DEVELOPMENT OF A METHOD FOR CRM SKILLS MEASUREMENT

Hiroka Tsuda*, Tomoko Iijima* and Fumio Noda*
***Japan Aerospace Exploration Agency**

Keywords: *Crew Resource Management, skill indicators, behavioral markers, Line Oriented Flight Training*

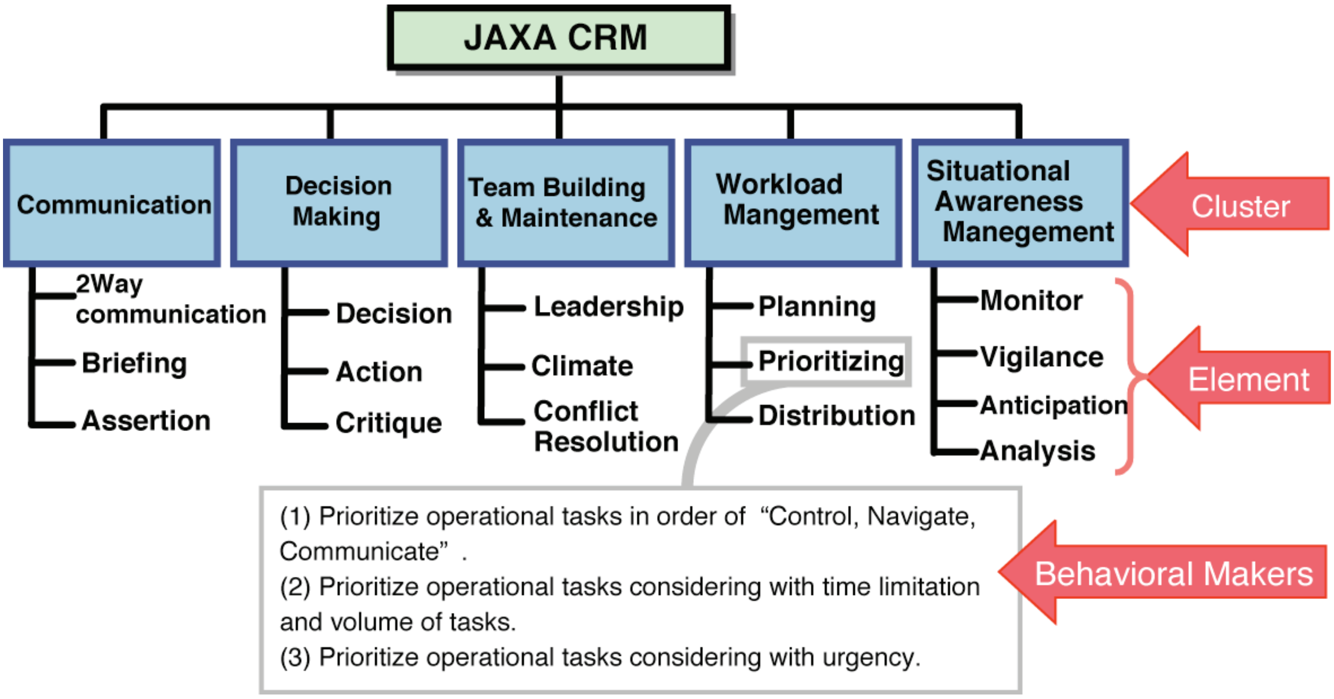
Abstract

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

1 Introduction

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 in 70% of fatal accidents, and so addressing these should yield further reductions in the accident rate. Here, “human factors” refers more to non-technical skills than technical skills; for example, effective communication among the flight crew.

Flight crews have access to a variety of resources, including human resources such as



(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.

Fig. 1. JAXA Proposed CRM Skills.

cabin crew and ground staff, hardware resources such as airports and other airplanes, and information resources such as manuals and information from Air Traffic Control. CRM (Crew Resource Management) is an approach to preventing aircraft accidents by using available resources in an effective manner.

After Helmric revealed that most human factors-related incidents are caused by inappropriate crew coordination, importance began to be placed in 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, and Japan Airlines began CRM training in 1986 [1].

It is considered that concrete behavioral indicators are necessary for conducting CRM training effectively, and so from 1999 to 2002 JAXA has been developing CRM Skills Behavioral Markers with support of airlines (Japan Air System, All Nippon Airways and Japan Airlines) that take into account the particular behavioral and psychological characteristics of Japanese crewmembers, which would be suitable for the Japanese flight crews operating in a domestic environment. Here “CRM skills” was defined to be the ability to carry out CRM.

