fail level, the raters agreed in 81% of evaluations (2 pilots x8 scenarios= 16). In the cases where the British Airways and Lufthansa groups showed discrepant ratings, the original design reference was consulted to determine the appropriate rating. The design reference was the behaviour specification from the original script which the pilot actors in the scenarios were supposed to demonstrate. One reason for the difference in the Pass/fail ratings between the two groups can be attributed to the difference in Standard Operating Procedures in the airlines concerned.

Participants

Fifteen experiment sessions were run involving 105 instructor pilots from 14 different airlines across Europe. The participants were all male with an average of 6 years as an instructor (st. dev= 6.2) with an average total flying hours of 10200 hours (st. dev= 3852).

Procedure

Groups of raters recruited from each airline participated in the experiment during one full day. The same pilot facilitator was involved in every session assisted by at least one consortium psychologist. All participants were already briefed about the background of the experiment and about the NOTECHS method by written material distributed in advance.

Due to difficulties in obtaining groups of instructor pilots, access for both training and experimentation was restricted to a single day. Therefore, the training sessions had to be designed to fit into a 3 hour time frame, followed by the 3 hour experimental session. This was sufficient for the basic experimental usability test, however, it was not intended to constitute a full or proper training requirement.

The raters received a short introduction to the JAR TEL experiment and were asked to fill out a background questionnaire to gather data about their professional background- such as age, experience of NTS evaluation, and English language ability. The raters then received behavioral observation training in the NOTECHS method and instructions for using the method during the experiment. This briefing was carried out in a controlled manner using the training video and an interactive question and answer session. At the end of the training video, raters further practised using the NOTECHS system to rate two more complex scenarios.

In the afternoon session, the eight test scenarios were shown, with the raters rating the Element, Category, and Pass/fail levels for both the Captain and F/O after each scenario.

After all experimental sessions had been run, the raters filled out an Evaluation Questionnaire, which contained 16 questions about their opinion of the NOTECHS system and the experimental method. Lastly, open discussions were conducted for debriefing on general feelings, to achieve knowledge on the context and to collect qualitative data for the understanding of the results.

RESULTS

The results cover a number of aspects of reliability and validity of the NOTECHS system, namely: the internal consistency, accuracy, inter-rater reliability and acceptability to the users.

Internal consistency

An assessment was made of the agreement between the Element level and the Category level, and between the Category level and the Pass/fail level. To evaluate the consistency between the Element and Category levels, an assessment was made of the absolute difference between the response to the Element and the response given to the corresponding Category. This was carried out by calculating the mean difference between each of the three or four Elements and their corresponding Categories. This was performed in the majority of situations in which at least one of the ratings at the Elements level was an observed rating, and the Category rating was not missing or rated Not observed.

This technique could not be used to compare the Categories with the Pass/fail response due to the dichotomy of the Pass/fail response. Therefore, it was necessary to collapse the Category level responses into a two point scale to allow a comparison of the consistency at the Pass/fail level. This was accomplished using the following method. If the Category was rated as acceptable, good, or very good this was considered to be consistent with a pass. Thus, a rating of very poor or poor at the Category level was considered to be a fail. As long as no more than one Category was not in line with the Pass/fail decision this was considered consistent

Figure 2 depicts the level of consistency between the Element and Category levels by showing the mean absolute difference across each of the eight scenarios for the four Categories, and the overall absolute difference across the Categories. It can be seen that the consistency is very high (a mean of less than 0.2 of a scale point between the Elements and Category) on all of the Categories except for Decision making. However, even for Decision making the mean absolute difference between the Elements and the Category is less than 0.5 on a five point scale.

[insert figure 2 here]

A two factor (Pilot, Categories) repeated measures analysis of variance was run using the mean absolute difference scores of the difference between the responses given at the Element level and the response given at the Category level. As would be expected from Figure 2, there was a significant main effect of both Pilot ($F_{(1,103)}$ = 54.74, p<.01) and Category ($F_{(2.2,230.5)}$ = 582.0, p<.01) and a significant interaction between the two factors ($F_{(2.5,256.8)}$ = 582.0, p<.01). The consistency between the Elements and corresponding Categories was significantly higher for the F/O (0.18) than the Captain (0.22). However, it should be indicated that despite the difference being significant, it is very small, with the significance resulting from the rather large sample size. Looking at each Category separately, it was found that all of the Categories were significantly different from each other except for Co-operation (0.10) and Situation awareness (0.11). From Figure 2, it can be seen that the interaction is due to the finding that for the first three Categories the inconsistency is greater for the Captain than the F/O, but for the Decision making ratings, the reverse is true. On the Decision making Category, the consistency is lower for the F/O than the Captain. This is confirmed by looking at the contrasts between the variables.

The consistency between the Categories and the Pass/fail level is shown in Figure 3. It can be seen that for all of the Categories, the consistency with the Pass/fail response was at least 75% across the eight scenarios. The overall consistency is a measure of the extent to which at least three out of the four Categories are consistent with the Pass/fail response.

