THE INTERNATIONAL JOURNAL OF AVIATION PSYCHOLOGY, *11*(1), 95–118

Behavioral Markers for Crew Resource Management: A Review of Current Practice

Rhona Flin
Department of Psychology
University of Aberdeen
King's College
Aberdeen, Scotland

Lynne Martin
NASA Ames Research Center
Moffett Field, CA

Developments in crew resource management (CRM) have progressed from the introduction of training programs to the evaluation of CRM skills, particularly for multicrew cockpits. European regulators responsible for flight operations and flight crew licensing (Joint Aviation Authorities, 1996, 1997) are introducing requirements for the training and assessment of pilots' nontechnical skills. This article reports a review of the literature and a survey of current practice in the development and use of behavioral marker systems for training and assessing nontechnical CRM skills in international and domestic (UK) airlines. In general, there appears to be a wide range of practice in the design and implementation of behavioral markers systems within CRM programs. Emerging issues relating to content validity of marker systems and rater reliability are likely to become the focus of both researchers' and pilots' interest.

Initial efforts to assess the value and impact of crew resource management (CRM) have consisted of standard training evaluation techniques based on pilots' opinions of the quality and relevance of the programs, in the United States (Gregorich & Wil-

Requests for reprints should be sent to Rhona Flin, Department of Psychology, University of Aberdeen, King's College, Old Aberdeen AB24 2UB, Scotland. E-mail: r.flin@abdn.ac.uk

helm, 1993) and in Europe (Maschke, Goeters, Hormann, & Schiewe, 1995; Naef, 1995). Additional research has indicated that pilots' attitudes toward accident related behaviors were improved by CRM training (see Helmreich, Merritt, & Wilhelm, 1999). These are important measures, but they are not a robust test of the effectiveness of CRM, which was introduced with the aim of improving flight safety and minimizing accident rates. The required proof that aviation accidents have been reduced as a result of CRM remains elusive (but, see Diehl, 1991), as incidence rates are very low, making any trends directly attributable to CRM difficult to detect (Helmreich et al., 1999). To assess whether CRM training is transferring to the flight deck, evaluation efforts are now targeted to observe the practice of CRM skills (Holt, Boehm-Davis, & Beaubien, in press; Salas, Fowlkes, Stout, Milanovich, & Prince, 1999) and the development of reliable, valid measures for assessing a crew's or a pilot's nontechnical skills.

Part of this drive for CRM assessment has come from the advanced qualification program (AQP) adopted by several U.S. airlines. These carriers undertake a full technical and nontechnical skills analysis, provide CRM and line oriented flight training (LOFT) for all flight crews, and undertake evaluation of CRM skills using line oriented evaluation (LOE) in full mission simulation (Birnbach & Longridge, 1993; Lanzano, Seamster, & Edens, 1997). They have developed core lists of CRM knowledge and skills (e.g., Boehm-Davis, Holt, & Seamster, in press) and may include key CRM behaviors on their flight deck checklists (e.g., Helmreich, 1996).

In the UK, human factors training and examination are required for a flight crew licence, and CRM training has been a mandatory requirement for commercial pilots since January, 1995. There is no requirement to formally assess CRM skills, and any evaluation is currently done on a voluntary basis by the airlines, although some guidance is available (Civil Aviation Authority [CAA], 1998). In Europe, Joint Aviation Requirements Flight Operations (JAR–OPS) and Joint Aviation Requirements Flight Crew Licensing (JAR–FCL) require CRM training and evaluation of CRM skills in multicrew operations (Joint Aviation Authorities [JAA], 1996, 1997; see Goeters, 1998). This indicates a need for a set of fundamental CRM (nontechnical) skills and the relevant markers to ensure an equitable system (CAA, 1998).

The term *behavioral markers* refers to a prescribed set of behaviors indicative of some aspect of performance. Typical behaviors are listed in relation to component skills and are now used for selection, training, and competence assessment in professions such as anesthesiology (Gaba et al., 1998). Although there is considerable interest by the civil aviation community in the identification and assessment of CRM skills, there is limited research literature on this topic. The aim of this study was to examine current practice in the development and use of behavioral markers for training and assessing CRM skills. A literature review was undertaken, followed by surveys of airlines in the UK and abroad.

LITERATURE REVIEW

The literature review concentrated on two main themes: the development of marker systems and the reliability and training of the raters using them.

Marker Systems: Research and Development

The seminal research on behavioral markers produced the Line/Los Checklist (LLC; Helmreich, Wilhelm, Kello, Taggart, & Butler, 1990), which is widely cited in the literature and is the basis of many airlines' CRM behavioral marker lists. It is used during in-flight observations to evaluate nontechnical CRM skills in human factors' line audits carried out for major airlines (Helmreich, Hines, & Wilhelm, 1996; Taggart, 1995). The behaviors included on the LLC have their origin in the analysis of accidents and incidents with identifiable human factors causation (e.g., Connelly, 1997), as well as supporting evidence from psychological research. Version, LLC4.4 (Helmreich, Butler, Taggart, & Wilhelm, 1997) elicits ratings for four phases of flight, under six categories of behaviors: team management and crew communications, situational awareness and decision making, automation management, special situations, technical proficiency, and overall observations. This gives a total of 28 behavioral marker elements and two overall evaluation measures, all of which are rated on a 4-point scale ranging from 1 (*poor*), 2 (*minimum expectations*), 3 (*standard*), to 4 (*outstanding*). It should be noted that the LLC is used to evaluate the crew's performance, rather than that of an individual pilot in a crew setting. (There is a separate section of the form for comments on a particular flight crew member.)

Using an earlier version of the LLC, Helmreich, Wilhelm, Gregorich, and Chidester (1990) found high degrees of variation in CRM performance ratings for crews flying different types of aircraft within the same airline. Across two airlines, different behaviors were linked (through their ratings) to superior performance. In Airline 1, inquiry, technical skills, advocacy, and decision making were correlated with ratings of above average performance, whereas for Airline 2, superior performance was associated with briefings and concern for the group. The researchers could not establish from their data whether these differences were due to true organizational differences or to different emphasis on particular aspects of the CRM courses during training.

Butler (1991) used the LLC and compared four U.S. airlines through 108 observations on overall technical efficiency and overall crew effectiveness. He found a wide range of performance, as well as significant differences between airlines. He also reported that during training to use the LLC, trainees' evaluations of the same crew's performance could vary widely, leading to his conclusion that standardization across raters is vital before the validity of CRM assessment can be properly gauged.