Figure 1 shows the CRM Skills proposed by JAXA. CRM Skills are classified into five clusters with 3–4 skills elements in each. Each skills element has two or more CRM Skills Behavioral Markers. The “Prioritizing” CRM Skills element has the three Behavioral Markers: (1) Prioritize operational tasks according to “Control, Navigate, Communicate”, (2) Prioritize operational tasks considering time limitations and task magnitudes, (3) Prioritize operational tasks considering urgency.

To assess the effect of CRM training and to provide feedback to a training program to reduce possible training inadequacies, we have developed a CRM Skills Measurement Method that can identify how far CRM skills have been learned [3]. The proposed method utilizes a subjective rating technique based on the JAXA CRM Skills Behavioral Markers. A series of experiments using line-oriented flight simulations was conducted to evaluate the effectiveness and usability of the method.

This paper describes the development of the CRM Skills Measurement Method and evaluation of its applicability.

2 Development of a CRM Skills Measurement Method

Development of the CRM Skills Measurement Method consisted of three phases as shown in Fig. 2: a preliminary study, design of a prototype rating sheet, and evaluation of the method.

In the preliminary study, we studied various airlines that incorporate CRM skills into LOFT[1] and/or LOE[2], and designed a first prototype (No. 1) of the CRM Skills Measurement Method based on the results of this survey. After several experiments using airline pilots and pilot interviews to evaluate this prototype, we developed the No. 3 measurement method prototype. In the third development phase, this No.3 prototype was then evaluated by LOFT experiments using a new LOFT scenario.

2.1 Design of CRM Skills Measurement Method

To design the CRM Skills Measurement Method, a survey was made of airlines that incorporate CRM skills into LOFT and/or LOE [4], and based on the survey results, an initial prototype of the CRM Skills Measurement

[1] LOFT (Line Oriented Flight Training)

Simulator-based training in which normal / abnormal / emergency situations are presented in simulated line operations to improve a crews’ ability to practice CRM in solving problems and completing a flight by themselves. Video recordings are usually used for review after training sessions, and individual crewmember performance is not evaluated.

[2] LOE (Line Operational Evaluation)

Similarly to simulators are used to present normal / abnormal / emergency situations, but the performance of individuals is evaluated under the AQP (Advanced Qualification Program) introduced by the US Federal Aviation Authority.