[insert figure 3 here]

A two factor (Pilot, Categories) repeated measures analysis of variance was run by adding the number of matches between the Categories (on the collapsed two point scale) and the response given at the Pass/fail level across the eight scenarios for the Captain and the F/O. Thus, if the response given for Decision making for the Captain was poor and the Pass/fail response was fail, this was considered to be a match. It was found that both the main effects of Pilot ($F_{(1,102)} = 1.07$, n.s.) and Category ($F_{(2.74,279.8)} = 2.49$, n.s.) were not significant. However, there was a significant interaction between the two variables ($F_{(2.8,285.3)} = 19.19$, $p<.01$). From examining the contrasts and looking at the graph of the interactions, it was found that the only situation in which there was a lack of interaction was between the Co-operation and Situation awareness Categories. For both of these Categories the match with the Pass/fail level was higher for the F/O than the Captain. However, the reverse was true for Leadership & management and Decision making creating an interaction (see Figure 3). In addition, there was an interaction between Leadership & management and Decision making because the difference between the Captain and F/O was significantly greater for Decision making than for Leadership & management.

Therefore, to summarise the consistency of the participants’ ratings of the Elements and Categories and Categories and Pass/fail appears to be fairly high. However, from looking at the absolute differences between ratings of the Elements and Categories it can be seen that the Decision making Category shows least consistency between the Elements and Categories, although this is not reflected in the consistency of ratings at the Category and Pass/fail level.

Accuracy

The raters’ scores were compared with the reference ratings for the Captain and F/O for each of the 8 scenarios. The accuracy at the Category level was assessed by calculating the absolute difference between the reference rating and the response given by the raters on the five-point scale. The Category level responses were also examined by collapsing the scale into a two point to examine the responses to the individual scenarios. This is a useful method to examine the data as it allows an examination to be made of the most crucial distinction in the scale between poor and acceptable. At the Pass/fail level the accuracy was assessed by simply summing the relative frequencies participants whose responses matched the reference rating.

Figure 4 shows the mean absolute difference between the reference rating and the participants responses for each of the four Categories averaged across the eight scenarios. It can be seen that the Co-operation Category was the most accurately rated, with the remaining three Categories all having a similar level of accuracy.

[insert figure 4 here]

A two factor (Pilot, Categories) repeated measures analysis of variance was run using the mean absolute difference scores of the difference between the responses given at the Category level and the reference rating. As would be expected from Figure 4, there was a significant main effect of both Pilot ($F_{(1,103)} = 33.3$, $p<.05$) and Category ($F_{(2.7,269)} = 34.2$, $p<.05$) and a significant interaction between the two factors ($F_{(2.7,277.6)} = 7.2$, $p<.05$). The accuracy was significantly superior for the F/O (0.55) than for the Captain (0.69). Examining the Category individually, it was found that they all were significantly different from each other except for Leadership & management (0.67) and Situation awareness (0.69), and Leadership & management and Decision making (0.62). It can also be seen that the greatest inaccuracy was on the Situation awareness and Decision making Categories. From looking at the contrasts between the variables, it was found that the interaction was due to the small difference (compared to the other variables) on the Co-operation Category between the rating accuracy of the Captain and the F/O, compared to the greater accuracy of rating the F/O when compared to the Captain on the remaining three Categories.

An examination was also made of the accuracy of raters separately for each of the eight scenarios. Figure 5 shows the accuracy using the absolute difference method for each of the eight scenarios. It can be seen that the greatest mean absolute differences from the reference ratings occurred for the Captain in scenarios 1, 2, and 4 and the F/O in scenarios 1 and 4. However, this was not reflected in the collapsed scale method (see Figure 6). This method indicates that scenarios 1, 3, and 8 were the most difficult for participants to rate accurately. Thus, in scenarios 2 and 4 the differences must not be across the poor/acceptable divide, but rather between acceptable, good, and very good or poor and very poor.

[insert figure 5 here]

[insert figure 6 here]

At the Pass/fail level, the mean accuracy of the raters was 83% across the eight scenarios for the Captain (st. dev= 11.4), and 84% for the First Officer (st.dev= 12.7). A repeated measures t-test indicated that this difference was not significant (t=-0.67, df=103, n.s.). The accuracy of raters at the Pass/fail level also shows that scenarios 1 and 3 for the Captain and scenarios 1, 3, and 8 for the F/O was lower than compared to the other scenarios (see Figure 7). However, a Chi-square test showed that a significantly larger number of raters matched the reference rating in seven of the eight scenarios than gave a response which disagreed with the reference rating (see Table 3). Only in scenario 3 for the Captain did a significantly larger proportion of raters disagree with the reference rating ($\chi^2 \geq 3.8$, df=1, p<.05) and in scenario 8 for the F/O the difference between agree and disagree was not high enough to be significant ($\chi^2 < 3.8$, df=1, n.s.; see Table 3).

[insert figure 7 here]

[insert table 3 here]

The agreement with the not observed reference rating was not included in the above analysis. This was due to the unique nature of this response. In only three occasions was the reference rating not observed. For the Decision making Category in scenario 3 for both the Captain (only 14% of participants agreed with the not observed reference rating) and the F/O (only 22% of participants agreed with the not observed reference rating) and for the Decision making Category for the Captain in scenario 7 (only 20% of participants agreed with the not observed reference rating). Therefore, it can be seen that there was a tendency for the participants to rate behaviours which were not judged to be present by the experts.

At the Pass/fail level, inspection of the results suggests that the Captains have been assessed independently from the F/Os. In the three scenarios where the reference rating for Captain is fail and for F/O is pass (scenarios 2, 6, and 8), the level of accuracy of the raters is generally 83-95% (see Figure 7). There is a higher level of disagreement in scenario 8 where 46% of raters also failed the F/O, but it is not possible to conclude whether this was related to the Captain’s rating.