Law and Wilhelm (1995) used the LLC4 to collect 1,495 instructor observations from two airlines. Some elements were rated across all phases of flight, whereas others were rated only during certain phases or were rarely rated. They found specific crew behaviors were differentially related to crew effectiveness at varying phases of the flight and suggested that data should be collected for each phase of flight. Significant differences were found between the ratings from the two airlines on 19 of the elements and also between fleets within the same company. They concluded that the LLC4 is sensitive enough to detect reliable teamwork and performance differences between and within organizations.

Besides the LLC, several other marker systems for assessing flight crew performance have been developed. With the aim of designing a prototype expert system for CRM assessment, Seamster and Edens (1993) asked six instructors to sort 60 LOFT concepts by identifying their reasons for grouping them together. Analysis revealed two clusters related to CRM assessment: (a) cognitive (problem identification, task prioritization, and workload management) and (b) interpersonal (teamwork, communication, group climate, and leadership–followership). A third cluster was technical assessment (procedures, technical skills, system knowledge, and maneuvers). Seamster and Edens suggested that this framework has applications for the training of CRM assessors: "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" (p. 126).

In the second phase of this project, Seamster, Edens, McDougall, and Hamman (1994) used observable behaviors associated with crew problems in proficiency checks and first look sessions. When 703 instructor remarks on CRM were categorized using slightly different labels, they showed that the four cognitive categories (situation awareness, workload management, planning, and decision making) made up a substantially greater percentage of crew problems (68%) than the four interpersonal categories (crew coordination, communications, leadership–followership, and group climate). From these results, they argued "that in both scenario development and scenario evaluation, there should not be an evenly distributed emphasis on CRM categories" (p. 3). With the help of eight instructors and captains, they were able to link subsets of observable crew behaviors to scenario event sets for transition or qualification training on the Boeing–737–300. This exercise showed that "when scenario event sets are specified and listed with likely crew behaviors, experienced pilots with some familiarity with the LOE concept can show substantial agreement on the primary observable behaviors to properly assess the related tasks. Therefore, it is likely that making CRM assessments based on observable behaviors will produce reliable assessments" (p. 10), although reliability was not tested directly.

A computer-based method was developed by Dutra, Norman, Malone, McDougall, and Edens (1995) that they called the *CRM assessment expert system*

tool. This was based on instructor remarks about crew problems observed during first look and proficiency check simulator sessions for B–737–300 and B–767 aircraft, for two air carriers. They segmented and coded 1,298 remarks into four assessment categories: (a) cognitive (decision making, situation awareness, workload management, and planning), (b) interpersonal (communications, group climate, crew coordination, and leadership–followership), (c) technical, and (d) other. A total of 579 segments was coded as CRM related (i.e., cognitive or interpersonal) with an associated 931 observable crew behaviors. They then linked these behaviors to 10 event sets they had chosen, and, through this, they reduced the observable behaviors to 90. After instructors had rated the centrality of the behaviors, this list was reduced to 54 elements. For each event they had chosen, they were able to specify five or six key CRM behaviors that the instructor should focus on during that stage of the LOFT flight. They found when event sets were clearly specified, subject matter experts showed agreement on the CRM behaviors that need to be observed to assess the related tasks. Thus, it was concluded that reliable and valid evaluations of CRM can be conducted using clearly defined observable behaviors.

The process of upgrading crew training for U.S. airlines using AQP continues to raise the profile of CRM skills in curriculum development, crew training, and crew performance assessment. Seamster, Prentiss, and Edens (1997) reviewed methods for identifying and specifying the primary CRM skills using standard behavioral skill analysis methods. They suggested that carriers need to improve their methods of analysis to produce a complete skill list with appropriate behaviors that can be rated for CRM assessment. In a related study, Lanzano et al., (1997) examined the results of a comprehensive task analysis for one fleet of a carrier involved in an AQP. This revealed 2,500 unique knowledge and skill entries in the program audit database. Two main categories of skills in such databases are psychomotor and cognitive. Their review showed that only 13 unique CRM elements were categorized as cognitive skill (3%), and they concluded that "carriers are identifying a very limited number of cognitive skills as being related to CRM" (Lanzano et al., 1997, p. 3). The single largest group of CRM elements (48%) was associated with unique knowledge components, demonstrating that the focus in CRM training has been primarily knowledge based rather than skill based. Their recommendations included: "When working with CRM skills, specify and adhere to a common level of detail" and "CRM skills restated as performance objectives should be linked to observable behaviors for training and assessment purposes during LOFT and LOE" (Lanzano et al., 1997, p. 5).

The studies reviewed previously suggest that there is some variability in the core content of CRM behavioral markers systems. Moreover, there are problems with some behavior lists that can make them difficult to use. Seamster, Hamman, and Edens (1995) found that when a marker contained more than one behavior, it was difficult to rate. To use their example: "team concept *and* environment for

open communications established *and/or* maintained" (p. 665). They suggested that only one behavior should be contained in each behavior statement, and a second should only be added if it is absolutely necessary to qualify the first. It is important that the wording of markers is concise and simple and that the verb of the statement refers to a clearly observable behavior, such as *monitor* or *ask*. This means that *made a decision* is not observable, whereas *communicates a decision* is. They emphasized that the designer should always remember that the tool has to be understood and used by instructors–evaluators who have a high workload during the assessment process and, therefore, should make each element as simple as possible.

Scenarios and Event Sets for Behavioral Ratings

Some marker systems have been designed to rate pilots' behaviors in response to predetermined scenario events. Fowlkes, Lane, Salas, Franz, and Oser (1994) produced a team performance measurement approach, targeted acceptable responses to generated events or tasks (TARGETs), for U.S. military cargo helicopter teams. This was based on a set of critical aircrew cooperation behaviors grouped into seven basic skill areas: mission analysis, adaptability–flexibility, leadership, decision making, assertiveness, situational awareness, and communication. Behaviors are linked specifically to stimulus events in a scenario (i.e., these are predefined into a set of acceptable behaviors, task responses; i.e., the TARGETs), and then they are rated as present or absent. In a training and evaluation study of six military aircrews, they concluded that the TARGETs measure had sensitivity (discrimination between crews) and an acceptable degree of rater reliability. However, 30% of events were managed correctly by all six crews and only two raters were used—one of whom had developed the TARGETs.