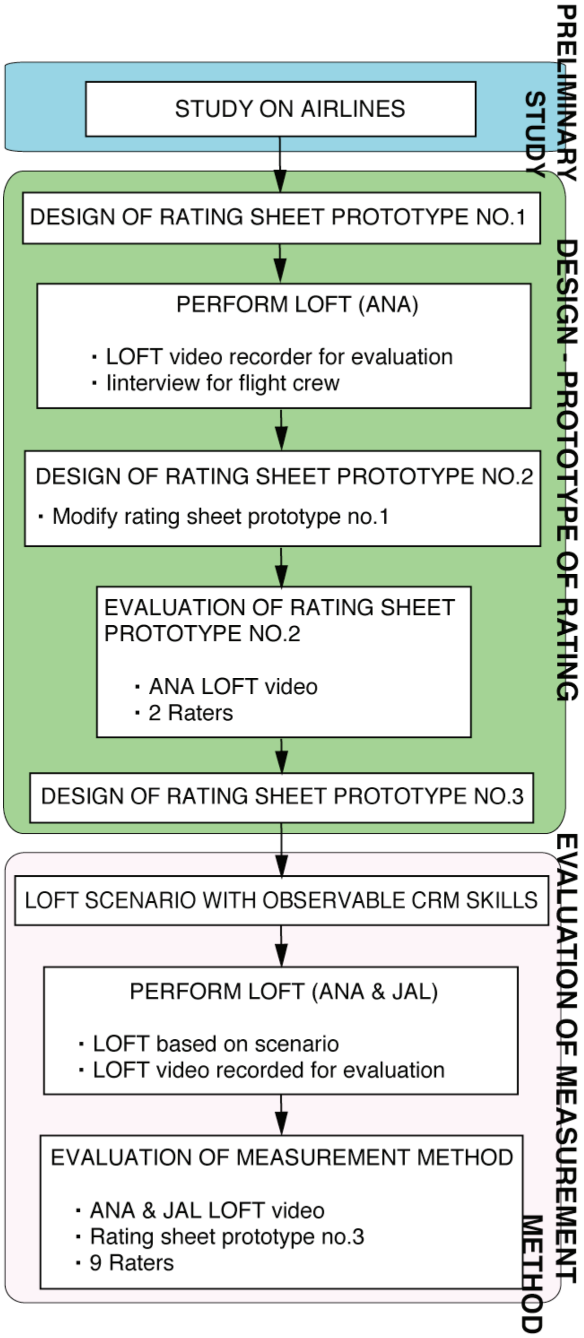

PRELIMINARY STUDY
STUDY ON AIRLINES
DESIGN - PROTOTYPE OF RATING
DESIGN OF RATING SHEET PROTOTYPE NO.1
PERFORM LOFT (ANA)
・LOFT video recorder for evaluation
・Iinterview for flight crew
DESIGN OF RATING SHEET PROTOTYPE NO.2
・Modify rating sheet prototype no.1
EVALUATION OF RATING SHEET PROTOTYPE NO.2
・ANA LOFT video
・2 Raters
DESIGN OF RATING SHEET PROTOTYPE NO.3
EVALUATION OF MEASUREMENT METHOD
LOFT SCENARIO WITH OBSERVABLE CRM SKILLS
PERFORM LOFT (ANA & JAL)
・LOFT based on scenario
・LOFT video recorded for evaluation
EVALUATION OF MEASUREMENT METHOD
・ANA & JAL LOFT video
・Rating sheet prototype no.3
・9 Raters

Fig. 2. Developing of CRM Skills Measurement Method.

Method, prototype No. 1, was developed for trial purposes. This method is used not for the evaluation of the individual crewmembers' performance but to evaluate a flight crew as a whole (a set of a captain and a copilot). The method used an initial prototype CRM Skills Rating Sheet (hereafter, Rating Sheet) on which "Raters" (training personnel who evaluate the crew's CRM skills) record scores for CRM skills they observe, at this stage on a three-point scale. The Rating Sheet was designed to be easy for Raters to use and analyze. 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, producing a second prototype. Major changes to the initial prototype were:

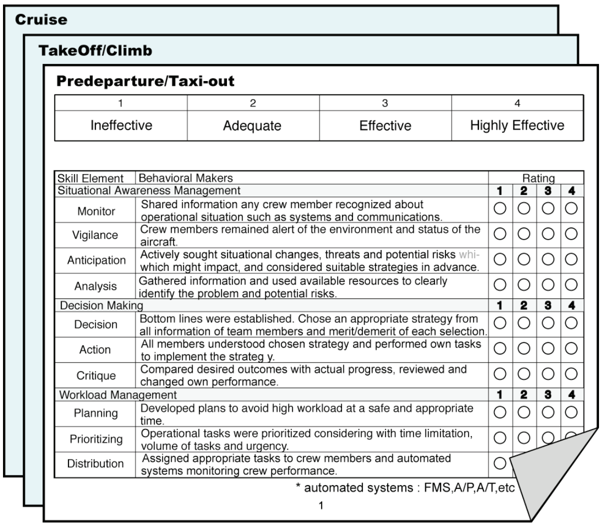
・The measurement standard was changed from a three-point rating score to a four-point rating score. (With only three points in the scale, it was too easy to select the central point in scoring, while prevented detailed analysis.)

・An "Observations" column was added to the Rating Sheet. This allows Raters, when they observe a crew carrying out a CRM skills, to first to record a checkmark, and to defer scoring until later.

To improve the second prototype, a preliminary rating experiment was conducted by applying it to a video recording of a sample LOFT session (the "LOFT VCR" below), with two Raters. Feedback from the Raters was then used to produce a third version of the rating sheet. One modification was to change the rating (score) column from a space for a numerical entry (the Rater having to write a number) to a sort of mark sheet.

On the Rating Sheet, the skills of Situational Awareness Management, Decision Making, and Workload Management were to be measured for each flight phase, while the Communication skill and Team Building and Maintenance skill were to be measured overall, i.e. over the entire flight. Fig. 3 shows part of the Rating Sheet.

In addition, an A3 size Observation Sheet on which flight profile and event sets are drawn (hereafter, Observation Sheet) was prepared. This Observation Sheet was designed so that Raters could write memos freely during the experiment.



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

Fig. 3. Example of CRM Skills Rating Sheet.

2.2 Evaluation experiments of CRM Skills Measurement Method

Experiments to evaluate the CRM Skills Measurement Method were conducted in the following manner. First, some sets of simulated LOFT (that is, a line-oriented flight simulation, but with the post-flight critique omitted) were conducted and the crew's behavior in the cockpit was recorded on video. Then, a Rater completed the prepared Rating Sheet while watching the LOFT VCR.

