Development of Method for CRM Skills Asessment

Hiroka Tsuda

Tomoko Iijima

Fumio Noda

Kohei Funabiki

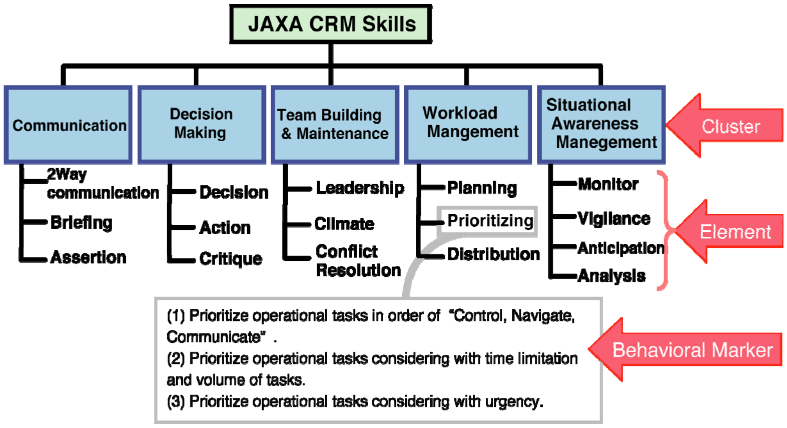
DEVELOPMENT OF METHOD FOR CRM SKILLS ASESSMENT

Hiroka Tsuda, Tomoko Iijima, Fumio Noda and Kohei Funabiki
Japan Aerospace Exploration Agency,
Tokyo, Japan

Crew Resource Management (CRM) is currently considered as one of the most effective methods for avoiding human errors or minimizing their effects. In training, measurement of the level of flight crews' CRM Skills is necessary in order to evaluate objectively which Skills have been adequately learned and which are lacking. The Japan Aerospace Exploration Agency (JAXA) has developed CRM Skills Behavioral Markers and CRM Skills Measurement Methods that can identify a crew's level of CRM Skills by which human errors and threats are managed. A series of simulated-LOFT (line oriented flight simulation training) were conducted to examine the applicability of the method.

While improvements in aircraft systems technology have dramatically reduced aircraft accident rates over the past few decades, at present accident rates have flattened out and so different approaches are required to further reduce accidents in the future. Human factors are now a primary causal factor of fatal accidents, and so addressing these should yield further reductions in the accident rate. After Helmric revealed that most human factors-related incidents are caused by inappropriate crew coordination, importance began to be placed on flight crew CRM training, and the first CRM training programs were started by airlines in United States in the 1980s. In Japan, flight crew CRM training was mandated by the Japan Civil Aviation Bureau (JCAB, 1998), and Japan Airlines began CRM training in 1986.

It is considered that concrete behavioral indicators are necessary for effective CRM training, and so from 1999 to 2002 JAXA has been developing CRM Skills Behavioral Markers with the support of airlines (Japan Air System, All Nippon Airways and Japan Airlines) that take into account the particular behavioral and psychological characteristics of Japanese crew members, which would be suitable for the Japanese flight crews operating in a domestic environment. Here, "CRM Skills" is defined to be the ability to carry out CRM. Fig. 1 shows the CRM Skills proposed by JAXA. These are classified into five clusters with three or four skills elements in each. Each skills element has two or more CRM Skills Behavioral Markers.



(1) Prioritize operational tasks in order of "Control, Navigate, Communicate" .
(2) Prioritize operational tasks considering with time limitation and volume of tasks.
(3) Prioritize operational tasks considering with urgency.

Figure 1. *JAXA proposed CRM Skills.*

Verifying the effect of CRM training is as important as conducting the training. To assess the effect of CRM training and to provide feedback to a training program to reduce possible inadequacies, we have developed a CRM Skills Measurement Method that can identify the extent to which CRM skills have been learned (Iijima *et al.*, 2003). Our proposed method utilizes a subjective rating technique based on the JAXA CRM Skills Behavioral Markers. This paper describes the development of the CRM Skills Measurement Method and the evaluation of its applicability.

