Developing a Method for Evaluating Crew Resource Management Skills: A European Perspective

Paul O’Connor, Hans-Jurgen Hörmann, Rhona Flin, Mike Lodge, Klaus-Martin Goeters & The JARTEL Group

Published online: 13 Nov 2009.


To link to this article: http://dx.doi.org/10.1207/S15327108IJAP1203_5
Developing a Method for Evaluating Crew Resource Management Skills: A European Perspective

Paul O’Connor
University of Aberdeen
Aberdeen, Scotland

Hans-Jurgen Hörmann
German Aerospace Centre (DLR)
Hamburg, Germany

Rhona Flin
University of Aberdeen
Aberdeen, Scotland

Mike Lodge
British Airways
London, England

Klaus-Martin Goeters
German Aerospace Centre (DLR)
Hamburg, Germany

The JARTEL Group
Rene Amalberti, Eric Andlauer, Andrea Boiardi, Marie-Claude Delsart, Adrian Ellis, Georgina Fletcher, Rhona Flin, Klaus-Martin Goeters, Hans-Jurgen Hörmann, Mike Lodge, Herman Nijhuis, Paul O’Connor, Lucio Polo, Susanna Russell, Jean-Jacques Speyer, Claude Valot, and Renate Verkaik

Requests for reprints should be sent to Rhona Flin, Psychology Department, Aberdeen University, Aberdeen, England, AB24 2UB. E-mail: r.flin@abdn.ac.uk
The European Commission in conjunction with the European Joint Aviation Authorities (Human Factors Project Advisory Group) has been sponsoring a series of studies investigating a culturally robust method for the evaluation of pilots’ nontechnical skills (NTS) for multicrew operations. This article outlines the development of a European NTS behavioral marker system for crew resource management evaluation called NOTECHS and presents preliminary results from an ongoing test phase of this system (Joint Aviation Requirements Translation and Elaboration of Legislation [JARTEL]). The JARTEL project has involved 105 instructors from 14 European airlines who were given a short training session to use the NOTECHS system. After the training phase, these instructor pilots used the system to evaluate the individual CRM skills of captains and first officers in 8 different video scenarios filmed in a Boeing 757 simulator. Issues relating to rater training, reliability, and accuracy, as well as the instructors’ opinions on the acceptability of the method, are discussed.

The area of nontechnical skills (NTS) evaluation in Europe has become increasingly important in the light of recent Joint Aviation Authorities (JAA) legislation that asks for the assessment of NTS (Joint Aviation Requirements for Flight Operations [JAR-OPS], 1999). Although the regulations make recommendations concerning what should be included in crew resource management (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 JAA Project Advisory Group for Human Factors tasked four research institutes (Dutch Aerospace Research Center, German Aerospace Research Centre, Institut de Medecine Aerospatiale du Service de Sante des Armees, and the University of Aberdeen) to develop an NTS assessment framework that became known as NOTECHS. The Joint Aviation Requirements and Translation Elaboration of Legislation (JARTEL) project was born out of the NOTECHS project with the aim of assessing the usability and validity of the set of behavioral markers established in NOTECHS through both experimental and operational evaluation.

This article provides some background on existing behavioral marker systems, outlines the NOTECHS framework, describes the preliminary testing of the system in an experimental setting, and discusses the implications of NOTECHS for training.

BEHAVIORAL MARKER SYSTEMS

Only a limited amount of research has been conducted into the development of behavioral marker systems for the evaluation of pilots’ NTS (for a review, see
Flin & Martin, 2001). The seminal research on behavioral markers comes from Helmreich’s group at the University of Texas/National Aeronautics and Space Administration (NASA)/Federal Aviation Administration 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 behaviors 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’ behavioral marker systems (Flin & Martin, 2001). The LLC system has been refined over the years on the basis of ongoing observational research (Clothier, 1991; Helmreich, 2000a, 2000b) and was recently integrated into the current version of Line Operations Safety Audit (LOSA). Version 9.0 (Helmreich, Klinect, & Wilhelm, 1999) elicits ratings in three broad categories (planning, execution, and review/modify plans) from four phases of flight (predeparture, takeoff 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. Inasmuch as LLC and, especially, LOSA were designed to provide a wide range of safety-related indications of the respective organization, the entity of analysis in these systems is not the individual crew member per se but the organization itself or parts of it (e.g., fleets and operational units).