2.2.1 Scenarios for simulated LOFT

Scenarios for simulated LOFT were generated based on [2], but with twelve event sets in each scenario instead of the original ten. The CRM skills to be observed in association with each Event Set were also generated. Each scenario was generated based on the following rules.

(1) The scenario should reproduce actual operations as much as possible.

(2) Each flight is assumed to be a full flight from takeoff to landing, although it is possible to shorten the cruise segment under certain circumstances.

(3) All possible outcomes are permissible – that is, the crew's behavior and choices are not forced in any specific direction.

(4) Sufficient time is built in to the scenarios to allow the flight crew to have discussions.

(5) Concepts of threats in the scenarios included the following.

(6) Instantaneous change of weather conditions.

(7) Minimum equipment list and permissible deferred repair standards are applied. If the scenario called for systems to be defective, they are assumed to be partially inoperative.

The prepared script for the instructors described the event set, communications with ATC, cabin crew, company radio, and ground staff, and specified which CRM skills were to be observed. Weather information (ATIS, METAR, TAF) and NOTAM information (SNOTAM) were also prepared. The co-operating airlines were requested to prepare flight plans and weight and balance sheets.

The scenarios in Federal Aviation Administration documents specify the Boeing 737 as the fleet aircraft, and use routings within the United States. To adapt these for Japanese use, the aircraft fleet types were changed to the Boeing 767 and Boeing 777, and Japanese routings, air space and airports were substituted.

Four types of scenario were created as mentioned above, and these are summarized in Table 1. Event sets 11 and 12 were derived from event sets 10 and 7 respectively by modification. An example of a correspondence between an Event Set and the CRM skills to be observed is shown in Table 2. A CRM skill to be looked for is given in square brackets (e.g. [Workload Management]) and concrete examples of corresponding behavior which may be observed are given below it. Figure 4 shows the flight profile of scenario #1.

Table 1. Four scenario components and five LOFT cases.

Scenario No. (Case No.)	Flight Route	Airplane	Event Set
#1 (Case #1, #3)	AXT-HND	B777, B767	1,2,3,4,5
#2 (Case #2)	SDJ-HND	B767	6,7,8
#3 (Case #4)	*XX*-HND	B767	10
#4 (Case #5)	KMQ-HND	B777	6,11,12

Table 2. Example of relation between Event Set and observable CRM skills.

Event Set No.	Event Set contents	CRM skills behavior
1 – Pre-departure/Taxi-out	・In Winter, Adverse Weather. Slippery Taxiway and Runway. ・Cb on departure route. ・No N1 Indication on #1 Engine.	[Planning] PF had aircraft deiced, anti-iced and planned for winter operations SOP. PF briefed CB. [Decision Making] PF analyzed destination WX and requested takeoff alternate.
2 – Takeoff/ Climb	・Cb in front. ・Route change. ・VOR L INOP.	[Situational Awareness] Crew discussed Cb location before it could become a problem. PF requested higher altitude. [Workload Management] Crew set clear priorities for tasks and their order.
3 – Climb	・#1 Engine fail. ・WX of Departure airport AXT is below Landing Minima.	[Workload Management] PF directed PNF to deal with engine problem. PNF performed needed C'k list and announced compliance. [Planning] PF stated that they cannot return to AKT. Crew assessed one engine landing with WX at diversion field.

*An alternate airport (Takeoff Alternate Airport) is designated within one hour before actual flight when the weather conditions at the departure airport are less than Landing Minima and more than Takeoff Minima.

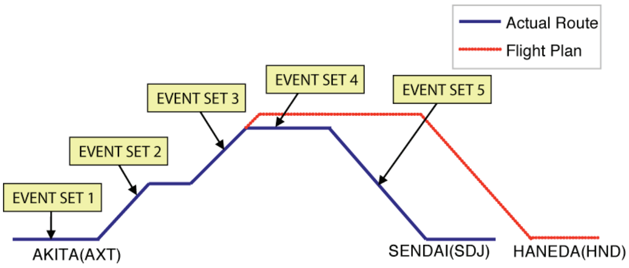


Fig. 4. Profile of Scenario#1 (Flight route is AXT-HND, five event sets happen en route).