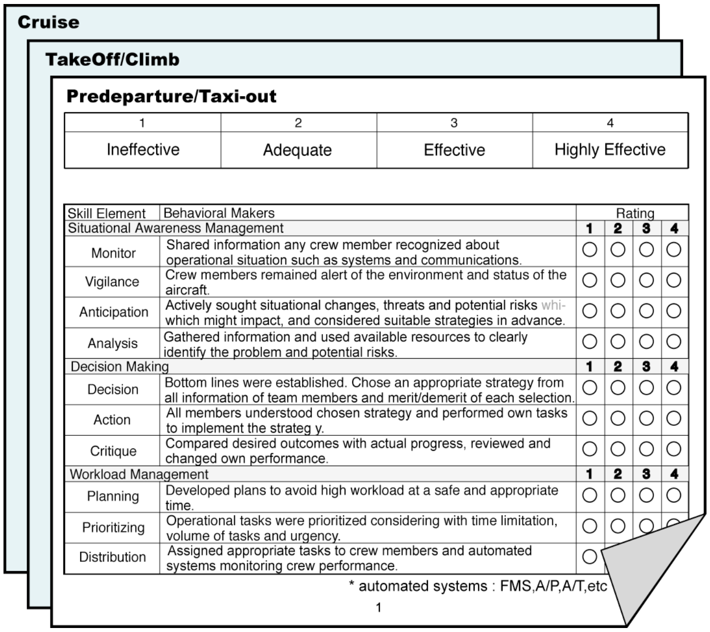
Design of CRM Skills Measurement Method

Development of the CRM Skills Measurement Method consisted of three phases: a preliminary study, design of a prototype rating sheet, and evaluation of the method. The Method includes CRM Skills, a Rating sheet, an Observation sheet, LOFT scenarios, and assessment of inter-rater-reliability.

CRM Skills Rating Sheet and Observation Sheet

To design the CRM Skills Measurement Method, a survey was made of airlines that incorporate CRM Skills into LOFT (Line Oriented Flight Training) and/or LOE (Line Operational Evaluation) (JAL, 2003), and based on the results, an initial version of the Measurement Method, prototype No. 1, was developed for trial purposes. The method used an initial prototype CRM Skills Rating Sheet on which "Raters" (training personnel who evaluate crews' CRM Skills) record scores for the CRM Skills they observe, at that stage on a three-point scale. Raters evaluate only actions that can be clearly observed based on the CRM Skills Behavioral Markers proposed by JAXA. After initial prototyping, the Measurement Method was refined by applying it to sample recorded LOFT sessions, resulting in prototype No. 2.

To improve the second prototype, a preliminary rating experiment was conducted by applying it to a video of a sample LOFT session (the "LOFT Videotapes" below), with two Raters. Feedback from the Raters was then used to produce a third version of the Rating Sheet, part of which is shown in Fig._2.



Cruise

TakeOff/Climb

Predeparture/Taxi-out

1	2	3	4
Ineffective	Adequate	Effective	Highly Effective

Skill Element	Behavioral Makers	Rating			
Situational Awareness Management		**1**	**2**	**3**	**4**
Monitor	Shared information any crew member recognized about operational situation such as systems and communications.	○	○	○	○
Vigilance	Crew members remained alert of the environment and status of the aircraft.	○	○	○	○
Anticipation	Actively sought situational changes, threats and potential risks whi-which might impact, and considered suitable strategies in advance.	○	○	○	○
Analysis	Gathered information and used available resources to clearly identify the problem and potential risks.	○	○	○	○
Decision Making		**1**	**2**	**3**	**4**
Decision	Bottom lines were established. Chose an appropriate strategy from all information of team members and merit/demerit of each selection.	○	○	○	○
Action	All members understood chosen strategy and performed own tasks to implement the strateg y.	○	○	○	○
Critique	Compared desired outcomes with actual progress, reviewed and changed own performance.	○	○	○	○
Workload Management		**1**	**2**	**3**	**4**
Planning	Developed plans to avoid high workload at a safe and appropriate time.	○	○	○	○
Prioritizing	Operational tasks were prioritized considering with time limitation, volume of tasks and urgency.	○	○	○	○
Distribution	Assigned appropriate tasks to crew members and automated systems monitoring crew performance.	○			

* automated systems : FMS,A/P,A/T,etc

1

Figure 2. *Example of CRM Skills Rating Sheet.*

Scenarios for Simulated LOFT

Scenarios for simulated LOFT were generated based on Seamster *et al.* (1998), but that document specifies the Boeing 737 and uses routings within the United States. To adapt these for Japanese use, the aircraft fleet types were changed to the Boeing 767 and 777, and Japanese routings, air space and airports were substituted. As a result, four types of scenario were created.