Another earlier marker system developed by Fowlkes, Lane, Salas, Franz, and Oser (1994) was a team performance measurement approach called Targeted Acceptable Responses to Generated Events or Tasks (TARGETs) for U.S. military cargo helicopter teams. TARGETs was based on a set of critical aircrew cooperation behaviors grouped into seven basic skill areas: mission analysis, adaptability and flexibility, leadership, decision making, assertiveness, situation awareness, and communication. In this system, for each stimulus event in a scenario, there is a predefined set of acceptable behaviors, each of which is rated as present or absent. As with the LLC, this is a measure of crew performance rather than individual performance. Fowlkes et al. 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 interrater reliability (IRR).

Many of the large airlines have also developed their own behavioral marker systems that are primarily used for training (van Avermaete & Kruijsen, 1998; Flin & Martin, 2001). To aid in this, a research group at George Mason University has provided assistance to companies in the development of a CRM skills-evaluation system and training for instructors to use it (George Mason University, 1996). Many of the smaller companies in Europe, however, do not have the time, resources, or expertise to develop their own systems. Thus, with the recent change in the JAA regulations described previously, it was recognized that there was a need for a basic, generic system that was not specific to any one company, country, or type of operation and allowed the pilots’ NTS to be assessed individually rather than as a crew.
THE NOTECHS FRAMEWORK

The NOTECHS project was sponsored by the European Community Directorate for Transport and the Environment (EC DGTREN) and the Civil Aviation Authorities of France, The Netherlands, Germany, and the United Kingdom and ran from March 1997 to March 1998 (van Avermaete & Kruijsen, 1998). The central goal of the NOTECHS project was to provide guidance for a feasible and efficient method for assessing pilots’ NTS by instructor pilots and examiners during training and check events in multicrew aircraft in countries across Europe. The method was to be based on common elements of NTS training and evaluation systems that were in use at European airlines such as Lufthansa and KLM (Royal Dutch Airlines). The first stage was to review existing systems of evaluating NTS. Flin and Martin (1998, 2001) surveyed 12 U.K. 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, and the Dutch CAA (van 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 appear to be related to decision making, situation awareness, leadership, and teamwork.

After the survey, a literature review of relevant research findings related to these key categories of NTS identified in the survey was performed (van 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, and that no single system provided 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/University of Texas Line/Line Oriented Simulation Checklist (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 to encompass the critical behaviors.
• The terminology used should reflect unambiguous everyday language for behavior, rather than psychological jargon.
• The skills listed at the behavior level should be directly observable in the case of social skills or could be inferred from communication in the case of cognitive skills.

The NOTECHS framework consists of a hierarchy of three levels: elements, categories, and pass/fail (Figure 1). On the basis of the individual behavior ratings at the element level, the user formulates the ratings at the category level, which finally leads to a pass or fail judgment (i.e., the recommendation of further training).

The primary category level can be divided into two social skills (cooperation and leadership and management skills) and two cognitive skills (situation awareness and decision making). This elemental set was based on theoretical models identified from the literature review (van 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 (Table 1). For each element, a number of positive and negative exemplar behaviors were included, again devised from the literature review and existing systems (Flin & Martin, 1998). The exemplar behaviors were phrased as generic (e.g., closes loop for communication) rather than specific (e.g., reads back to air traffic control).

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 that is inherent in all four categories. A category of personal limitations (e.g., stress and fatigue) was also rejected because of the difficulty in observing it except in the most extreme of cases.

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

THE JARTEL PROJECT

The JARTEL project can be divided into two main stages: (a) the experimental phase, in which the usability and cultural robustness of the NOTECHS framework was tested by using video scenarios and (b) the operational phase, in which the system was used in real life by instructors to assess NTS. This article concentrates on the experimental phase of the project to assess the usability of the NOTECHS system in a controlled setting.

To test the NOTECHS framework, a 5-point rating scale at the element and category levels was used. Criteria were developed for deciding which rating should be awarded at the element and category levels (Table 2).
If a behavior was not observed, then a not-observed response was recorded. At the pass/fail level, a rating of fail was given if the pilot displayed overall behavior related to the NTS that endangered or could endanger flight safety, and a rating of pass was given if the pilot’s overall behavior did not endanger flight safety. Specific exemplar behaviors were not specified at each of the 5 points for each exemplar element because 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 performed by determining the extent to which ratings at the category level were in line with those at the element and pass/fail levels.

• 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 section). 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.

• Interrater agreement was measured using an index, developed by James, Demaree, and Wolf (1984, 1993), called the within-group IRR 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 among the raters, then \(r_{wg} = 1\). This measure was selected because it has been used to assess the IRR of behavioral marker systems in aviation by Law and Sherman (1995) and Hamman, Beaubien, and Holt (1999).