Seamster, Hamman, and Edens (1995) also advocated the use of event sets within LOFT–LOE, for which specific behavioral markers are written. They argued that identification of observable behaviors can be better focused if a clearly definable unit of action or time is specified and used to delimit the observable crew behaviors. An example is provided of an LOE worksheet showing different event sets and specific CRM behaviors for each set, rated on a 4-point scale ranging from 1 (*unacceptable*), 2 (*minimally acceptable*), 3 (*standard*), to 4 (*above standard*). The scenario event set allows instructors–evaluators to focus on particular CRM categories at given times in the session, depending on the phase of flight and the objectives of the event set. This may reduce instructor workload by allowing attention to be directed on a few key CRM categories rather than having to monitor for all categories continually. These can then be carried forward to assessments to increase interrater reliability. They recommended that the classification used for a marker system should divide CRM behaviors into groupings that facilitate the assessment pro-

cess. Broadly, CRM behaviors divide into two behaviors: interpersonal elements and mental activities. The former are directly observable, whereas for the latter, the instructor–evaluator will often have to make inferences based on interpersonal or technical behaviors to assess mental activities. Therefore, according to Seamster, Hamman, and Edens (1995), instructors or evaluators should not be asked to observe where the crew is engaged in mental actions (e.g., making a decision); rather, they should be asked to observe actions (e.g., specific crew communications) that indicate that a decision has been made. Interpersonal factors can be directly observed through crew communication and coordination of tasks.

Rating CRM Skills

One of the fundamental concerns for ensuring the quality of any system for rating pilots' CRM behavior is the reliability of the raters' judgements. A degree of bias or systematic error can be expected in any performance rating task. Hamman and Holt (1997) identified several reasons for rating errors: personal interpretation, memory errors, scale use, and biases due to motivation. Different judges may be particularly susceptible to certain biases, which raises a second problem: interrater reliability. Brannick and Prince (1991) explained, " … if judges cannot be trained to be interchangeable, then feedback to air crews will depend more upon the particular instructor than on the team's behavior" (p. 1). These measurement errors would also apply in the case of an examiner rating the performance of a pilot's technical skills, and they underline the importance of ensuring the validity and reliability of any system for rating nontechnical skills in which the markers are likely to be rather less specific than instrument readings or control positions.

Thus, a critical factor in the implementation of any rating device is the training of those tasked to use the system (Taggart, 1991). Antersijn and Verhoef (1995) discussed the process of gaining acceptance for a new nontechnical skills rating instrument. The nontechnical skills were subdivided into five main categories: work attitude, information management, leadership, stress management, and cooperation. They stressed that such tools should be practical and visible and the users should be part of the development of the system. When Royal Dutch airlines (KLM) introduced their Feedback and Appraisal System, they prepared their instructors through an advanced instruction course in which the system was presented in practical sessions. During assessments, the definitions and descriptions of the behavioral markers were available to all those taking part—from instructor to flight engineer. Antersijn and Verhoef concluded, from a survey of 118 instructors and 194 pilot–flight engineers, that the Feedback and Appraisal System KLM developed was a success. They also found that some of the elements in the system were used more often in different environments; that is, some elements are more useful during normal flights compared to simulator training.

Williams, Holt, and Boehm-Davis (1997) looked at the interrater reliability of instructors–evaluators (examiners) rating pilot performance. They found that interrater reliability was very low prior to training. However, training during which instructors and examiners discussed consistency and other aspects of reliability led to an improvement of interrater reliability, although Williams et al. do not quantify this. Training was of three types: (a) familiarization with the rating scales, (b) frame of reference training in which instructors were informed about organizational standards to provide them with a common baseline from which to work, and (c) awareness training. The third method was the least effective. Williams et al. had 60 participants rate videotaped LOEs using 3- and 4-point CRM and technical behavior scales. The researchers looked at agreement, systematic differences, congruency, consistency, and sensitivity between raters, and suggested a set of baseline and benchmark measures for rater evaluation. In general, they found all measures of agreement before training were low, but the 4-point rating scale was related to more stable and consistent differences among participants. They extended this study and produced clear guidelines on training and assessing the reliability of instructors' and evaluators' CRM rating skills (Holt et al., in press).

Law and Sherman (1995) noted the sparseness of reliability and validity data for CRM skill assessments. Trained observer ratings show reduced halo bias, increased rater accuracy, and a reduction in errors. This indicates that it is important to teach trainers to use the instruments if they are to be employed consistently and yield results of any value. However, these positive effects do fade with time, although Law and Sherman suggested that if the skills are frequently practiced, this fade may not occur as refresher training can ameliorate increases in errors over time. They suggested an index of agreement through which the reliability of CRM ratings can be evaluated and compared. In a genuine training situation, 24 evaluator trainees viewed a video of a two-person crew in a full-fidelity, simulated line flight. Ten other evaluator trainees viewed a different crew in the same simulation. Trainees used Helmreich et al.'s (1997) LLC4 to provide judgements of 27 elements for four phases of flight and gave two overall ratings. The agreement between raters was high with the two groups showing overall effectiveness ratings that correlated at 0.82 and 1. Technical rating agreement was lower, with the correlations between ratings of 0.78 and 0.94. Thus, raters were in more agreement about crew performance effectiveness than they were in crew technical proficiency. Law and Sherman discussed some limitations of their findings, including the absence of a control group for the experiment, and that although trainees were in high agreement about the overall performance of crews, their ratings of the 27 CRM elements varied. This suggests that raters do not necessarily agree exactly on when a specific CRM behavior should be rated for a given phase of flight (see also Brannick & Prince, 1995).

Seamster, Edens, and Holt (1995) evaluated CRM ratings at a session level, compared to event level, to assess how experienced instructors and check pilots as-

sess crew performance of videotaped LOE sessions. A total of 32 aircrew instructors used four instruments containing CRM and technical elements to rate five scenario event sets. The highest correlations were obtained when instructors rated the entire scenario using a 5-point scale (0.47), as opposed to the lowest correlation from rating of event sets (0.19). They suggested that the higher correlation may indicate a halo effect, making the event set rating more discriminatory. They concluded that the effectiveness of a particular method of assessment may depend on the exact rating scales being used, making it critical to specify both the assessment method and the rating being used when conducting CRM research.

Summary

The CRM literature suggests that there is a general acceptance of CRM training and a growing acknowledgment of the need to assess these nontechnical skills (although this appears to be mainly in the larger carriers, especially those in the United States, operating an AQP). Using a rating system based on behavioral markers to assess CRM skills is a constructive way, as it will allow a greater degree of objectivity in training feedback sessions and evaluation. However, the research to date has only begun to provide answers to the core issues.

Unit of assessment. The unit of assessment in most research studies, particularly in the work of Helmreich et al. (1997) and Helmreich et al. (1996), is the flight deck crew rather than an individual pilot in a crew setting. It is obviously of interest to airlines to see how their crews are performing, but for individual licensing, individual assessments will be required, as they are currently for technical skills.