A prepared script for the Raters described the event set, weather information, NOTAM, weight & balance, communications with ATC, cabin crew, company radio, and ground staff, and specified which CRM Skills were to be observed.

Selection of Raters

Assessment of inter-rater-reliability is an important issue in the evaluation of flight crews' CRM Skills. In these experiments, Raters were selected based on the following requirements.

(1) Job experience in a CRM training-related department of an airline. Knowledge of CRM skills is indispensable.
(2) Aircrew experience in at least one "glass-cockpit" airplane type.
(3) Ability to understand the proposed CRM Skills Behavioral Markers before the experiments.
(4) Aircrew_experience of the aircraft fleet used in the simulated LOFT is not necessary, since aircraft type-specific Standard Operating Procedures (SOPs) are irrelevant to CRM Skills.

Nine Raters (A to I) ware selected from Japanese airlines. Their experience is shown in Table 1. X, Y, and Z are their airlines, and the left column shows the airplane types with which they had crew experience.

Table 1. *Classification of nine Raters by Airline and Airplane Type.*

	X-Airline	Y-Airline	Z-Airline
B747-400	A, B	C	
B767	G	D	
B777	H	F	
Others	E		I

Simulated LOFT Experiments

Five sets (Cases #1–#5) of simulated LOFT sessions using the four scenarios were flown on B777 and B767 flight simulators, and the sessions were recorded to LOFT Videotapes. Using the five LOFT Videotapes, experiments to evaluate the Measurement Method were conducted in the following manner.

(1) The CRM Skills Behavioral Markers, Measurement Method and experiment procedure were explained to the Raters.
(2) Scoring was on a four-point scale: 1 denotes Ineffective, 2 Adequate, 3 Effective, and 4 Highly Effective. '3' is the reference standard.
(3) Only the degree of CRM skills practice is to be evaluated; whether or not a crew follows SOPs is irrelevant.
(4) The skills of the crew itself should be evaluated, not the skills of individual crewmembers.
(5) Comments should be recorded regarding CRM skills that are judged to be better or worse than the reference standard.
(6) A CRM Skills entry may be left blank in the case where Behavioral Markers cannot be observed in the crew.

After watching a video, each Rater completed the Rating Sheet and was then interviewed to determine the reasons for his scorings and to obtain general comments on the Measurement Method.

Results

Overall

Table 2 shows the average of rating (score) and the average of standard deviation (SD) across all Skills calculated for each case. The average rating is the highest for Case #1, and its SD is the second smallest. As already mentioned, Case #5 shows the greatest variance.

Fig._3 shows a plot of the average rated scores across all cases. It can be seen that Cases #1 and #3, which are based on same Scenario #1, were rated relatively consistently, while Cases #4 and #5 show greater variance of the average scores awarded by the Raters. When looking at the relative scores between the cases rated by each Rater, consistency is observed for 8 out of the 9 Raters (excepting G), excepting Case #5.

Features of each CRM Skills Behavioral Marker

Average rating and Average SD

For each Behavioral Marker, the scores of the nine Raters and five cases were totaled and the average rating and average variance were calculated. While the average rating for any Behavioral Marker was concentrated at the standard score level three (from 2.932 to 3.111), the average variance extended from 0.054 to 0.402. From this, it is understood that there is a difference in ratings between Raters or between cases.

The distributions of average rating and average variance for each CRM Skills Behavioral Marker are plotted in Fig._4. As expected, the figure shows a strong correlation between average rating and average variance (r=0.505); that is, average variance tends to grow for Behavioral Markers which receive high scores.