To summarise, overall there was a high level of agreement between the participants and the experts at the Category level. However, at the Pass/fail level, on the more ambiguous scenarios (particularly scenario 3) the proportion of raters matching the expert’s reference rating was reduced. Also, there was a tendency for the participants not to use the Not observed rating, even when the reference rating was not observed .

Inter-rater agreement

The within-group inter-rater reliability coefficient (r_{wg}) was used to analyse the inter-rater agreement at both the Category (using the five point scale) and Pass/fail levels (using the two point scale). Figure 8 shows the mean r_{wg} and the standard deviation across the eight scenarios at the Category level. It can be seen that the value of r_{wg} was fairly high for both the Captain and the F/O. For each of the Categories, the variance of the rating distributions was a mean of 76% smaller than the variance associated with a random response pattern.

[insert figure 8 here]

An examination of the mean r_{wg} scores for each scenario shows that there is little variation, with r_{wg} varying from between about 0.64 to approximately 0.87 for both the Captain and the F/O (see Figure 9). However, at the Pass/fail level there is a large variation in r_{wg} across the eight scenarios (see Figure 10). In general, there is either very high agreement (scenarios 2, 4, 5, 6, and 7) among raters or a very low level of agreement (scenarios 1, 3, and for FOs also 8).

[insert figure 9 here]

[insert figure 10 here]

As with the agreement with the reference ratings, at the Category level there were fairly high levels of inter-rater agreement. However, at the Pass/fail level the agreement between the raters was either very high, or very low.

User acceptability

The feedback about the NOTECHS rating system gleaned from the Evaluation Questionnaire was that the majority of raters were very satisfied with the system and thought it was useful. Over 95% of the sample thought that it was acceptable to evaluate pilots on their NTS. Further, of the 53% of raters who were familiar with other NTS rating systems, 82% thought that the NOTECHS system was superior. The vast majority of raters thought the division into four Categories and 15 Elements was satisfactory (88%), only 7% of raters thought some Categories or Elements were superfluous, and 98% thought the 5-point rating scales were satisfactory. Thus, the raters appeared to be very satisfied that the NOTECHS framework is a suitable system for assessing NTS behaviour in multi-pilot aircrew.

DISCUSSION

Internal consistency

The internal consistency of the system was high, with the ratings at the Category level generally being reflected by those at the Element and Pass/fail levels. This is a reassuring result as "one of the most difficult aspects in becoming proficient in CRM assessment is not in learning the individual elements but in compiling those elements into a meaningful hierarchy so that their relationship is understandable as well as usable" (Seamster & Edens, 1993: 126). As a result of this experimental work, it has been decided that the Elements will not be explicitly rated in the operational phase of the experiment in which instructors will be using the NOTECHS system to evaluate pilots in a simulator or in a line-flight.

Accuracy

At the Category level, the absolute difference method demonstrated that the participants found slightly greater difficulty in accurately assessing the Captain when compared to the F/O. Also, overall the Situation awareness and Decision making Categories were found to be the most difficult to rate accurately.

There was a tendency for raters not to use the Not observed rating, with the participants rating behaviours which the expert rating pilots did not judge to have occurred. This has implications for the training of instructors to use the system. However, it should be stressed that the videos were only very short in duration and, in the operational environment in which the system is designed to be used, instructors will be watching the crews for much longer than in video scenarios.

The individual scenarios were also examined. At the Category level, scenarios 1, 3, and 8 had the lowest levels of accuracy. At the Pass/fail level, the raters had greatest difficulty with scenarios 1 and 3 for both pilots and scenario 8 for the F/O. The difficulty in judging these scenarios is also echoed by the ratings of the two groups of experts who were used to calculate the reference ratings. The only occasions in which the two groups differed in their responses at the Pass/fail level were for the F/O in scenario 1, the Captain in scenario 3, and the F/O in scenario 8. Closer inspection of these scenarios reveals the complexity in their judgement (see method section for an outline of the scenarios). In scenario 1 the rater must decide how to separate the behaviours and responsibilities of the two pilots, in scenario 3 no conclusion is shown, and in scenario 8 the rater must judge how assertive the F/O can be without aggravating the situation further.

It seems likely that the short amount of training was not sufficient to allow the raters to judge these scenarios. Also the difference between pass and fail needs to be outlined more clearly.

Inter-rater agreement

At the Category level, the level of inter-rater agreement was high, the variance of the rating distribution was approximately 80% smaller than the variance associated with a random response pattern. Also, when the scenarios were examined individually, there was little variation in the mean inter-rater agreement for the Captain and the F/O. Generally, the values fell within the inter-rater reliability bench mark for agreement proposed by Williams, Holt and Boehm-Davis (1997) of r_{wg}= 0.7 to 0.8 when the Categories or the scenarios were examined separately. Thus, at the Category level there was a consistently high level of inter-rater reliability. However, the same was not true at the Pass/fail level. Again, the lowest level of agreement between raters was on scenarios 1 and 3 for both pilots and scenario 8 for the

F/O. However, generally the remaining inter-rater reliability scores were very high. The same explanation as with the accuracy of the raters can be used to explain the difficulty in judging these scenarios.

User acceptability

The raters in the experiment were generally positive about the NOTECHS system. In fact, a large proportion of those who were familiar with other NTS rating systems thought that NOTECHS was superior.

