2011

Development of a CRM Skills Measurement Method Including Threat and Error Management Concept

Tomoko Iijima

Hiroka Tsuda

Fumio Noda

DEVELOPMENT OF A CRM SKILLS MEASUREMENT METHOD INCLUDING THREAT AND ERROR MANAGEMENT CONCEPT

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

The Japan Aerospace Exploration Agency (JAXA) has developed a Crew Resource Management (CRM) Skills measurement method that includes a Threat and Error Management (TEM) concept and identifies a crew's level of CRM Skills by the way in which they manage human errors and threats. To validate the method, a CRM Skills measurement experiment was carried out by four raters using Line Oriented Flight Training (LOFT) scenarios. To increase inter-rater reliability, the raters collated their results to develop "True Scores". The experiment identified factors contributing to individual scoring differences between the raters and provided information for improving the CRM Skills rating sheet, inter-rater reliability training and LOFT scenarios.

CRM training is considered to be one of the most effective means of reducing human errors or minimizing their effects. Demands for greater operational safety and economy are will require more effective and efficient training and to achieve this, CRM Skills measurement will be necessary to evaluate objectively those skills that have been learned and to identify those that are inadequate. To allow CRM Skills training to be incorporated into pilot training and line operations, JAXA has developed behavioral markers by which CRM Skills may be identified (Iijima *et al.*, 2003) and a CRM Skills measurement method to assess the effectiveness of CRM training and to continue its improvement (Noda *et al.*, 2005, Tsuda *et al.*, 2006).

CRM Skills measurement relies upon the subjective scoring of crew behaviors observed in LOFT scenarios by "raters", who are typically pilots assigned to an airline's training department. However, since the scores are subjective, there is an issue of variability between different raters' scores for the same observed behavior. For example, Tsuda *et al.* found individual differences between nine raters who participated in CRM Skills measurement experiments. The main factor contributing to these individual differences was found to be differing rater viewpoints when observing crew behaviors. This indicates that inter-rater reliability training is necessary to standardize raters' viewpoints and their rating criteria.

The sixth annex of the International Civil Aviation Organization (ICAO) treaty now requires as an international standard that flight crew training must include human abilities and limitations, including Threat and Error Management (TEM), and assess competency in these areas. To meet this requirement, it is considered that a CRM Skills measurement method that includes a TEM concept is necessary.

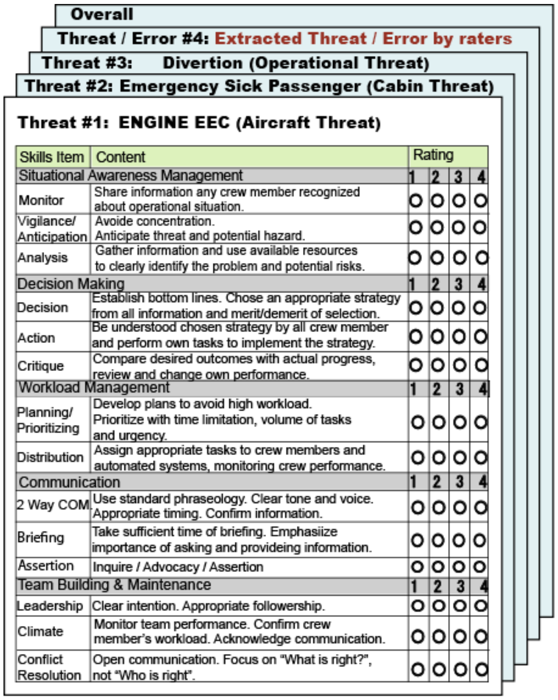
We therefore propose a new CRM Skills rating technique in which raters measure CRM Skills for each threat included in LOFT scenarios. The measurement adopts a TEM concept. To reduce individual scoring differences, raters are directed to evaluate only crew behavior when managing or mismanaging threat/errors, and to standardize their evaluations they hold discussions to derive a "True Score" result for each CRM skill item (Baker *et al.*, 1999). This paper describes the proposed CRM Skills measurement method and the verification of its validity.