Correlations between Behavioral Markers

It was assumed that the evaluated score for a single Behavioral Marker might influence the score for other Behavioral Markers. The correlation coefficients between Behavioral Markers calculated from all the gathered data. were analyzed to examine the extent of this influence.

"Total Team Performance" has a relatively strong correlation with "Leadership", "Climate" and "Assertion". Correlation was observed not only between Behavioral Markers that belong to the same Element, such as "Monitor" versus "Leadership", but also between Markers in different Elements, such as "Leadership" versus "Climate" or "Distribution" versus "Prioritizing".

On the other hand, there was hardly any correlation in the combinations of “Planning” versus “Assertion”, “Critique” versus “Anticipation”, and “Two-Way Communication” versus “Briefing”.

Table 2. *Average Rating and Average of Standard Deviation (SD) for each Case#.*

LOFT No.	Scenario No.	Average Rating	Average SD
Case #1	1	3.100	0.331
Case #2	2	2.983	0.327
Case #3	1	3.035	0.399
Case #4	3	2.916	0.401
Case #5	4	2.982	0.503

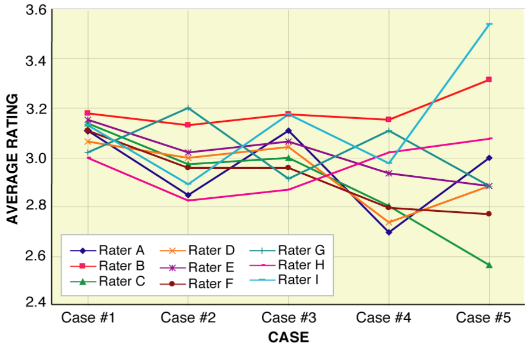


Figure 3. *Average Rating for Each Case.*

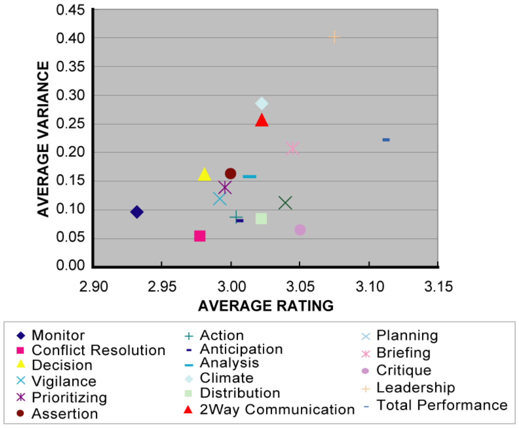


Figure 4. *Scatter Plot of Average Rating VS Average Variance.*

Number of Empty Behavioral Markers Columns

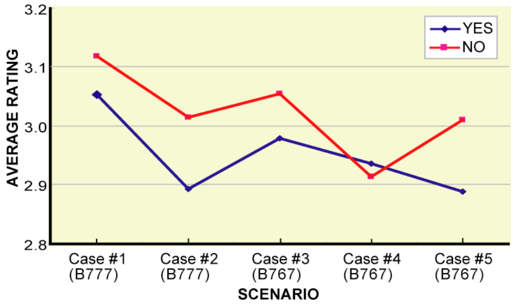
Raters were permitted to leave a column in the CRM Skills sheet blank in the case that the corresponding crew behavior was not observed, or for other reasons. This implies that skills with more empty columns in the Rating Sheet are more difficult to observe.

Relatively high numbers of empty columns were recorded for the Behavioral Markers “Analysis”, “Critique”, and during the Takeoff/Climb and Cruise flight phases. The following narrative comments related to the empty columns were obtained from interviews with Raters.

- Some items were difficult to score because they were not visually prominent in the video record.
- Understood that “Vigilance” is identical to “Anticipation”, only one of these was scored.
- “Critique” was not scored but was included in “Communication” in “Overall”.

Moreover, it is considered that difficulty in identifying transitions between flight phases caused more empty columns during take-off and cruise. For example, some cases contained a missed approach but the timing of the transition is not clearly apparent in the video recordings.