General discussion

At the Category level, the NOTECHS system has provided to be a usable and reliable assessment method for both the Captain and F/O in a controlled experimental condition. The results are promising for the next phase of the project to test the system in an operational setting. However, at the Pass/fail level the results were more mixed, with the raters having some difficulty reliably rating the three ambiguous scenarios, but performing with a very high level of reliability on the majority of the scenarios. As described above, the ambiguous scenarios had particular characteristics which were difficult to judge, and with more intensive training, and more complete sequences of interactive behaviour it is likely that the reliability would be improved. However, the purpose of the study was not to test the short amount of training that was delivered to the participants. Rather, the aim was to assess the NOTECHS system as a method of assessing the non-technical skills of commercial pilots. As would be expected, there was not complete agreement between the participants. However, this is unlikely to be true even in technical checks. Nevertheless, even after the short training that was given, the level of consistency appears to be high. A global Pass/fail decision is not inherent to the NOTECHS system and therefore of less relevance to the method and NTS-evaluation. Also, for the purpose of training the Pass/fail rating is of less importance than the ratings at the Category level.

Implications for training

The NOTECHS system provides a framework which allows an individual pilot’s NTS to be assessed. While it is recognised that many major carriers have already developed and are using behavioural marker systems, this does not appear to be the majority of operators. A survey of 11 UK airlines in 1997 showed that only 5 of them had developed a CRM behavioural markers list, and none of these were used for formal CRM assessment (Flin & Martin, 2001). Moreover, of the 104 training Captains who participated in the current experiment, only 53% were familiar with a NTS rating system, and only 31% had any pervious experience of evaluating NTS.

Therefore, there is an obvious need for a valid and reliable generic behavioural marker system which can be made available to those airlines which do not have the resources or expertise to develop their own systems. This will allow instructors to give structured feedback on pilots’ NTS, reinforce the importance of NTS to pilots, and fulfil the requirement to comply with the recent JAA legislation which asks for the assessment of NTS (JAR-OPS; Joint Aviation Requirements Flight Operations, 1999). It is anticipated that the NOTECHS system will allow these goals to be met.

It is also recognised that the instructors would require a more intensive period of training and calibration to use the system rather than the very short training given to the participants for rating the videos. The training should be designed to ensure that trainers are able to use the behavioural markers accurately and consistently, by reducing the likelihood of judgement biases and improve inter-rater reliability (Flin & Martin, 2001). Baker, Mulqueen, and Dismukes (1999) review a number of different approaches to training instructors to assess non-technical skills. Work in this area is of great relevance to the NOTECHS system, as if it is to be widely used, it will be necessary to establish an effective method of training instructors.

CONCLUSION

It is recognised that the NOTECHS system has only been subject to preliminary testing in an experimental setting. The next stage of the project is to evaluate the NOTECHS system in an operational setting. As mentioned above, the system will be used in much the same way as before, except the Elements will not be rated, and the scoreform has been adapted slightly. Prior instructor training will also be more intensive. However, there is evidence that some airlines are already adopting and modifying the NOTECHS method within their own training departments, which may provide opportunities for further evaluation.

In the opinion of the JAR TEL consortium, the results of the experimental phase of the JAR TEL project are quite encouraging for the further development and ultimate implementation of the NOTECHS method. The very high level of acceptance and approval shown by the instructor groups in the various areas of Europe gives grounds for some optimism in convincing the aviation fraternity that a method such as NOTECHS has a valuable part to play in the quest for enhanced levels of air safety.

REFERENCES

Avermaete, van, J.A.G. & Kruijsen, E. (1998) (Eds.) *The evaluation of non-technical skills of multi-pilot aircrew in relation to the JAR-FCL requirements*. (Project report: CR-98443). Amsterdam, The Netherlands: NLR.

Baker, D.P., Mulqueen, C. & Dismukes, K.R. (1999) Training pilot instructors to assess CRM: The utility of frame-of-reference (FOR) training. In *Proceedings of the International Aviation Training symposium* (pp. 291-300) Oklahoma City, Oklahoma.

Clothier, C. (1991) Behavioural interactions across various aircraft types: Results of systematic observations of line operations and simulators. In R. Jensen (Ed*.) Proceedings of the Sixth International Symposium on Aviation Psychology* (pp. 332-337) Columbus: Ohio State University.

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

Flin, R. & Martin, L. (1998) *Behavioural markers for crew resource management*. (Civil Aviation Authority, Paper 98005). London: Civil Aviation Authority.

Flin, R. & Martin, L. (2001) Behavioral markers for Crew Resource Management: A review of current practice. *International Journal of Aviation Psychology, 11,* 95-118.

Fowlkes, J., Lane, N., Salas, E., Franz, T. & Oser, R. (1994) Improving the measurement of team performance: The TARGETs methodology. *Military Psychology*, *6*, 47-61.

George Mason University (1996) *Improving Crew Assessments.* Training materials to accompany a FAA sponsored workshop on evaluator calibration. Fairfax, VA: Author.

Hamman, W.R., Beaubien, M.J, & Holt, R.W. (1999) Evaluating instructor/evaluator inter-rater reliability from performance database information. In R. Jensen (Ed*.) Proceedings* *of the 10th International Symposium on Aviation Psychology*, (pp. 1214-1219) Ohio: Ohio State University.

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

Helmreich, R. (2000) *The Line Operations Safety Audit (LOSA*). (Version 9). Austin: NASA/University of Texas/Federal Aviation Administration Aerospace Group.

Helmreich, R. (2000) Managing threat and error: Data from line operations. In B. Haniad (Ed) *Proceedings of the Fifth Australian Aviation Psychology Symposium*, Sydney, Australia

Helmreich, R., Butler, R., Taggart, W. & Wilhelm, J. (1995) *The NASA/University of Texas/Federal Aviation Administration Line/LOS checklist: A behavioural marker-based checklist for CRM skills assessment. Instructions for using the LLC*$_{v4}$. (Tech. Paper 42-02). Austin, TX: NASA/University of Texas/Federal Aviation Administration Aerospace Group.