Design of CRM Skills Measurement Method

To develop a CRM Skills measurement method, it is necessary to develop a CRM Skills rating sheet, effective LOFT scenarios that exercise CRM Skills for TEM, and to address inter-rater reliability training.

CRM Skills Rating Sheets

The proposed CRM Skills measuring method is a subjective scoring technique in which raters evaluate a crew's CRM Skills from their threat/error management behaviors and for each skill assign a numerical score on a four-point scale: 1 = "Ineffective", 2= "Adequate", 3= "Effective" and 4 = "Highly Effective". Scores are recorded on a CRM Skills rating sheet along with written narrative comments. The purposes of the CRM Skills rating sheet are to reduce score differences due to differences in individual rater viewpoints when observing threat management behaviors, and to conduct measurements appropriate to the TEM concept. To compare measuring CRM Skills on a per threat basis with measuring skills on a per flight phase basis, we developed two corresponding CRM Skills rating sheets shown in Figures 1 and 2.



Overall

Threat / Error #4: Extracted Threat / Error by raters

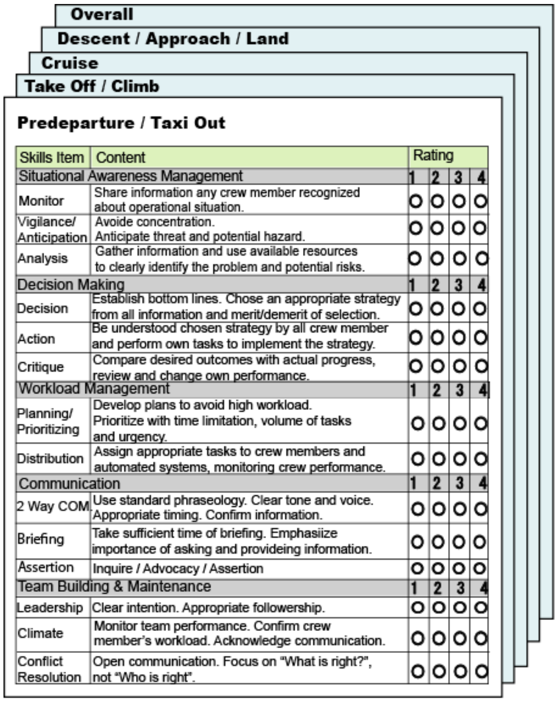
Threat #3: Divertion (Operational Threat)

Threat #2: Emergency Sick Passenger (Cabin Threat)

Threat #1: ENGINE EEC (Aircraft Threat)

Skills Item	Content	Rating			
Situational Awareness Management		1	2	3	4
Monitor	Share information any crew member recognized about operational situation.	○	○	○	○
Vigilance/ Anticipation	Avoide concentration. Anticipate threat and potential hazard.	○	○	○	○
Analysis	Gather information and use available resources to clearly identify the problem and potential risks.	○	○	○	○
Decision Making		1	2	3	4
Decision	Establish bottom lines. Chose an appropriate strategy from all information and merit/demerit of selection.	○	○	○	○
Action	Be understood chosen strategy by all crew member and perform own tasks to implement the strategy.	○	○	○	○
Critique	Compare desired outcomes with actual progress, review and change own performance.	○	○	○	○
Workload Management		1	2	3	4
Planning/ Prioritizing	Develop plans to avoid high workload. Prioritize with time limitation, volume of tasks and urgency.	○	○	○	○
Distribution	Assign appropriate tasks to crew members and automated systems, monitoring crew performance.	○	○	○	○
Communication		1	2	3	4
2 Way COM	Use standard phraseology. Clear tone and voice. Appropriate timing. Confirm information.	○	○	○	○
Briefing	Take sufficient time of briefing. Emphasiize importance of asking and provideing information.	○	○	○	○
Assertion	Inquire / Advocacy / Assertion	○	○	○	○
Team Building & Maintenance		1	2	3	4
Leadership	Clear intention. Appropriate followership.	○	○	○	○
Climate	Monitor team performance. Confirm crew member's workload. Acknowledge communication.	○	○	○	○
Conflict Resolution	Open communication. Focus on "What is right?", not "Who is right".	○	○	○	○