Identification of CRM skills and associated behavioral markers. The assessment of CRM has to be founded on an established set of CRM skills. Core CRM concepts are often subdivided into two categories: cognitive skills (e.g., decision making, situation awareness, and workload management) and social skills (e.g., leadership and team work). These concepts appear to be used fairly consistently, but labeling differs across research studies, airlines, and fleets. Devising a behavioral markers system for those skills presents a number of challenges. The sheer number of nontechnical behaviors that could potentially be assessed must be reduced to match a set of critical CRM skills. These can be condensed by being more specific about the event or the scenario, as each airline and fleet will choose different scenarios. Thus, the behaviors being recorded in the industry could cover a very wide range—even allowing for synonyms and equivalent terms.

Assessment method. Although there is an extensive psychological literature on the measurement of individual and team performance (Aiken, 1996; Brannick, Salas, & Prince, 1997), there have been very few studies that systematically compare rating scales for CRM performance, particularly for the rating of individual pilots.

Reliability. Studies of rater reliability highlight the importance of rater training to achieve calibration against established standards and minimize bias effects (Williams et al., 1997). Differences between ratings based on entire sessions, compared to ratings for given stimulus events or event sets, have been found. In the case of the latter, specific behaviors can be specified and research teams have argued for the merits of this approach (Fowlkes et al., 1994), although in practical terms, scenario development costs would have to be taken into account. Helmreich et al. (1996) suggested that it is advantageous to rate crew behaviors in relation to phase of flight.

Rater training. Training in facilitation, debriefing, and the use of CRM rating scales is critical to ensure fair and accurate assessments of nontechnical skills. A number of research projects have been undertaken and guidance released; for instance, the use of the LLC (Helmreich, Wilhelm, Kello, et al. 1990) or the conduct of LOFT and LOE sessions (Dismukes, Jobe, & McDonnell, 1997; McDonnell, Jobe, & Dismukes, 1997). Raters conducting licence evaluations will, themselves, need to be assessed and qualified (see Royal Aeronautical Society, 1998, for competence standards for CRM trainers and evaluators).

There are other questions relating to assessment methods that have not been examined in the research studies; for example, comparison between line- and simulator-based evaluations and the extent to which technical and nontechnical skills assessment can be integrated, although work is beginning in this area (Lanzano et al., 1997). There are other psychometric issues that should be considered in the design of the rating scales; for example, the number of scale points, whether these have behavioral anchors, and how the scale should be scored and standardized. These issues have not been tested systematically for CRM behavior scales. Some of these questions were addressed in the two surveys reported next.

METHOD

The aim of the first survey was to assess current views on behavioral marker systems in companies in the UK with or without experience of using a marker system. The aim of the second survey was to gather evidence on the use of marker systems

from a small sample of international (non-UK) airlines with experience in their use. On the basis of the literature review, a semistructured interview schedule (for the UK) and a questionnaire (for international) was designed that asked about CRM and, in particular, the use of behavioral markers in training and assessment (see the Appendix for a list of the questionnaire items and Flin & Martin, 1998, for more details).

UK Sample

The sample comprised a cross section of 11 commercial operators, including large multifleet operations and smaller operators—both fixed wing and rotor wing (Air Atlantique, Bond Helicopters, Bristow Helicopters, Britannia, British Airways, British International Helicopters, British World Airways, Magec, Monarch, Shell Aircraft, and Virgin). The Royal Air Force (RAF) was also included as they had developed a generic CRM program that was compatible to those used by the airlines. Participants were involved in the development and delivery of CRM in their organization. Interviews were by telephone (6) or face to face (6), the former taking about 45 min and the latter just over 1 hr.

International Survey

Thirteen airlines known to use behavioral markers were contacted by telephone, letter, or at the 1997 Aviation Psychology Symposium in Ohio and sent the questionnaire. Nine returned it (69% response rate). These were Aer Lingus, Air Canada, Ansett, Atlantic Coast, Braathens SAFE, Cathay Pacific, Delta, Northwest, and SAS. Some information was also provided by American Airlines and Continental. The European Evaluation of Nontechnical Skills of Multipilot Aircrew in Relation to the JAR–FCL Requirements (Van Avermaete & Kruijsen, 1998) research group provided information for KLM, Lufthansa, and Air France. This gave a final sample of 14 airlines, although some of these are incomplete responses; hence, the changing sample numbers in the discussion that follows.

RESULTS

A broad content analysis was undertaken of the interviews and questionnaire responses. These are presented separately for the UK and international samples as this contrasts a sample of mixed-size operators in one country with a sample of experienced users from larger carriers in the international community. The following discussion is subdivided into five themes: details of behavioral markers scales, training, use of the marker systems, pilots' opinions, and future developments (with a simple summary provided in Table 1).

TABLE 1
Summary of Key Questions and Responses From Airlines

Key Issues	*UK*	*International*
BMS available?	5/12	14/14
BMS used for assessment?	0/5	12/14
Can pilots fail check based on BMS ratings?	Not currently	6/12
Need for BMS training for instructors and examiners?	Yes	Yes
Pilot views on BMS?	Not asked	Generally positive
Will BMS be used in the future?	Yes	Yes

Note. BMS = behavioral marker system.

Details of the Behavioral Markers Scale (UK)

Of the 12 UK organizations surveyed, only 50% (5 airlines, plus the RAF) had any kind of behavioral marker scale (one of which had just abandoned its use). None of these were used for formal CRM assessment. Instead, they were designed for CRM training and to structure LOFT feedback sessions. In one system they were described as "an aide memoire of a range of behaviors." These 5 airlines were aware that they might be required to evaluate CRM skills more formally at some future date. Those organizations who had CRM skills scales had consulted a number of examples before beginning to develop their own. For some, this development was minimal—adopting an existing list and changing the odd word or phrase. Others had carried out in-house research to produce their own behavioral markers framework. All participants developing behavioral markers scales had reviewed Helmreich, Wilhelm, Kello, et al.'s (1990) LLC form.

Copies of five of the six marker frameworks were obtained. Figures in brackets represent categories or elements or behaviors. They were called LOFT–human factors (8/37); team skill–pointer (8/58); LOFT–HF/CRM performance markers (6/36), performance markers (8/32), and CRM behavioral markers (3/16). The CRM skill categories typically included situational awareness, decision making, communication, leadership or followership, crew relations, workload management, or some variation of these. However, the five systems were structured differently and contained different categories, elements, and behaviors. Only one had a rating scale, which listed three behaviors for each element: one unsatisfactory, one standard, and one above standard. In two systems, the markers were for an individual pilot's behavior. In the other three, they were worded in terms of crew skills but could also be used to give feedback to a particular pilot. The general opinion of the LLC (Helmreich et al., 1997; Helmreich, Wilhelm, Kello, et al., 1990) was that it was an excellent research tool, but it did not entirely meet the needs of UK practitioners.