Helmreich, R., Butler, R., Taggart, W. & Wilhelm, J. (1997) The NASA/University of Texas/Federal Aviation Administration Line/LOS checklist: A behavioural-based checklist for CRM skills assessment (Version 4.4)[computer software]. Austin, TX: NASA/University of Texas/Federal Aviation Administration Aerospace Group.

Helmreich, R. L., Klinect, J. R., & Wilhelm, J. A. (1999*). The line operations safety audit (LOSA) observer's manual, version 7.0* . (Tech. Rep. 99-0). Austin, TX: NASA/University of Texas/Federal Aviation Administration Aerospace Group.

Helmreich, R., Wilhelm, J., Kello, J., Taggart, E. & Butler, R. (1990*) Reinforcing and evaluating crew resource management: Evaluator/LOS instructor manual* (Tech. Manual 90-2). Austin, TX: NASA/University of Texas/Federal Aviation Administration Aerospace Group.

James, L.R., Demaree, R.G. & Wolf, G. (1984) Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, *69*, 85-98.

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

Joint Aviation Authorities (1999*). JAR-OPS. 1 Subpart N (NPA-OPS-16) Crew Resource Management- Flight crew*. Hoofddorp, The Netherlands: Author.

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

Seamster, T., & Edens, E. (1993). Cognitive modelling of CRM assessment expertise: Identification of the primary assessors. In L. Smith (Ed.), *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting* (pp. 122-126) San Diego, CA: Human Factors and Ergonomics Society.

Williams, D.M., Holt, R.W. & Boehm-Davies, D.A. (1997). Training for inter-rater reliability: Baselines and benchmarks. In R. Jensen (Ed..)*, Proceedings of the 9th International Symposium on Aviation Psychology* (pp. 514-519) Columbus: Ohio State University.

Figure 1. The NOTECHS hierarchy of levels

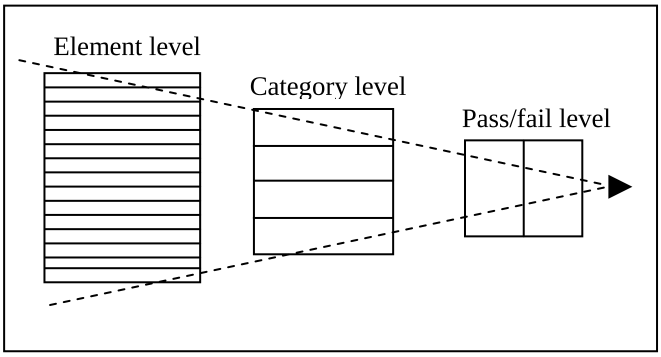


Table 1. The NOTECHS framework

Categories	***Elements***
Co-operation	Team building and maintaining Considering others Supporting others Conflict solving
Leadership and managerial skills	Use of authority Maintaining standards Planning and co-ordinating Workload management
Situation awareness	System awareness Environmental awareness Assessment of time
Decision making	Problem definition / diagnosis Option generation Risk assessment Outcome review

Table 2. Definition of NOTECHS ratings

Rating	*Definition*
Very poor	Behaviour directly endangered flight safety
Poor	In other conditions the behaviour could endanger flight safety
Acceptable	Behaviour does not endanger flight safety, but needs improvement
Good	Behaviour enhances flight safety
Very good	Behaviour optimally enhances a flight safety and could be an example for other pilots

Figure 2. Mean and standard deviation of the absolute differences between Element and Category levels.

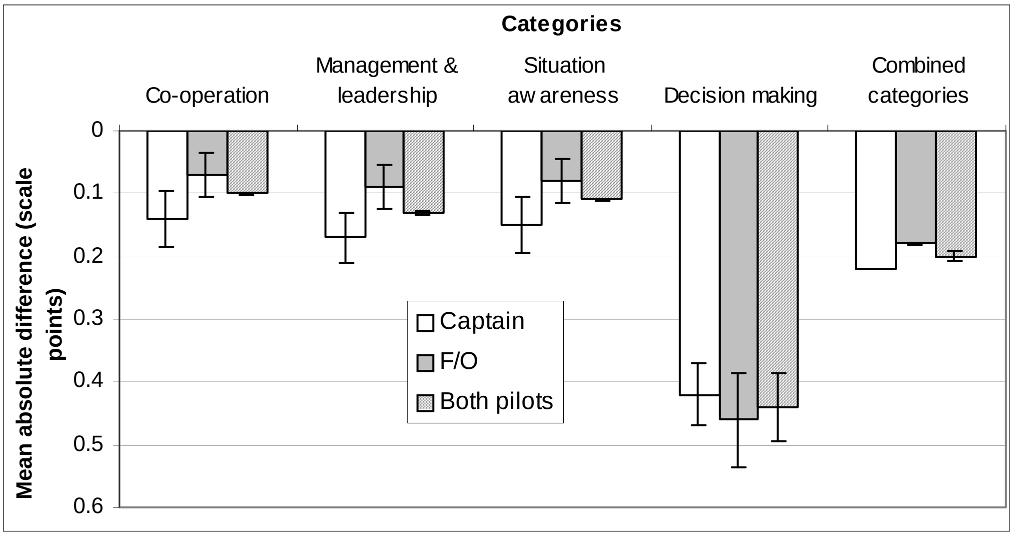


Figure 3. Mean and standard deviation of the consistency between the Category and Pass/fail level.