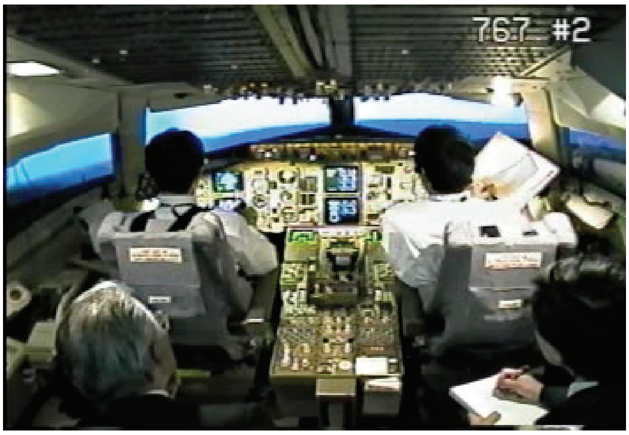


Fig. 5. LOFT Simulation.

2.2.2 LOFT video recording in the simulator

Five sets of simulated LOFT sessions using the four scenarios were flown on B777 and B767 flight simulators. As in normal LOFT, the flight crews were not told the detailed scenario contents, such as implemented events, in advance.

Figure 5 shows a video still from a session in progress. An instructor (normally called a "facilitator" in LOFT), who controls the flight simulator, and the experimenter occupy the two seats behind the flight crew. The instructor played all non-flight crew roles such as air traffic controller, ground staff and cabin crew. Each session lasted roughly two hours.

2.2.3 Selection of Raters

The Raters for the experiments 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). Persons who had previously participated in this research as subject flight crew were excluded.

(2) Air crew experience in at least one "glass-cockpit" airplane type.

(3) Ability to understand the proposed CRM Skills Behavioral Markers before the experiments.

It was assumed that the Raters do not have to have aircrew experience of the aircraft type used in the LOFT sessions, since was is thought that observing whether the subject crew adheres to SOPs is irrelevant to CRM skills measurement.

Table 3. 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

Nine Raters (A to I) were selected from Japanese airlines. Their experience is shown in Table 3. X, Y, and Z are their airlines, and the left column shows the airplane types with which they had crew experience.

2.2.4 Experiments

The CRM Skills Behavioral Markers and experiment procedure (concrete Rating Method) were explained to the Raters before the experiments. The following issues were emphasized in the briefing.

(1) Scoring is on a four-point scale: 1 denotes Ineffective, 2 Adequate, 3 Effective, and 4 Highly Effective. '3' is the reference standard.

(2) Only the degree of CRM skills practice is to be evaluated; whether or not the crew follow SOPs is not relevant.

(3) The skills of the crew itself should be evaluated, not the skills of individual crewmembers.

(4) Comments should be recorded regarding CRM skills that are judged to be better than or worse than the reference standard.

(5) A CRM Skills entry may be left blank in the case where Behavioral Markers cannot be observed in the crew.

The experiment apparatus consisted of the following.

- Video recordings of five simulated LOFT sessions
- Four LOFT scenarios
- A table of CRM Skills Behavioral Markers
- Observation Sheet
- CRM Skills Rating Sheet

After each Rater had completed the Rating Sheet after watching the video, he was interviewed to determine the reasons for his scorings and to obtain general comments on the CRM Skills Measurement Method.

Two or more Raters participated in the experiment together in sets. Each set of Raters watched the videos together in the same room, then completed their own Rating Sheets independently. The experiment took two days for each set of Raters, with two video sessions for the first day, then three for the second day.

3 Results

The results obtained from the experiment are presented below.

3.1 Overall

Figure 6 shows a plot of the average rated scores across all cases.

From the figure, 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), but excepting Case #5.

Table 4 shows average of rating (score) and average of standard deviation (SD) across all Skills calculated for each case. The average rating was the highest for Case #1, and the SD is second smallest. As already mentioned, Case #5 shows the greatest variance.

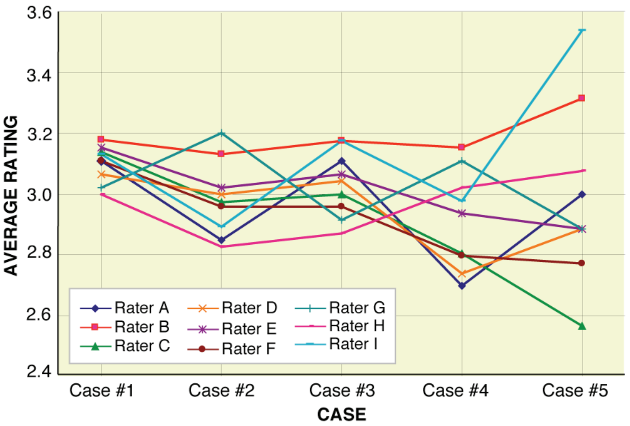


Fig. 6. Average Rating for Each Case (sorting by Rater and Case#).