Raters’ Aircraft Type Experience



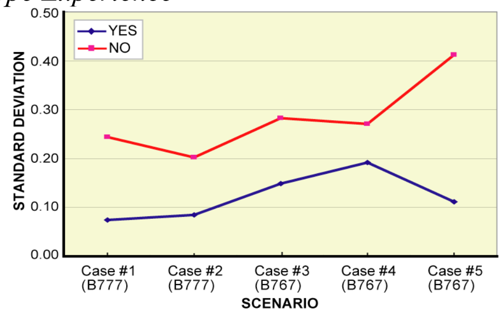


Figure 5. *Average Rating Sorted by Experience YES/NO.* Figure 6. *Average SD Sorted by Experience YES/NO.*

An analysis was conducted to investigate the effect of Raters’ experience with the type of aircraft used in the scenario. The average and standard deviation of ratings in each group (w/ or w/o experience) are plotted in Figs. 5

and 6 respectively. Significant differences at a 5% threshold level were observed for both average and STD between experienced and non-experienced groups (average: p=0.013 and SD: p=0.014). The correlation of differences between cases was strong for averages (r=0.665) but not for SDs (r=0.169).

From these results, it is concluded that a Rater with experience of the aircraft type used in the LOFT tends to award lower scores, and variance between Raters in the experienced group is small.

Raters' Comments

After the Raters had completed the Sheets, they were asked to comment freely on the experiment. The obtained comments are summarized below:

(1) CRM Skills Rating Sheet

- Some CRM Skills Behavioral Markers should be unified into a single rating item to make it easy to evaluate crew behavior.
- "Communication" and "Team Building & Maintenance" may be rated not Overall but rather for each flight phase, as for other skills. Rating a skill "Overall" disguises differences between flight phases.

(2) Background of the Raters

- Rating might be easier if the events sets in the scenarios are known to the Raters in advance, while it is impossible in LOSA (Line Operational Safety Audit).
- Raters' knowledge of the SOPs of the type of aircraft is not essential, but it does help the rating task.
- Some raters with flight inspector experience may easily award low scores (such as 2). On the other hand, other Raters may find it hard to do so because they may imagine that low scores might affect the certification of the crew.

(3) Method of scoring

- Four degrees of rating level is too little/much.
- There seem to be two ways of rating an event related to crew error. When a crewmember makes an error and notices (corrects) it by himself, Raters may either evaluate him negatively for making the error, or may evaluate him positively for correcting it.
- If the rating is done long after viewing the video, Raters are allowed to think about the reasons for the crew's behavior and it if becomes understandable, and then the score tends to be higher.

(4) Others

- There seem to be too many event sets. It was therefore felt that the scenario was designed more to look at technical skills and the evaluation of SOP practice than CRM Skills.
- The experiment provided a good opportunity to know the status of operations and recommended CRM Skills of other companies.

Discussion

Correlation between Behavioral Markers

When the score average variances of the Raters are examined it becomes clear that some Behavioral Markers have large variance. There were some Behavioral Markers for which the Rating Sheet scores were often left blank. The causes of this are considered to be not only differences in individual Raters' judgment, but also the format of the CRM Skills Rating Sheet itself.

At the early stage of this research, we presented Raters with concrete examples of crew behaviors that could be expected to be observed corresponding to CRM Skills Behavioral Markers. In response to this, it was suggested that it might not be clear how to evaluate skills if behaviors other than the examples provided were observed, and that with detailed examples of crew behavior anyone could be a Rater without training. As the result of this feedback, only guidelines on scoring were presented to Raters for this experiment, with neither detailed examples of crew behaviors nor how to guidance on how to evaluate them.

In this experiment, Raters commented that while each Behavioral Marker was to be scored separately in the Rating Sheet, some crew behaviors corresponded to more than one Behavioral Marker and in such cases, the Behavioral Markers should be unified to form a single column. For example, "Vigilance" and "Anticipation", "Monitor" and "Analysis", "Planning" and "Prioritizing", "Conflict Resolution" and "Briefing". These Behavioral Markers were often left blank in Rating Sheet. For a Behavioral Marker which is strongly correlated with another, if its Rating Sheet column is left blank then is possible to guess the score from that of its correlated Behavioral Marker.