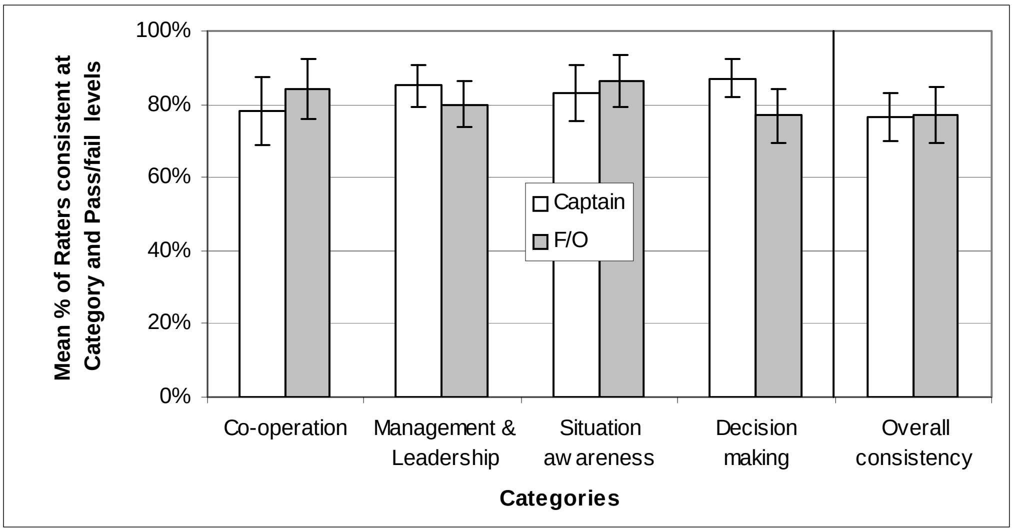
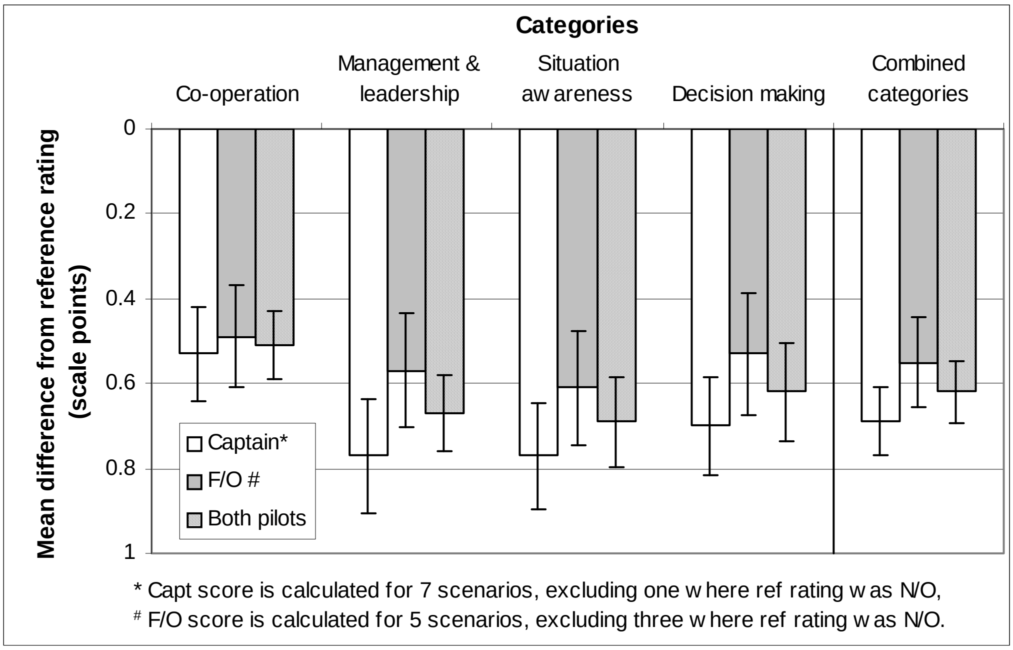


Figure 4. Mean and standard deviation of the absolute difference between reference rating and raters’ responses at the Category level.



* Capt score is calculated for 7 scenarios, excluding one w here ref rating w as N/O,
[#] F/O score is calculated for 5 scenarios, excluding three w here ref rating w as N/O.

Figure 5. Mean absolute difference between reference rating and raters’ responses at the Category level for each scenario.