Table 4. Average rating and average of standard deviation (SD) for each Case#.

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

Table 5. Average rating and average of SD for each Rater.

Rater	Average rating	Average SD
A	2.948	0.460
B	3.170	0.414
C	2.966	0.318
D	2.952	0.305
E	3.024	0.284
F	2.905	0.448
G	3.024	0.452
H	2.948	0.405
I	3.091	0.464

Table 5 shows average of rating and average of SD across all Skills calculated for each Rater. Looking at individual Raters, Rater B gave the highest average of ratings (that is, Rater B awarded the highest scores through the five cases). Rater I’s ratings had the greatest variance, while Rater E’s had the lowest. Rater A and Rater H consistently awarded low scores.

3.2 Features of each CRM Skills Behavioral Marker

3.2.1 Average rating and Average SD

The results of nine Raters and five cases of each Behavioral Markers were totaled, and the Average rating and Average variance were calculated. Tables 6 and 7 show these in ascending order.

While the average rating for any Behavioral Marker is concentrated on a standard three point level from 2.932 to 3.111, the average variance extends from 0.054 to 0.402.

Table 6. Behavioral Markers (CRM skills element) Average rating (ascending order).

	CRM skills element		Average rating
F/P	SA	Monitor	2.932
Over All	TB	Conflict Resolution	2.978
F/P	DM	Decision	2.981
F/P	SA	Vigilance	2.992
F/P	WM	Prioritizing	2.996
Over All	Com	Assertion	3.000
F/P	DM	Action	3.004
F/P	SA	Anticipation	3.006
F/P	SA	Analysis	3.014
Over All	TB	Climate	3.022
F/P	WM	Distribution	3.022
Over All	Com	2Way Communication	3.022
F/P	WM	Planning	3.040
Over All	Com	Briefing	3.044
F/P	DM	Critique	3.051
Over All	TB	Leadership	3.075
Over All		Total Team Performance	3.111

Table 7. Behavioral Markers (CRM skills element) Average variance (ascending order).

	CRM skills element		Average variance
Over All	TB	Conflict Resolution	0.054
F/P	DM	Critique	0.065
F/P	SA	Anticipation	0.080
F/P	WM	Distribution	0.084
F/P	DM	Action	0.087
F/P	SA	Monitor	0.097
F/P	WM	Planning	0.112
F/P	SA	Vigilance	0.121
F/P	WM	Prioritizing	0.140
F/P	SA	Analysis	0.158
Over All	Com	Assertion	0.163
F/P	DM	Decision	0.163
Over All	Com	Briefing	0.207
Over All		Total Team Performance	0.222
Over All	Com	2Way Communication	0.257
Over All	TB	Climate	0.286
Over All	TB	Leadership	0.402

* F/P Flight Phase, SA: Situational Awareness,
DM: Decision Making, WM: Workload Management,
Com: Communication, TB: Team Building

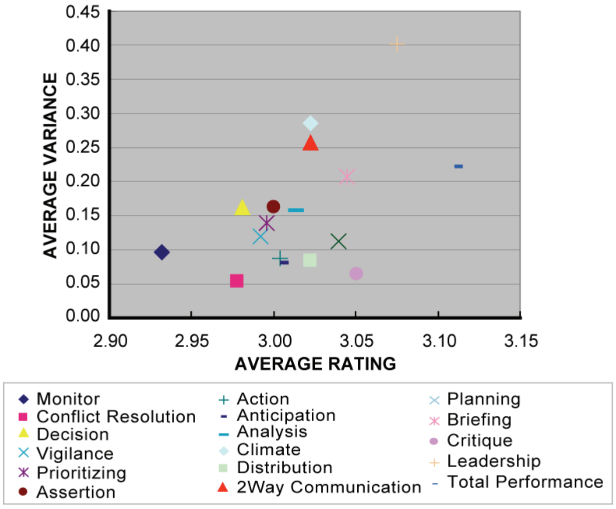


Fig.7. Scatter Plot of Average Rating VS Average Variance.

From this, it is understood that there is a difference in ratings between Raters or between the cases.

The distributions of average rating and average variance for each CRM Skills Behavioral Marker are plotted in Fig. 6. As expected from Figure 7, there is 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.