• Acceptability of the system was assessed using the feedback from the raters on the evaluation questionnaire.

### TABLE 2
Definition of NOTECHS Ratings

<table>
<thead>
<tr>
<th>Rating</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very poor</td>
<td>Behavior directly endangered flight safety</td>
</tr>
<tr>
<td>Poor</td>
<td>In other conditions the behavior could endanger flight safety</td>
</tr>
<tr>
<td>Acceptable</td>
<td>Behavior does not endanger flight safety but needs improvement</td>
</tr>
<tr>
<td>Good</td>
<td>Behavior enhances flight safety</td>
</tr>
<tr>
<td>Very good</td>
<td>Behavior optimally enhances a flight safety and could be an example for other pilots</td>
</tr>
</tbody>
</table>
METHOD

Before performing 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. To minimize the risk of the raters typecasting 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 preexperiment briefings given to the participants.

Eight scenarios were used in the main experiment (average length 7 min; range 3–15 min), chosen from a total of 15 that were filmed. The scenarios were designed by a training captain and a psychologist from the consortium to demonstrate a range of realistic situations, and although the scenarios were not scripted to the level of prescribing exactly what should be said, each scenario had a set of design references that were levels of NTS for each behavior category (on the 5-point scale) that the pilot actors were supposed to illustrate.

1. Descent: The F/O is the pilot flying and a passenger problem is reported by the cabin crew. The action centers 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 predeparture 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 underconfident F/O and achieves a very positive result after good training input.

6. A night approach in the mountains: The captain decides to carry out a visual approach through high terrain and triggers a ground proximity warning system (GPWS) warning. The F/O takes control and prevents an accident.

7. An automatic approach in CAT III conditions (foggy): 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 benchmarks or reference data were required for the analysis process 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 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 after a group discussion arrived at a consensus rating for each of the categories and pass/fail judgments.

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 × 4 categories × 8 scenarios = 64). This was across the acceptable/poor divide in only 6% of the total evaluations made; thus, the remaining discrepancy was either at the poor/very poor level or the acceptable/good/very good level, which are fine-grained judgments to make. At the pass/fail level, the raters agreed in 81% of evaluations (2 pilots × 8 scenarios = 16). In the cases in which 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 behavior specification from the original script that 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 the 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 men with an average of 6 years as an instructor (SD = 6.2) and an average total of 10,200 flying hr (SD = 3,852).

Procedure

Groups of raters recruited from each airline participated in the experiment during 1 full day. The same pilot facilitator was involved in every session and assisted by at least one consortium psychologist. All participants were already briefed on
the background of the experiment and on the NOTECHS method by written material distributed in advance.

As a result of 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 accommodate a 3-hr time frame, followed by the 3-hr 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 JARTEL experiment and were asked to fill out a background questionnaire to gather data about their professional background such as age, experience in 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 performed in a controlled manner using the training video and an interactive question-and-answer session. At the end of the training video, raters further practiced 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. Last, open discussions were conducted for debriefing on general feelings to achieve knowledge on the context and collect qualitative data for the understanding of the results.

**RESULTS**

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

**Internal Consistency**

An assessment was made on 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 performed by calculating the mean difference among each of the three or four elements and their corresponding categories 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 because of the dichotomy of the pass/fail response. Therefore, it was necessary to collapse the category level responses into a 2-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 ($M < 0.2$ of a scale point between the element and category) on all of the categories except for decision making, but even for decision making the mean absolute difference between the element and the category is less than 0.5 on a 5-point scale.

A 2-factor (pilot and categories) repeated measures analysis of variance (ANOVA) was run using the mean absolute difference scores of the difference among 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 among the elements and corresponding categories was significantly higher for the F/O (.18) than the captain (.22). It should be noted that

**FIGURE 2** Mean and standard deviation of the absolute differences between the element and category levels.
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 cooperation (.10) and situation awareness (.11). From Figure 2, it can be seen that the interaction is attributable 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. In the decision-making category, the consistency is lower for the F/O than the captain. This is confirmed by looking at the contrasts among the variables.

The consistency among 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 of the four categories are consistent with the pass/fail response.

A 2-factor (pilot and categories) repeated measures ANOVA was run by adding the number of matches among the categories (on the collapsed 2-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, \( F(1, 102) = 1.07, ns \) and category, \( F(2.74, 279.8) = 2.49, ns \), 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

![Figure 3](image-url)
the cooperation and situation awareness categories. For both of these categories the match with the pass/fail level was higher for the F/O than for the captain. However, the reverse was true for leadership and management and decision making, thus creating an interaction (Figure 3). In addition, there was an interaction between leadership and management and decision making because the difference between the captain and F/O was significantly greater for decision making than for leadership and management.