Details of the Behavioral Marker Scale (International)

All 14 airlines had some type of CRM behavioral marker system used in training and, in some cases, for assessment. Sample names were crew effectiveness pointers, crew effectiveness markers, observable crew behaviors, essential skill list, check assessment system, and crew performance indicators. These marker systems had been developed from 1979 (typically from 1990 onward), using in-house expertise from their pilots, trainers, and psychologists. When asked for the research basis of their marker system, most referred to the LLC, and some companies had adopted it with minimal alteration. Other sources were also mentioned—several airlines had collected in-house data from pilot performance, pilot opinions, and incident data. Some airlines had studied marker systems from other airlines or military air forces before designing their own version. This suggests that the majority of marker systems had been based on the LLC, although airlines tend to adapt this for their own purposes. Two airlines mentioned the importance of developing a generic system. Reasons for this were twofold. First, where there are international regulations (e.g., in Europe), the behavioral markers systems in use should assess pilots on an equivalent basis. Second, confusion can arise when different behavioral markers systems are in circulation—raising questions in the pilot community as to the respective validity of these systems.

The 13 behavioral markers lists provided varied widely in design and included from 7 to 30 behaviors. Nine of the systems considered technical skills with nontechnical skills. The main CRM categories were broadly similar and were sometimes subdivided into social (interpersonal) and cognitive (mental) skills. The principal categories typically included leadership, communication, team coordination, decision making, situation awareness, and workload management—although the labels for these categories often differed. Some lists included additional categories such as stress, automation management, or crew self-evaluation. No two systems had the same content or format. All included a rating scale, ranging from 2-point to 6-point scales, with a 5-point scale being the most common. Typical labels would be: unsatisfactory, satisfactory, average or standard, above average, and excellent (or equivalent terms). In two companies, a 5-point numerical scale was used. Some airlines used other labels such as +/– (which indicate presence or absence of a behavior or pass or fail). The reasons given for choosing the scale in use included that it was a proven design (e.g., version of LLC) or that the format provided the information they required.

Behavioral markers systems had been introduced within airlines in the main through courses (either CRM training, LOFT, or other courses), written reference manuals, or a mixture of methods. Nine airlines had modified their original system as a result of experience, ranging from minor adjustments like decoding psychological terminology, to complete overhaul and reworking of all the markers in line with company-wide opinions. One airline had to improve its training for instructors and

evaluators, and they incorporated the markers onto the instructor's grade sheet. Only two airlines made changes because the system was rejected by the pilots. Six airlines reported that they still had problems with their current system (e.g., over complex or introduced too quickly). Of these, four airlines have modified the system from the original, and two are planning further modifications.

Training for Rating CRM Behaviors (UK)

Some participants felt that all pilots should receive training explaining the underlying CRM skills concepts of the behavioral markers scales under development. Others felt that instructors and evaluators needed to receive this training and that pilots only required a comprehensive manual and an overview to be presented in CRM initial and refresher courses. Some airlines are developing facilitation training for their instructors and examiners, whereas others are relying on consultants to provide this training for them. Participants commented that much of the resistance they have encountered from their trainers, with respect to using behavioral markers, is based on their lack of confidence at being able to use the system. Facilitator training should include how to generate and conduct debriefs—getting pilots to talk and to share their experience. Airlines who are developing behavioral markers have all planned a phase of user verification, in which instructors and examiners will be asked to test, run, and critique the behavioral markers scale by using it during training such as LOFT.

Training for Rating CRM Behaviors (International)

Ten out of the 13 airlines answering this question provided a course to train their instructors and examiners to use the behavioral markers system. For 50% of the sample, this was a dedicated course, whereas for the others it was part of an instructor's and examiner's training course covering a range of skills and techniques, such as facilitation and feedback. Courses varied from 1 to 3 days and tended to be workshop based with many practical exercises and video segments that would be used for practice rating. Several respondents were aware of potential difficulties in achieving standardized (calibrated) assessment and emphasized the need for instructors and examiners to be properly trained in the use of behavioral markers to assess CRM skills, particularly if jeopardy assessments were involved.

Use of the Behavioral Markers System (UK)

The general concern of participants was that the system should be fair, especially if a pilot's licence is on the line. This raises a raft of issues about how assessment should take place and opportunities for pilots to retrain where necessary. However, most participants felt that assessment was a useful progression of the CRM training. Many

participants felt that CRM skills had been assessed for years under the guise of airmanship. If a pilot lacks these skills then he or she could fail his or her licence revalidation test or their operator proficiency check under the present assessment system. The need for observations to be concrete and repeatable is emphasized. Some participants felt that outcomes of assessments should not have *fail* as an option but should, instead, recommend further training. However, this raises questions of how many times a pilot should be recommended for further training before the organization concludes that the pilot does not have the required skills to hold a licence. Participants felt that if CRM becomes a jeopardy assessment then providing retraining and opportunities to resolve unsatisfactory performance will become central. These issues also pertain to technical skills assessment for licensing.

Some participants felt that nontechnical behavior on a flight deck cannot be prescribed—there are so many individual differences that safe behavior can be achieved in hundreds of different ways. Their view was that a behavioral markers scale is an indication of desirable behaviors or a guide rather than a prescription. Apart from the reliability or validity issues discussed previously, another concern was whether one scale could be used to assess pilots who had been trained on different CRM courses.

Most respondents believed that the only way one can judge whether pilots have understood enough of what has been taught on CRM courses to put it into practice is to observe their performance, either on the line or during a realistic simulation. Another issue raised was whether it is appropriate to assess an individual pilot's nontechnical skills rather than the skills of the flight deck crew who are working together. It is likely that this issue will need to be addressed by a systematic comparison of individual versus crew CRM ratings at some future date.

Some airlines have plans for using the data they collect from assessments to improve their CRM programs. By reviewing the performance of pilots during assessment, they hope to identify generally weak areas and be able to tailor their refresher programs to concentrate on these skills. Plans such as this will close the loop on CRM training and tie the programs in with the areas of greatest need. This should lead to a continual improvement of CRM skills.