3.2.2 Correlations between Behavioral Markers

The correlation coefficients between Behavioral Markers were calculated from all the gathered data. It was assumed that the evaluated score for one Behavioral Marker might influence the score for other Behavioral Markers, and this analysis was conducted to verify the extent of this influence. The ten most highly correlated Markers are listed in Table 8, and the ten most weakly correlated in Table 9.

Table 8 shows that “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”. Moreover, it is observed that the Behavioral Markers which are correlated are scored not for each flight phases but as “Over all” skill ratings.

Table 8. Combination of Behavioral Markers that correlation is strong.

Combination of Behavioral Markers		correlation
Leadership,	Total Team Performance	0.656
Climate,	Total Team Performance	0.551
Climate,	Leadership	0.506
Assertion,	Total Team Performance	0.438
Monitor,	Leadership	0.426
Distribution,	Prioritizing	0.412
Leadership,	Conflict Resolution	0.390
Briefing,	Conflict Resolution	0.377
Leadership,	2Way Communication	0.361
Monitor,	Total Team Performance	0.350

Table 9. Combination of Behavioral Markers that correlation is weak.

Combination of Behavioral Markers		correlation
Planning,	Assertion	0.001
Critique,	Anticipation	-0.001
Briefing,	2Way Communication	-0.004
Critique,	Prioritizing	0.011
Briefing,	Decision	0.012
Vigilance,	Briefing	0.014
Analysis,	2Way Communication	0.029
Action,	Anticipation	-0.030
Critique,	Planning	0.031
Climate,	Vigilance	0.037

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

3.2.3 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. Behavioral Markers with one or more empty columns are listed in Table 10.

Table 10. Number of blank columns in Behavioral Markers.

Markers	P/T	T/C	Crz	D/A/L	O/A	Σ
Vigilance		1	1	1	/	3
Anticipation	2	1	4	1	/	8
Analysis	2	4	3	1	/	10
Critique	1	3	3	3	/	10
Planning		1	1		/	2
Prioritizing	1		1		/	2
Distribution	1	1	1	1	/	4
Assertion	/	/	/	/	1	1
Leadership	/	/	/	/	1	1
Conflict Resolution	/	/	/	/	1	1
Σ	7	11	14	7	3	42

*P/T:Predeparture/Taxi-Out , T/C:Takeoff/Climb , Crz:Cruise , D/A/L:Descent/Approach/Land , O/A:Over All

Table 10 shows that 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.

・ Understanding that "Vigilance" is identical to "Anticipation", only one of these was scored.

・ "Critique" was not scored but was included in "Communication" in "Overall".

On the other hand, it is considered that difficulty in identifying transitions between flight phases caused more empty columns during take-off and cruise. For example, in some cases, a crew conducted a missed approach but the timing of the transition is not clearly recorded in the video record.

3.3 Correlation between Raters

The correlation between Raters was calculated from all the Rating results of the nine Raters and five cases. The purpose of this is to guess extent to how the scoring tendencies of each Rater correspond to others. For instance, even if different Raters award different scores (a severe Rater or a tolerant Rater), a crew's behavior for a particular CRM skill may be judged to be adequate or inadequate if there is positive correlation between the ratings of two Raters. When a negative, opposite, correlation is seen, it means that two Raters came to opposite conclusions on the adequacy or inadequacy of the CRM skills action. The five strongest correlations between combinations of two Raters are shown in the left of Table 11 and five combinations with the weakest correlations are shown in right.

Table 11. Correlation between Raters.

descending order	
A G	-0.395
D H	0.372
C F	0.323
C I	-0.311
E F	0.295

ascending order	
B D	-0.002
B E	0.008
A I	0.010
D F	0.014
A D	0.022

A negative correlation is seen between Rater A and Rater G, and a positive correlation between Rater D and Rater H. However, these correlations between Raters are weak as the maximum is 0.395. This means the standardization of the Raters' evaluations was unsuccessful.

3.4 Raters' Aircraft Type Experience

An analysis was conducted to investigate the effect of Raters' experience with the type of aircraft used in the scenario. Table 12 shows groups of Raters by their aircraft type experience.

The average and standard deviation of ratings in each group were obtained and are plotted in Figs. 8 and 9 respectively.

Significant differences at a 5% threshold 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).

Table 12. Group of Raters by aircraft type experience.