Therefore, to summarize the consistency of the participants’ ratings of the elements and categories and categories and pass/fail appears to be fairly high. However, looking at the absolute differences between ratings of the elements and categories, it can be seen that the decision-making category shows the 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 eight 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 5-point scale. The category level responses were also examined by collapsing the scale into a 2-point scale 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 of 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 cooperation category was the most accurately rated, with the remaining three categories all having a similar level of accuracy.

A 2-factor (pilot and categories) repeated measures ANOVA 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, $F(2.7, 277.6) = 7.2$, $p < .05$. Accuracy was significantly superior for the F/O (.55) than for the captain (.69). Examining the categories individually, it was found that they all were significantly different from each other except for leadership and management (.67) and situation awareness (.69), and leadership and management and decision making (.62). It can also be seen that the greatest inaccuracy was in the situation awareness and decision-making categories. In
regard to the contrasts among the variables, it was found that the interaction was due to the small difference (compared with the other variables) in the cooperation category between the rating accuracy of the captain and the F/O compared with the greater accuracy of rating the F/O when compared with 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 for the F/O in Scenarios 1 and 4; however, this was not reflected in the collapsed scale method (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 among acceptable, good, and very good or poor and very poor.

At the pass/fail level, the mean accuracy of the raters was 83% across the eight scenarios for the captain (SD = 11.4), and 84% for the F/O (SD = 12.7). A repeated measures $t$ test indicated that this difference was not significant $t(103) = -0.67$, ns. 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 were lower.
than compared with the other scenarios (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 that disagreed with the reference rating (Table 3). Only in Scenario 3 for the captain did a significantly larger proportion of raters disagree with the reference rating, $\chi^2(1) \geq 3.8, p < .05$;
in Scenario 8 for the F/O, the difference between agree and disagree was not high enough to be significant, $\chi(1) < 3.8$, ns (Table 3).

The agreement with the not-observed reference rating was not included in the analysis described previously. This was because of 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

FIGURE 7 Percentage of raters agreeing with the reference rating at the pass/fail level.

TABLE 3
Number of Participants Who Agreed and Disagreed With the Pass/Fail Reference Rating and Goodness-of-Fit Statistic

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Captain Agree</th>
<th>Captain Disagree</th>
<th>$\chi^2$</th>
<th>Significant</th>
<th>First Officer Agree</th>
<th>First Officer Disagree</th>
<th>$\chi^2$</th>
<th>Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>74</td>
<td>30</td>
<td>18.6</td>
<td>&gt;.05</td>
<td>75</td>
<td>28</td>
<td>21.5</td>
<td>&gt;.05</td>
</tr>
<tr>
<td>2</td>
<td>86</td>
<td>18</td>
<td>44.5</td>
<td>&gt;.05</td>
<td>101</td>
<td>3</td>
<td>92.4</td>
<td>&gt;.05</td>
</tr>
<tr>
<td>3</td>
<td>36</td>
<td>68</td>
<td>9.3</td>
<td>&gt;.05</td>
<td>65</td>
<td>39</td>
<td>6.5</td>
<td>&gt;.05</td>
</tr>
<tr>
<td>4</td>
<td>98</td>
<td>6</td>
<td>81.4</td>
<td>&gt;.05</td>
<td>102</td>
<td>2</td>
<td>96.2</td>
<td>&gt;.05</td>
</tr>
<tr>
<td>5</td>
<td>99</td>
<td>4</td>
<td>87.6</td>
<td>&gt;.05</td>
<td>103</td>
<td>0</td>
<td>Not valid</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>96</td>
<td>7</td>
<td>76.9</td>
<td>&gt;.05</td>
<td>90</td>
<td>11</td>
<td>61.8</td>
<td>&gt;.05</td>
</tr>
<tr>
<td>7</td>
<td>102</td>
<td>1</td>
<td>99.0</td>
<td>&gt;.05</td>
<td>102</td>
<td>1</td>
<td>99.0</td>
<td>&gt;.05</td>
</tr>
<tr>
<td>8</td>
<td>96</td>
<td>6</td>
<td>79.4</td>
<td>&gt;.05</td>
<td>55</td>
<td>47</td>
<td>0.63</td>
<td>ns</td>
</tr>
</tbody>
</table>
be seen that there was a tendency for the participants to rate behaviors that were not judged to be present by the experts.