Fig. 1 *Per Threat CRM Skills Rating Sheet*



Overall

Descent / Approach / Land

Cruise

Take Off / Climb

Predeparture / Taxi Out

Skills Item	Content	Rating			
Situational Awareness Management		1	2	3	4
Monitor	Share information any crew member recognized about operational situation.	○	○	○	○
Vigilance/ Anticipation	Avoide concentration. Anticipate threat and potential hazard.	○	○	○	○
Analysis	Gather information and use available resources to clearly identify the problem and potential risks.	○	○	○	○
Decision Making		1	2	3	4
Decision	Establish bottom lines. Chose an appropriate strategy from all information and merit/demerit of selection.	○	○	○	○
Action	Be understood chosen strategy by all crew member and perform own tasks to implement the strategy.	○	○	○	○
Critique	Compare desired outcomes with actual progress, review and change own performance.	○	○	○	○
Workload Management		1	2	3	4
Planning/ Prioritizing	Develop plans to avoid high workload. Prioritize with time limitation, volume of tasks and urgency.	○	○	○	○
Distribution	Assign appropriate tasks to crew members and automated systems, monitoring crew performance.	○	○	○	○
Communication		1	2	3	4
2 Way COM	Use standard phraseology. Clear tone and voice. Appropriate timing. Confirm information.	○	○	○	○
Briefing	Take sufficient time of briefing. Emphasiize importance of asking and provideing information.	○	○	○	○
Assertion	Inquire / Advocacy / Assertion	○	○	○	○
Team Building & Maintenance		1	2	3	4
Leadership	Clear intention. Appropriate followership.	○	○	○	○
Climate	Monitor team performance. Confirm crew member's workload. Acknowledge communication.	○	○	○	○
Conflict Resolution	Open communication. Focus on "What is right?", not "Who is right".	○	○	○	○

Fig. 2 *Per Flight Phase CRM Skills Rating Sheet*

Selection of Raters

We selected four raters, identified here as A, B, C and D, for the CRM Skills measuring experiment. Each rater was an experienced captain (average flying time: 9,125 hours, average pilot in command time: 3,025 hours) who had worked in a CRM training-related department of an airline. Rater B had experience as a LOFT instructor and rater D had experience as a check airman. Rater A had aircrew experience of the aircraft type in one of the LOFT scenarios (scenario 3) mentioned below. All the raters learned the proposed CRM Skills behavioral markers, the scoring procedure and the scenario contents before the experiments.

Simulated LOFT Scenarios

Three simulated LOFT scenarios to measure CRM Skills were selected from existing recordings of LOFT exercises. Figure 3 shows the threat codes (Klinect *et al.*, 2001, e.g. Aircraft Threat) included in each scenario. Scenario 3 includes many kinds of threat types while the others have fewer. Scenarios 1 and 2 include three Aircraft threats. In scenario 2, Aircraft threats appear continuously during the Takeoff/Climb phase. In scenario 1, on the other hand, although Aircraft threats appear continuously during the Descent / Approach / Land phase, an Arrival threat is inserted between Aircraft threats in this phase.