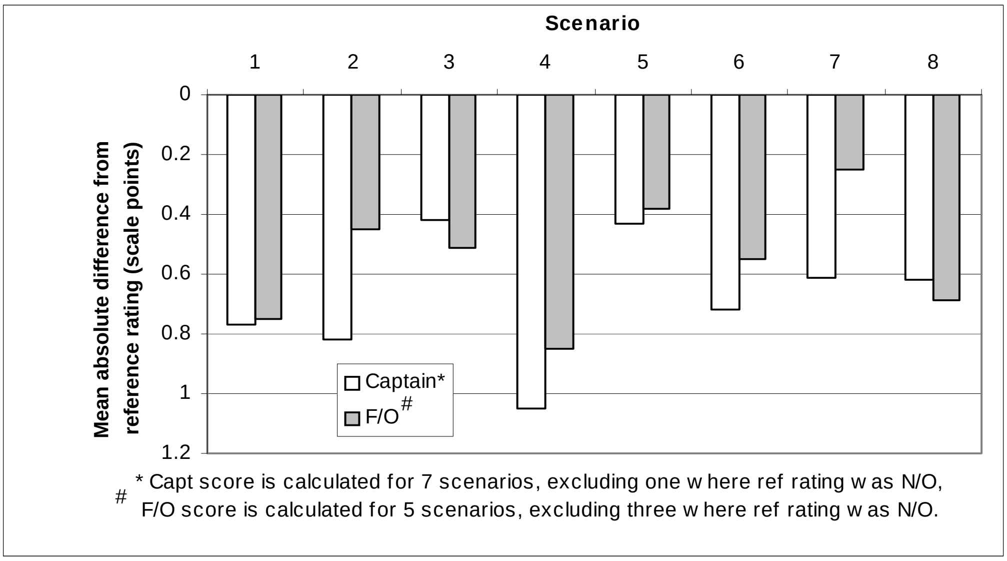


Figure 6. Mean percentage of raters agreeing with the reference rating at the Category level for each scenario using the collapsed two-point scale.

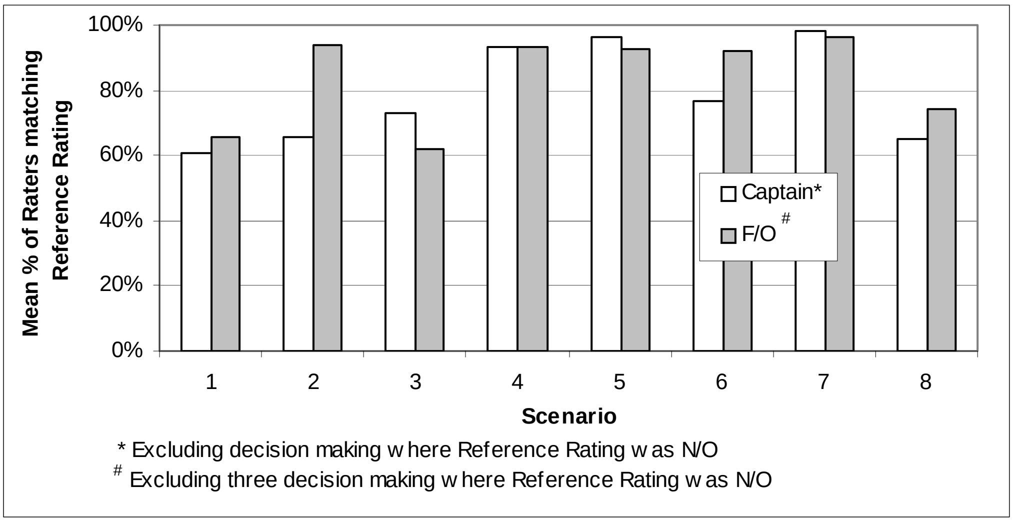


Figure 7. Percentage of raters agreeing with the reference rating at Pass/fail level.

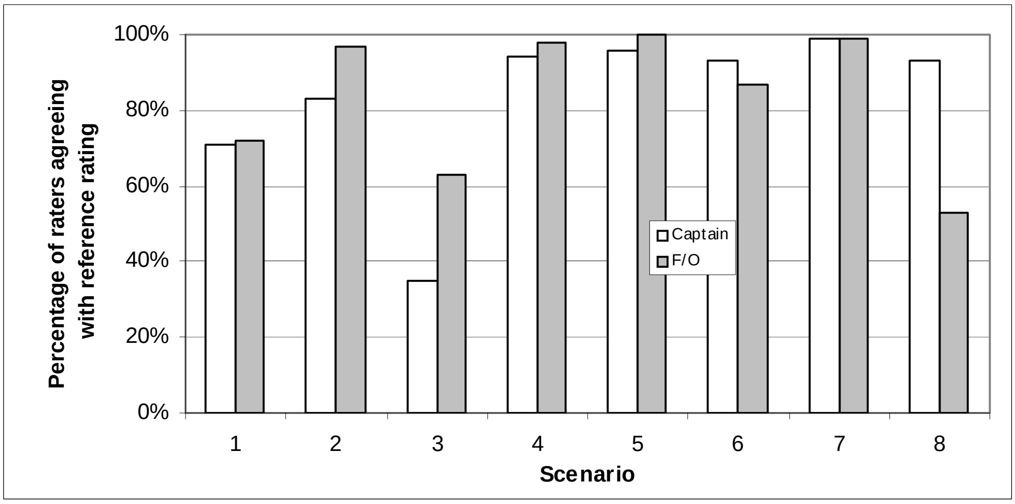


Table 3. Number of participants who agreed and disagreed with the Pass/fail reference rating and goodness-of-fit statistic.

	Captain				First Officer			
Scenario	Agree	Disagree	Chi^2	Sig.	Agree	Disagree	Chi^2	Sig.
1	74	30	18.6	>.05	75	28	21.5	>.05
2	86	18	44.5	>.05	101	3	92.4	>.05
3	36	68	9.3	>.05	65	39	6.5	>.05
4	98	6	81.4	>.05	102	2	96.2	>.05
5	99	4	87.6	>.05	103	0	Not valid	
6	96	7	76.9	>.05	90	11	61.8	>.05
7	102	1	99.0	>.05	102	1	99.0	>.05
8	96	6	79.4	>.05	55	47	0.63	n.s

Figure 8. Mean inter-rater agreement at the Category level.