Use of the Behavioral Marker System (International)

Of 12 airlines, 9 reported introducing their pilots to the behavioral markers system through training courses, whereas 3 use written information to do so. In 11 airlines, their behavioral markers systems were used for training (from ab-initio training to recurrent courses), as well as assessment. Several U.S. airlines had fleets in the process of transitioning to the AQP, which requires assessment of CRM skills. Of the total sample of 14 airlines, 12 said that they used behavioral markers for some kind of assessment, 50% of whom (6) reported that the pilots could fail a check based on their CRM skills. Of the 6 airlines who said they

would not fail on the basis of CRM skills, 2 reported that they would be moving to this position. Two other airlines stated that they would not fail a check on CRM skills because the airline retrains the pilot in question until they do pass the assessment. Of those 6 airlines who may fail a pilot on the basis of their assessment, all reassess the pilot at a later date. In the interim, all offered further training, although this ranges from repeating the original course to going through a different course or having training specifically designed in conjunction with the fleet training captain. One airline includes an interview with the fleet manager as part of the review of an unsuccessful assessment. Some airlines have a concern that understanding of the behavioral markers system and its use is not great enough for it to be used in assessment, that pilots do not yet fully understand what the aim of the system is, and have not yet received sufficient training to be assessed using the markers. A more negative outlook on this problem is that assessment can be introduced without sufficient CRM training for pilots and, therefore, does not follow an increase in awareness or use of CRM skills.

A number of airlines were concerned about retraining; for instance, how to design further training and whether this should be different from the original training. This is the issue of whether it is necessary or practicable to develop nontechnical skills retraining more suited to the learning style of the individual who has to retake the course. Other airlines were concerned that although behavioral markers are indicators that the individual is having problems, they are not diagnostic and, therefore, further work has to be undertaken to specify exactly where the pilot is having difficulties. Only one airline reported that they also use the behavioral markers system to assess the CRM courses they run, with the aim of improving the course through this application of the markers.

Following the use of the behavioral markers system, all airlines debriefed the pilots about the CRM ratings that they had been given and gave them the opportunity to discuss the points made. They may be asked to read and sign the assessment sheet, and two airlines said that they gave pilots copies of this record. In one airline, the pilots were asked to complete a crew self-evaluation using the same CRM rating scale. Three airlines did not retain the rating data. The other airlines filed the behavioral markers report, either in a general database or on the pilot's personal file. In general, ratings were de-identified before entry onto a fleet or company database. Following the feedback session, three airlines offered their pilots the opportunity to take further training or practice.

Pilots' Views on the Use of Behavioral Markers (International)

When asked what feedback had been received from pilots who had taken part in courses or assessments in which behavioral markers were used, the responses were

generally favorable (this question was not posed to the organizations in the UK, in which there was less experience of these systems). Eight out of nine airlines reported that their pilots received the behavioral markers system positively: Three reported high acceptance, five reported acceptance, and one said rejection. The following comments are indicative:

> Only positive feedback—I think a lot of our pilots up till now have felt that CRM is difficult to get a grip on, a 'buzz word.' The behavioral markers system shows them that CRM is about concrete, practical skills, something they can do and something that makes sense.
>
> CRM LOFTS were nonjeopardy so pilots appreciated the opportunity to practice CRM skills. The future assessments will be just the same as other skills assessments.
>
> Pilots enjoy and express that they learn more in our LOE–LOFT training. In general, they have a positive approach to evaluating crew effectiveness.

One respondent reported considerable resistance to the introduction of behavioral markers, particularly among the instructors:

> This is mainly because of their lack of confidence in the [name of marker system]. Their feelings seem to have some merit as considerable variation in assessments was found. We also have evidence that building an evaluation consensus amongst instructors after a day of watching a few videotapes for training purposes is by no means the same as achieving an enduring and objective set of performance criteria.

When asked of problems perceived by pilots with regard to the behavioral markers system, one respondent mentioned that any problems they did encounter were due to the system being used in a way that emphasized identification and correction of mistakes rather than an acceptance that errors are inevitable and that the critical CRM behaviors are due to error trapping and mitigation. A different airline revealed difficulties balancing emphasis on assessment against flight safety and error management. It seems that using behavioral markers in both jeopardy assessment and training in this airline caused a degree of conflict.

Future Developments

The response to this question was almost uniform (the UK and international), with airlines reporting that they intend to use their behavioral markers systems more in the future. Included in this broadening of perspective are aims to integrate technical with nontechnical skills, make the system simpler, and improve instructors' and ex-

aminers' skills. Furthermore, some airlines envisaged the need to extend the use of behavioral markers systems to other aviation domains such as maintenance. The following comments were offered:

> I hope, in the future, that knowledge about behavioral markers will be a natural part of every instructor's competence, and I hope they will help them to have better use of video debrief as a tool to teach pilots CRM.
>
> We would like to expand it to more specific assessment (e.g., to be able to identify more specific areas the candidate would be able to improve on).
>
> Measurement of behaviors are critical to addressing human factors issues which feature in all accidents.
>
> Continued use as more aircraft fleets train using AQP.
>
> I think it is important to develop behavioral markers that different airlines can agree upon, and feel that NASA–University of Texas has done a very good job with their behavioral markers.
>
> We have extended the behavioral markers system to other company domains—flight attendants, meterologists, and mechanics receive human factors training.
>
> We are still developing CRM-wise, but we feel that our crews have sufficient understanding and practice that CRM skills can be assessed like all others.
>
> There are great differences between behavioral markers systems in use. There seems to be confusion between the use of criteria to help instructors structure their briefings and the notion that cockpit behavior can be reduced to a checklist of observable behaviors. Behavioral markers may create a new assessment reality without positively affecting pilot performance or impacting upon cockpit management.

A summary of some of the key questions is given in Table 1.

DISCUSSION

Participants from the UK operators raised a number of issues concerning the introduction of behavioral markers that may reflect their more limited level of experience of such systems. A central and important point was that they should not be introduced in isolation. For markers to be understood and useful, they should be part of an education program that explains their role in CRM training and assessment. Although CRM awareness training is mandated in the UK, this does not mean that all airlines have chosen to develop their programs any further than this. The airlines with more advanced CRM programs have already introduced marker systems and were using these in LOFT feedback sessions.

Moving to CRM skills assessment involves a number of stages: scale development, scale trialling, training of assessors, and system evaluation. An organization moving through all these stages will need to ensure pilots' familiarity with the concepts and allow the time necessary for cultural adjustment. The terminology of a behavioral rating scale and the way it is set out will influence the acceptance of it by pilots, instructors, and examiners as a valid component of the assessment process. The needs of the users must be taken into account in the design of any scale that should be viewed as a tool and come complete with a user manual.