It is therefore understood that while there is no need to improve the CRM Skill Behavioral Markers themselves, there is a need to review and restructure the measurement items in the CRM Skills Rating Sheet. However, unification of some Behavioral Markers into a single measurement item makes the evaluation of CRM skills coarser, and is not always necessarily better from the viewpoint of identifying a crew's inadequate CRM Skills.

Standardization of Raters

As already discussed, Raters' type experience with the aircraft in the simulated LOFT affected their scoring behavior. Although Raters were instructed that they should not score execution of SOPs, their knowledge of the aircraft SOPs did in fact influence their ratings. Standardization of Raters should therefore be conducted taking into account their type experience. In this experiment, no limitations or requirements were stipulated on Raters' flight crew backgrounds, and no standardization was conducted prior to the experiment. Although the authors had assumed that that adequately developed CRM Skills Behavioral Markers would require no standardization in advance, the experiment result highlighted the necessity of Rater standardization.

A method for Rater standardization widely_used by world airlines is as follows. Two or more Raters watch a recorded LOFT session together, and then compare their own rating scores with each other and with a Standard Score while discussing. Repeating this procedure minimizes scoring variation between Raters. In the present experiment, some Raters commented that it was very effective to know the opinion of other Raters. However, contrary to this, Case #5 was scored after the greatest amount of discussion but showed the highest variation between Raters' scores. The reason may be to do with the following comments concerning Case #5: "I had became accustomed to the experiments, so it came to be able to evaluated that I thought", and, "By comparing with the past scores of other Raters, I noticed that my own scores had been relatively high, so I reduced my scores in this case." Consequently, it is concluded that Raters without their own firm rating criteria were easily influenced by the opinions of other Raters, and a familiarization with use CRM Skills evaluation method is necessary before the rating session.

For Rater standardization, the method mentioned above seems not to be only effective for standardizing Raters' scoring criteria, but is also effective for familiarization with how to rate before the actual rating session.

Degree of Scoring Level

Many Raters commented that the current four-degree scale of scoring is confusing. The current scale of "1" to "4" gave Raters an impression that "2" means "unacceptable", and it was difficult to decide whether to score a "2" or a "3" if minor deficiencies were observed for a skill. The Rating Sheet is one measurement tool for evaluating crew CRM Skills levels, and its main objective is to extract skills that require improvement. It is thought that current four level scoring scale should be revised to allow Raters to score "2" more easily. However, if the Sheet is used only for LOFT, where it is assured that the score record is immediately discarded after the training session, the current four degrees is perfectly acceptable.

Conclusion

JAXA has developed a CRM Skills Measurement Method to evaluate the effectiveness of flight crew CRM Skills training. The method was developed by means of a survey, interviews and several simulated LOFT experiments. Nine Raters evaluated crew CRM Skills performance using this Measurement Method in LOFT experiments.

Analysis of the ratings and consideration of Raters' comments indicate that the concept of the CRM Skills Measurement Method is sound and that suitable CRM Skills Behavioral Markers are available, but there is room for improvement in the Rating Sheet and in the method by which the rating is carried out. The importance of inter-Rater-reliability was recognized, and insights into CRM Skills Measurement Method were also obtained.

References

Japan Civil Aviation Bureau_(JCAB) (1998). Order of execution of CRM training to aircrew. Flight Standards Order Vol._410, Engineering Department, Ministry of Land, Infrastructure and Transport .

Seamster, T. L., Edens, E. S., McDougall, W. A., and Hamman, W. R. (1998). Observable Crew Behaviors in the Development and Assessment of Line Operational Evaluations (LOE's), FAA OP-US/6821.

Iijima T.,_Noda F., Sudo K., Muraoka K. and Funabiki K._(2003). Development of CRM skills behavioral markers, TR-1465, National Aerospace Laboratory Report.

Flight Crew Technical Service, Flight Operations, Japan Airlines Co., Ltd. (JAL)_(2003). Research Report of LOFT/LOE and CRM skills.

Helmrech, R. L., Klinect, J. R, Wilhelm, J. A and Sexton, J. B. (2001). The Line Operations Safety Audit (LOSA), Proc. of 1st LOSA week.

Acknowledgments

This research is supported by ANA, JALI, JALJ and other the staff of other airlines. The authors wish to thank them all.