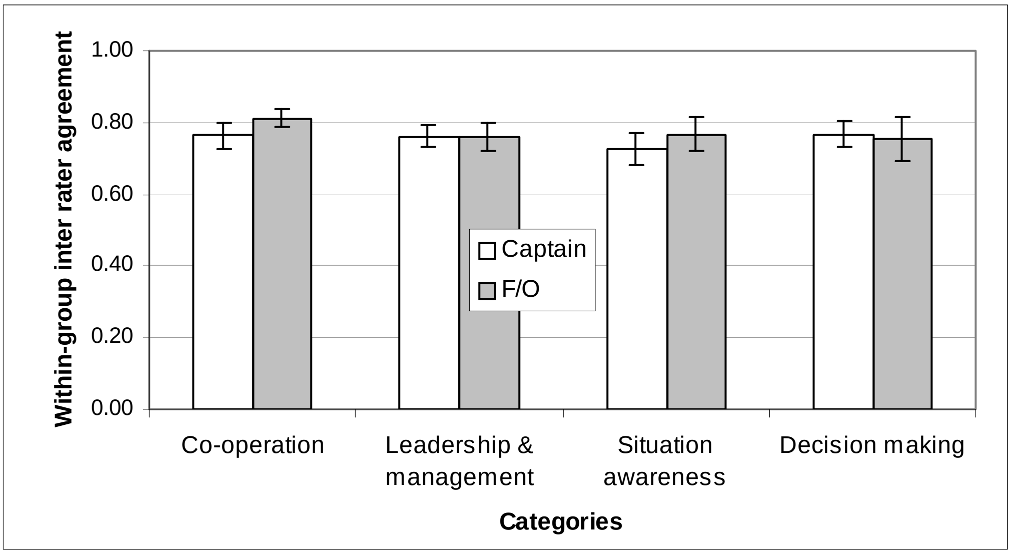


Figure 9. Mean inter-rater agreement for each scenario at the Category level.

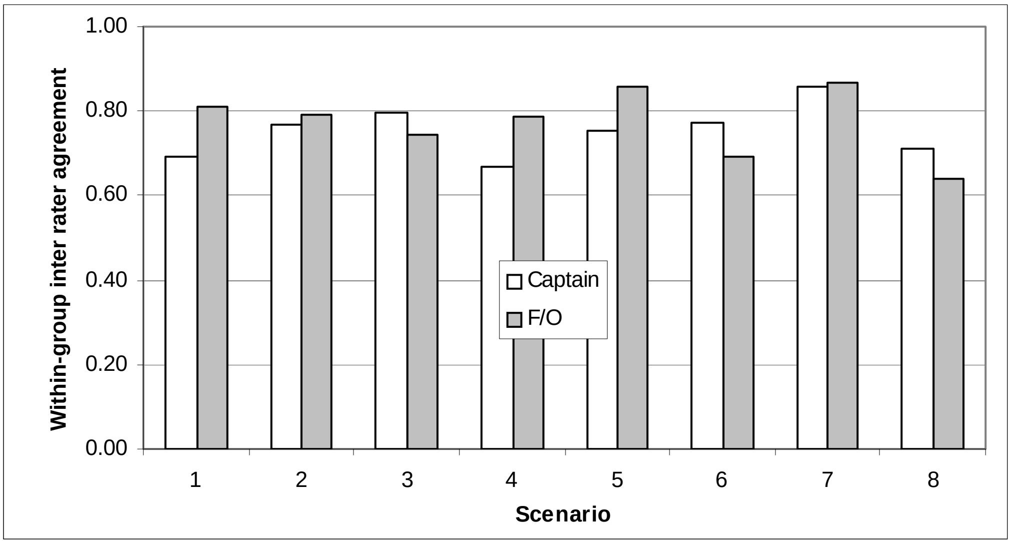
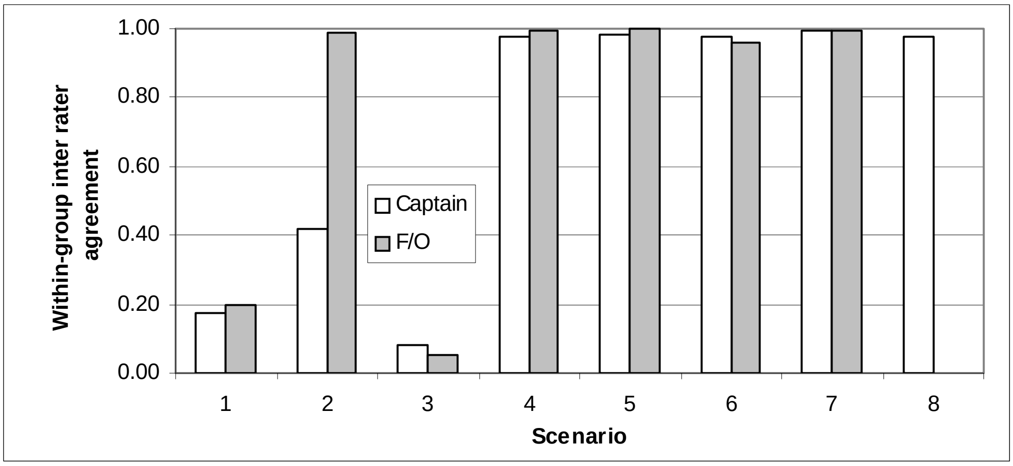


Figure 10. Inter-rater agreement for each scenario at the Pass/fail level.


Within-group inter rater agreement
□ Captain
▨ F/O
Scenario