Some CRM instructors–examiners who will be using behavioral markers systems appear to lack confidence in their ability to use or develop these tools. In many cases, this stems from an unfamiliarity with both the style of behavioral markers presentation and their practical application. Another view expressed was the lack of guidance provided by official bodies and, in some cases, a lack of support from management. The survey results showed that in the sample from the UK, the development of behavioral markers for CRM skills was not widespread (5 out of 12 were using marker systems). However, only large carriers were sampled for the international group; thus, the two samples are not matched. The use of marker systems may be found to be variable in other countries when airlines with smaller fleets and domestic carriers are sampled.

The international sample showed a range of development and use of behavioral markers systems. Although all these airlines (14) reported having a behavioral markers system of some sort that is used in training, fewer (12) used a rating scale for assessment of pilots' nontechnical skills, and fewer still (6) used this system as part of a formal check. Thus, less than half (43%) of the airlines utilized a behavioral markers system through all stages of training and assessment, suggesting that these systems are still in a development phase. It should again be emphasized that these responses were from a restricted sample of larger airlines, and smaller operators were underrepresented.

Most participants seem to be aiming for a behavioral markers scale that combines the positive properties of a research tool like the LLC and a notecard reminder tool like a checklist. They emphasized the need for a system that was simple, easy to understand, and use. It was felt that successful use of the behavioral markers would rest on the users' understanding of the underlying concepts and, therefore, proper rater and facilitation training was essential. The importance of keeping the instructors and examiners informed and involved with the development of behavioral markers was also emphasized. These pilots hold senior positions in their companies, and their views are influential in the pilot community. As the people who will be asked to use the systems, both in practice and potentially during assessment, it is vital that they not only know how to use them but that they support their introduction. In essence, the instructors' and examiners' acceptance of behavioral markers is crucial to pilots' acceptance of any CRM evaluation system.

EPILOGUE

New European Developments

In 1997, the European project, NOTECHS, was initiated by the Joint Aviation Authorities Research Committee Human Factors Project Advisory Group to provide background information for the JAR–FCL in relation to the evaluation of a pilot's nontechnical skills. Nontechnical skills were defined as pilots' attitudes and behaviors in the cockpit not directly related to aircraft control, system management, and standard operating procedures. The goal of the project was to develop a methodology for assessing pilots' nontechnical skills during flight and simulator checks. The scope of the project related principally to JAR–FCL (Part 1, Subpart F, Paragraph 240), as well as relevant sections of JAR–OPS (JAA, 1996, 1997).

The NOTECHS project was undertaken by a consortium of psychologists from four partner teams: Netherlands (NLR), Germany (DLR), France (IMASSA), and the UK (University of Aberdeen), as well as pilots from KLM. The project objectives were to review the use of nontechnical skills marker systems and to either provide a preliminary endorsement of one particular system or to develop a draft nontechnical marker system based on existing systems and previous research. On completion of the review of existing systems and previous research, the NOTECHS group produced a new draft standard for the assessment of individual pilot's nontechnical skills. The draft nontechnical skills standard (behavioral marker system; see Flin, Goeters, Hormann, & Martin, 1998; Van Avermaete & Kruijsen, 1998) is being evaluated on a European cross-cultural basis in a new phase of research sponsored by the Transport Directorate of the European Commission (EC DGVII). This project, which began in 1998, involved the NOTECHS consortium plus additional research organizations (SOFREAVIA, France and DERA, UK) and three companies: British Airways, Airbus, and Alitalia. The project is called Joint Aviation Requirements Translation and Elaboration, and the group will report on an empirical study of the use of the NOTECHS marker system in a number of European countries by 2001.

ACKNOWLEDGMENTS

This study was commissioned by the Safety Regulation Group of the British Civil Aviation Authority.

We thank the survey participants who took the time and effort to complete our questionnaire and to supply copies of their marker schemes.

The views expressed are those of the authors and should not be taken to represent the position or policy of the funding body.

REFERENCES

Aiken, L. (1996). *Rating scales and checklists.* New York: Wiley.

Antersijn, P., & Verhoef, M. (1995). Assessment of non-technical skills: Is it possible? In N. McDonald, N. Johnston, & R. Fuller (Eds.), *Applications of psychology to the aviation system* (pp. 243–250). Aldershot, England: Averbury.

Birnbach, R., & Longridge, T. (1993). The regulatory perspective. In E. Wiener, B. Kanki, & R. Helmreich (Eds.), *Cockpit resource management* (pp. 263–281). San Diego, CA: Academic.

Boehm-Davis, D., Holt, R., & Seamster, T. (in press). Airline resource management programs. In E. Salas, C. Bowers, & E. Edens (Eds.), *Improving teamwork in organizations: Applications of resource management training.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Brannick, M., & Prince, C. (1995). Reliability of measures of aircrew skills across events and scenarios. In R. Jensen (Ed.), *Proceedings of the 8th Symposium of Aviation Psychology* (pp. 603–606). Columbus: Ohio State University.

Brannick, M., Salas, E., & Prince, C. (Eds.). (1997). *Team performance, assessment and measurement.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Civil Aviation Authority. (1998). *Flight crew CRM training standards* (Aeronautical Information Circular 114/1998: Pink 178). Gatwick, England: Author.

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

Diehl, A. (1991, November). *Does cockpit management training reduce aircrew error?* Paper presented at the 22nd International Seminar for Air Safety Investigators, Canberra, Australia.

Dismukes, K., Jobe, K., & McDonnell, L. (1997). *LOFT debriefings: An analysis of instructor techniques and crew participation* (NASA Tech. Mem. 110442). Ames, IA: NASA.

Dutra, L., Norman, D., Malone, T., McDougall, W., & Edens, E. (1995). Crew resource management/assessment: Identification of key observable behaviours. In R. Jensen (Ed.), *Proceedings of the 8th Symposium of Aviation Psychology* (pp. 562–567). Columbus: Ohio State University.

Flin, R., Goeters, K.-M., Hormann, J., & Martin, L. (1998, September). *A generic structure of non-technical skills.* Paper presented at the European Association of Aviation Psychology conference, Vienna.

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

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.

Gaba, D., Howard, S., Flanaghan, B., Smith, B., Fish, K., & Botney, R. (1998). Assessment of clinical performance during simulated crises using both technical and behavioural ratings. *Anesthesiology, 89,* 8–18.

Goeters, K.-M. (Ed.). (1998). *Aviation psychology: A science and a profession.* Aldershot, England: Ashgate.

Gregorich, S., & Wilhelm, J. (1993). Crew resource management training assessment. In E. Wiener, B. Kanki, & R. Helmreich (Eds.), *Cockpit resource management* (pp. 173–198). San Diego, CA: Academic.

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

Helmreich, R. (1996, October). *The evolution of crew resource management.* Paper presented at the IATA Human Factors Seminar, Warsaw, Poland.