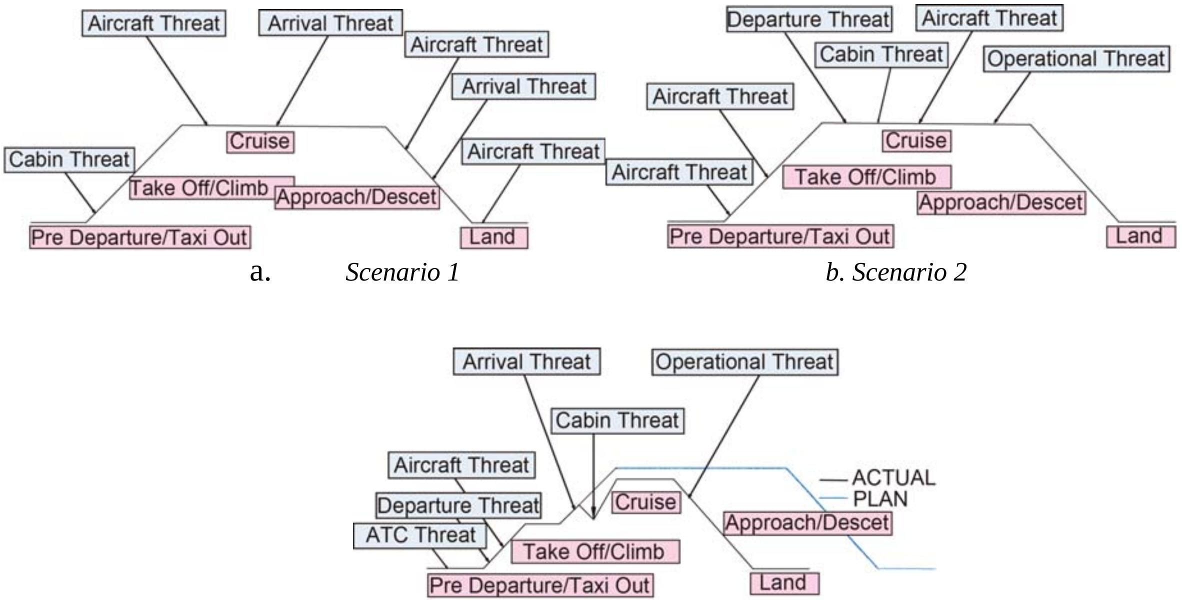
Experimental procedure

The experiment was carried out in three steps:

(1) The raters watched video recordings of the three simulated LOFT sessions and completed both types of CRM Skills rating sheets (per flight phase and per threat). The raters were also asked to make notes as appropriate on a CRM Skills observation sheet (Noda *et al.*, 2005, Tsuda *et al.*, 2006) while watching the recordings.

(2) After step (1), we examined the ratings and selected the scenario which had the least differences between the four raters' scores. Based on the rating scores of this scenario, the four raters then discussed those CRM Skills items which they had scored differently. This discussion was a trial to allow the raters to compare their scoring rationales and to decide a "True Score" for each Skill item on which they all agreed.

(3) After step (2), the raters again watched the recording of the scenario with the greatest rating differences, and again completed the two types of rating sheets to verify the validity of the True Score discussions.



a. *Scenario 1* *b. Scenario 2*

c. Scenario 3

Figure 3 *Scenario Structure of simulated LOFT sessions*

Results

Validity of the Proposed CRM Skills Rating Sheet for Each Threat

From step (2) of the experiment procedure, the difference between raters' scores was found to be greatest for scenario 2, and so scenario 2 was scored again after the raters had discussed their rationales. Figure 4 shows average values of standard deviation (SD) of scores of all CRM Skills items for each rater and case. Cases 1, 2, 3 and 4 in Fig. 4 indicate respectively scenarios 1, 2, 3 and the results of the second rating of scenario 2.

It is apparent from the figure that all average SD values are lower for the per threat CRM Skills rating sheet than the per flight phase rating sheet. The unpaired *t* test showed that this difference was statistically significant for cases 1 (P=0.037<0.05) and 3 (P<0.000), but not for cases 2 (P=0.098>0.05) and 4 (P=0.114>0.05).

Raters' comments indicate that the main advantages and disadvantages of each type of CRM Skills rating sheet are as follows:

(1) Per flight phase rating

- When the same CRM Skill is observed several times in the same flight phase, the scores for each instance of the Skill could potentially cancel out (nullify) each other. For example, if a rater observes both effective and ineffective behaviors for the "Monitoring" skills during the same flight phase, it is possible that his score will be the average (balance) of these behaviors, and the effective and ineffective behaviors might not appear in the final analysis.
- The "per flight phase" approach allows an overall evaluation of a crew's skills. On the other hand, the "per threat" approach is limited to evaluating crew behaviors when managing (or mismanaging) threats and errors and does not give an overall evaluation.

(2) Per threat rating

- It is possible to rate in detail.
- Timing of evaluation is sometimes difficult because some threats are persistent. For example, in the case of a passenger being taken ill, the rater might be confused as to precisely when to evaluate the crew's management behavior (when the threat first appears, or at some point later) because the crew may continue to address the threat at a later point in time.
- With the per threat rating sheet, it can be difficult to score CRM skills that a crew exercises or fails to exercise because the skill may not be related to any threat on the sheet.