At the pass/fail level, inspection of the results indicates that the captains were assessed independently from the F/Os. In the three scenarios in which 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% to 95% (Figure 7). There is a higher level of disagreement in Scenario 8 in which 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 summarize, overall there was a high level of agreement between the participants and the experts at the category level. However, at the pass/fail level in 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.

Interrater Agreement

The $r_{wg}$ was used to analyze the interrater agreement at both the category (using the 5-point scale) and pass/fail levels (using the 2-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.
An examination of the mean $r_{wg}$ scores for each scenario shows that there is little variation, with $r_{wg}$ varying from approximately .64 to .87 for both the captain and the F/O (Figure 9). However, at the pass/fail level there is a large variation in $r_{wg}$ across the eight scenarios (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 and 3 and for F/Os also Scenario 8).

Similar to the agreement on the reference ratings, there were fairly high levels of interrater agreement on the category level; however, at the pass/fail level the agreement between the raters was either very high or very low.

![Figure 9](image_url)  
**FIGURE 9** Mean interrater agreement for each scenario at the category level.

![Figure 10](image_url)  
**FIGURE 10** Interrater agreement for each scenario at the pass/fail level.
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. More than 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 behavior 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, p. 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 experienced slightly greater difficulty in accurately assessing the captain when compared with assessing the F/O. 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 behaviors that the expert-rating pilots did not judge to have occurred. This has implications for the training of instructors to use the system. It should be stressed, however, that the videos were very short in duration and that 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 judgment (see Method section for an outline of the scenarios). In Scenario 1, the rater must decide how to separate the behaviors 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.

Interrater Agreement

At the category level, the level of interrater agreement was high, and 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 interrater agreement for the captain and the F/O. Generally, the values fell within the IRR benchmark for agreement proposed by Williams, Holt, and Boehm-Davis (1997) of \( r_{wg} = .7 \) to .8 when the categories or the scenarios were examined separately. Thus, at the category level there was a consistently high level of IRR. However, the same was not true at the pass/fail level. Again, the lowest level of agreement among raters was on Scenarios 1 and 3 for both pilots and Scenario 8 for the F/O. Generally, the remaining IRR scores were very high. The same explanation as for 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 proved 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. 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 for the majority of the scenarios. As described previously, the ambiguous scenarios had particular characteristics that were difficult to judge. With more intensive training and more complete sequences of interactive behavior, it is likely that the reliability could be improved. The purpose of the study was not to test the short amount of training that was delivered to the participants but rather to assess the NOTECHS system as a method of assessing the NTS of commercial pilots. As would be expected, there was not complete agreement among 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 that allows an individual pilot’s NTS to be assessed. Although it is recognized that many major carriers have already developed and are using behavioral marker systems, this does not appear to be the case for the majority of operators. A survey of 11 U.K. airlines in 1997 showed that only 5 of them had developed a CRM behavioral markers list, and none of these were used for formal CRM assessment (Flin & Martin, 2001). Moreover, of the 104 training captains who participated in this experiment, only 53% were familiar with an NTS rating system, and only 31% had any previous experience of evaluating NTS.

Therefore, there is an obvious need for a valid and reliable generic behavioral marker system that can be made available to those airlines that 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 fulfill the requirement to comply with the recent JAA legislation that asks for the assessment of NTS (JAR-OPS, 1999). It is anticipated that the NOTECHS system will allow these goals to be met.

It is also recognized 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 behavioral markers accurately and consistently by reducing the likelihood of judgment biases and improving IRR (Flin & Martin, 2001). Baker, Mulqueen, and Dismukes (1999) reviewed a number of different approaches to training instructors to assess NTS. Work in this area is of great relevance to the NOTECHS system because if it is to be widely used, it will be necessary to establish an effective method of training instructors.
CONCLUSIONS

It is recognized 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 previously, the system will be used in much the same way as before except the elements will not be rated and the score form adapted slightly. Prior instructor training will also be more intensive. 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 JARTEL consortium, the results of the experimental phase of the JARTEL 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.

ACKNOWLEDGMENTS

This study forms part of the JARTEL project (Contract No. AI-97-SC.2037) sponsored by EC DGTREN. The views expressed are those of the authors and should not be taken to represent the position or policy of the funding body. We thank all the airlines and especially the instructors who participated in the experimental sessions. We also thank the European Cockpit Association and JAA representatives, who provided advice and support, and the EC JARTEL Project Officer C. North.

REFERENCES


Manuscript first received June 2001