	Group of Raters	
	Experienced	Not-experienced
Case #1 (B777)	F, H	A,B,C,D,E,G,I
Case #2 (B777)	F, H	A,B,C,D,E,G,I
Case #3 (B767)	D, G	A,B,C,E,F,H,I
Case #4 (B767)	D, G	A,B,C,E,F,H,I
Case #5 (B767)	D, G	A,B,C,E,F,H,I

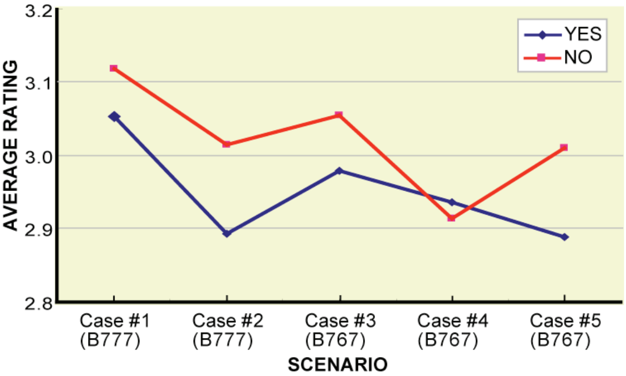


Fig.8. Average Rating Sorted by Experience YES/NO.

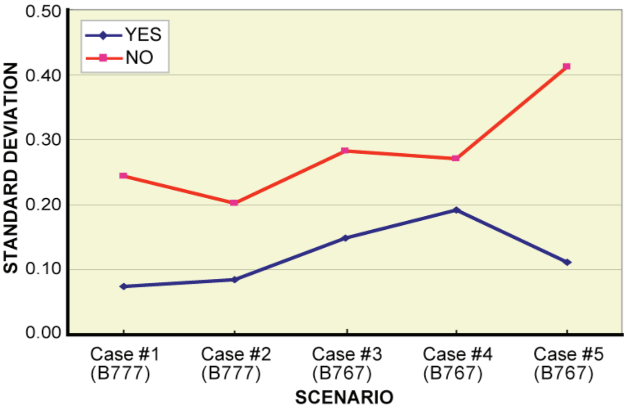


Fig. 9. Average SD Sorted by Experience YES/NO.

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

4 Raters' Comments

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

(1) CRM Skills Rating Sheet

・In Rating Sheets, some CRM Skills Behavioral Markers should be unified into one 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.

・Raters' knowledge of the SOPs of the type of aircraft in the training is not indispensable, 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 flight crew makes a error and notices (corrects) it by himself, Raters should evaluate him negatively looking at making the error, or should evaluate positively looking at correcting the error.

・If the rating is done long after viewing the video, Raters are allowed to think about the reason of crew's behavior and crew's behavior becomes understandable, becomes , and then the score tends to be higher.

[3] LOSA (Line Operational Safety Audit)
LOSA is a means of observing and evaluating crew performance by utilizing trained observers seat on cockpit jump seats during line flight [5].

(4) Others

- There seem to be too many events. Therefore, it was 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.

5 Discussion

5.1 Correlation between Behavioral Markers

Although looking at the score averages the ratings of the Raters seem to agree for each Behavioral Marker, when average variance is examined it becomes clear that some Behavioral Markers have large variance. Moreover, 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' judgments, but also due to 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 evaluate them.

In this experiment, Raters commented were 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 this case, 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 one, if its Rating Sheet column is left blank then is possible to guess the score from that of its correlated Behavioral Marker. In other words, a set of Behavioral Markers that have a strong correlation between them can be unified into a single 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.

5.2 Standardization of Raters

As shown in Figure 6 and Table 11, little correlation was observed between the Raters. As already discussed in section 3.4, the Raters' type experience with the aircraft in the LOFT simulations affected their scoring behavior. Although the Raters were told that they should not score the crew's execution of SOPs, the Raters' 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 variations between Raters. In the present experiment, some Raters commented that it was very effective to know the opinion of other Rater. 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 it thought", and, "By comparing with the past scores of other Raters, I noticed that 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 affected by the opinions of other Raters, and of rating using CRM skills evaluation method is necessary before the rating session.

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

5.3 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 crews 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.

6 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 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.

Acknowledgments

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

References

[1] Japan Civil Aviation Bureau, *Order of execution of CRM training to aircrew.* Flight Standards Order Vol.410, Engineering Department, Ministry of Land, Infrastructure and Transport , 1998.

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

[3] Iijima T.,Noda F., Sudo K., Muraoka K. and Funabiki K., *Development of CRM skills behavioral markers*, TR-1465, National Aerospace Laboratory Report, 2003.

[4] Flight Crew Technical Service, Flight Operations, Japan Airlines Co., Ltd. *Research Report of LOFT/LOE and CRM skills*. 2003.

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