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

Helmreich, R., Hines, W., & Wilhelm, J. (1996, September). *Issues in crew resource management and automation use: Data from line audits.* Paper presented at the 6th CRM Industry Workshop, Charlotte, NC.

Helmreich, R., Merritt, A., & Wilhelm, J. (1999). The evolution of crew resource management training in commercial aviation. *International Journal of Aviation Psychology, 9,* 19–32.

Helmreich, R., Wilhelm, J., Gregorich, S., & Chidester, T. (1990). Preliminary results from the evaluation of cockpit resource management training: Performance ratings of flight crews. *Aviation, Space and Environmental Medicine, 61,* 576–579.

Helmreich, R., Wilhelm, J., Kello, J., Taggart, W., & Butler, R. (1990). *Reinforcing and evaluating crew resource management: Evaluator/LOS instructor reference manual* (Tech. Manual 90–2). Austin: NASA/University of Texas.

Holt, R., Boehm-Davis, D., & Beaubien, J. (in press). Evaluating resource management training. In E. Salas, C. Bowers, & E. Edens (Eds.), *Improving teamwork in organizations: Applications of resource management training.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Joint Aviation Authorities. (1996). *JAR–OPS. Flight operations.* Hoofddorp, The Netherlands: Author.

Joint Aviation Authorities. (1997). *JAR–FCL: Part 1. (Aeroplane) flight crew licensing requirements.* Hoofddorp, The Netherlands: Author.

Lanzano, J., Seamster, T., & Edens, E. (1997). The importance of CRM skills in an AQP. In R. Jensen (Ed.), *Proceedings of the 9th Symposium of Aviation Psychology* (pp. 574–579). Columbus: Ohio State University.

Law, J., & Sherman, P. (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 Symposium of Aviation Psychology* (pp. 608–612). Columbus: Ohio State University.

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

Maschke, P., Goeters, K., Hormann, H., & Schiewe, A. (1995). The development of the DLR/ Lufthansa crew resource management training. In N. Johnston, R. Fuller, & N. McDonald (Eds.), *Aviation psychology: Training and selection* (pp. 23–31). Aldershot, England: Avebury.

McDonnell, L., Jobe, K., & Dismukes, K. (1997). *Facilitating LOS debriefings: A training manual* (NASA Tech. Memorandum 112192). Ames, IA: NASA.

Naef, W. (1995). Practical application of CRM concepts: Swissair's human aspects development program. In R. Jensen (Ed.), *Proceedings of the 8th Symposium of Aviation Psychology* (pp. 597–602). Columbus: Ohio State University.

Royal Aeronautical Society. (1998). *Guide to performance standards for instructors of crew resource management (CRM) training in commercial aviation.* London: Author.

Salas, E., Fowlkes, J., Stout, R., Milanovich, D., & Prince, C. (1999). Does CRM training improve teamwork skills in the cockpit? Two evaluation studies. *Human Factors, 41,* 326–343.

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.

Seamster, T., Edens, E., & Holt, R. (1995). Scenario event sets and the reliability of CRM assessment. In R. Jensen (Ed.), *Proceedings of the 8th Symposium of Aviation Psychology* (pp. 613–618). Columbus: Ohio State University.

Seamster, T., Edens, E., McDougall, W., & Hamman, W. (1994). *Observable crew behaviors in the development and assessment of line operational evaluations* (Rep. No. DTFA01–93–C–00055). Washington, DC: Federal Aviation Administration.

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

Seamster, T., Prentiss, F., & Edens, E. (1997). Methods for the analysis of CRM skills. In R. Jensen (Ed.), *Proceedings of the 9th Symposium of Aviation Psychology* (pp. 500–504). Columbus: Ohio State University.

Taggart, W. (1991). Advanced CRM training for instructors and evaluators. In R. Jensen (Ed.), *Proceedings of the 6th Symposium of Aviation Psychology.* Columbus: Ohio State University.

Taggart, W. (1995). The NASA/University of Texas/Federal Aviation Administration Line/LOS Checklist: Assessing system safety and crew performance. In R. Jensen (Ed.), *Proceedings of the 8th Symposium of Aviation Psychology* (pp. 688–693). Columbus: Ohio State University.

Van Avermaete, J., & Kruijsen, E. (1998). *The evaluation of non-technical skills of multi-pilot aircrew in relation to the JAR–FCL requirements* (NOTECHS Group Project Rep. No. NLRCR98443). Amsterdam: National Aerospace Laboratory.

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

Manuscript first received May 1999

APPENDIX
International Survey Questionnaire

This questionnaire asks about the behavioural markers (crew performance indicators) system, for non-technical/CRM skills, that is in place in YOUR AIRLINE. Please could you complete all items as fully as possible? All information you include will be de-identified before it is included in any of our reports or publications.

General information about the marker system YOUR AIRLINE uses

Name of behavioural markers (BM) system?
Who developed the BM system?
If YOUR AIRLINE developed the system itself—what year was development begun?
What was the research basis (e.g., Helmreich's NASA/University of Texas work, LLC4)?
When was the system first used in YOUR AIRLINE?
How was the initial concept of the BM system introduced?
What changes have you made to your BM system since it was introduced?

Details of behavioral marker system

List the markers in your BM system, or enclose a copy of your BM sheet.

Does it have a rating scale?
How many points are on the scale? 2, 3, 4, 5, 6, or other?
What are the labels for the points? e.g., poor, standard, excellent?
Why was this type of scale chosen?
Does the system consider technical skills with non technical skills?

Training Instructors or Examiners to use the BM system

What training do instructors or examiners receive to use the BM system?
How long are the courses for?
Do instructors or examiners follow a protocol when using the BM system?

Status of BM system

What information is given to pilots to explain how the BM system will be used?
What is the BM system used for within YOUR AIRLINE?
If the BM system is used during pilot training, please give brief details below.
If your BM system is used during checks, could a pilot fail this check based on their CRM skills?
Are pilots reassessed if they fail the check?
When are they reassessed?
What feedback do pilots receive about their ratings on the BM scale?
Following feedback, do pilots have an option for further training or practice?
How is the output or result from using the BM system filed? e.g., does a report go onto a pilot's personal file?
Are there any further consequences not dealt with above?

Pilots' views

What feedback have you received from pilots who have taken part in courses or assessments where YOUR AIRLINE's BM system is used?
What is the level of acceptance of the BM system by pilots? (Please tick one)
Are you aware of any problems perceived by your pilots with YOUR AIRLINE's BM system? If yes, what are they?

Future developments

How do you think BM will develop and be used in YOUR AIRLINE in the future?

Any other comment

Add any further comments you would like to make on behalf of YOUR AIRLINE.