Validity of Discussions to Introduce "True Score"

In step (2) of the experiment procedure, the raters discussed their scoring of cases 1 and 3 in order to derive a "True Score" for each CRM skill item. After this discussion, the raters again observed and rated scenario 2, and the result was analyzed as case 4. The differences between cases 2 and 4 are therefore due to the discussion between the raters and the raters becoming familiar with the scenario, since both cases used the same LOFT scenario.

As is apparent from Fig. 4, the values of variance for case 4 are greater than for case 2 for both types of rating sheet. However, the paired *t* test reveals that these differences are not statistically significant for either per threat rating (*P*=0.399>0.05) or per flight phase rating (*P*=0.382>0.05). It is possible, though, that there are concrete differences between cases 2 and 4 that cannot be identified from only the average value of SD. To explain the changes from case 2 to case 4, Figure 5 shows the proportions of each score for these cases. It is clear from the figure that case 2 has the greater proportion of "3 point" and "blank" scores for each item. These results indicate that raters' scores tend to be biased towards middle values. In case 4, the proportion of "3 point" scores decreases, but the proportions of "2 point" and "1 point" scores increase, indicating that scoring tendencies change from a "Central tendency" (Baker *et al.*, 1999) to a more varied evaluation.

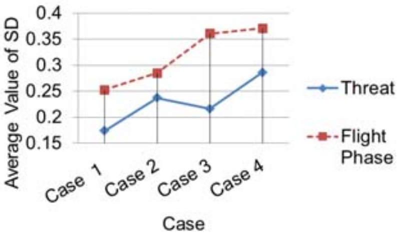
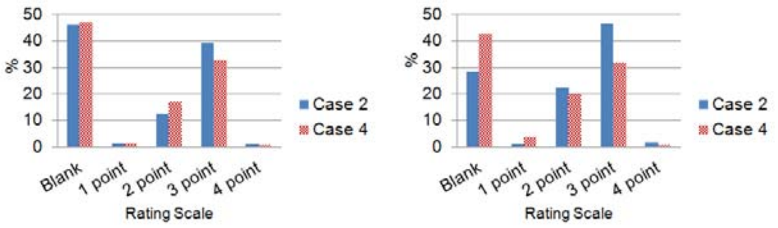


Fig. 4 *Average SD for each case*



a. *Per Threat Rating Sheet* b. *Per Flight Phase Rating Sheet*

Fig. 5 *Percent of Rating for Case 2 vs. Case 4(twice of scenario 2)*

Effect of Included Threat Codes in Simulated LOFT Scenarios

To investigate whether the variance of ratings is affected by the threat types (threat codes) included in the scenarios, Table 2 shows the correspondence between these two quantities. As shown in the table, the average values for the four raters are greatest for scenario 3 (case 3), which includes equal numbers of all threat types. However, we cannot say that it is generally true that scenarios that have equal numbers of each threat code will give greater average scores, because we did not compare scores with another scenario using the same crew who carried out scenario 3.

It is considered that the timing of insertion of a threat into a scenario is important as well as the type of threat. Both scenarios 1 and 2 include three Aircraft threats. While Aircraft threats appear continuously in scenario 2 as shown in Fig. 3, scenario 1 has another type of threat inserted between two Aircraft threats, so Aircraft threats do not appear continuously. The average rating values were lower and variance values were greater in scenario 2, in which the Aircraft threats were presented continuously.

Table 2 *Threat Code (Threat Type) vs. Average Value of Rating and SD*

Threat Code Case	Aircraft	Departure/Arrival	Cabin	Operational	ATC	Average Value of Rating		SD
Case 1	3	2	1	0	0	Phase[*1]	3.17	0.467
						Threat[*2]	3.12	0.432
Case 2	3	1	1	1	0	Phase	2.69	0.540
						Threat	2.71	0.544
Case 3	1	2	1	1	1	Phase	3.39	0.490
						Threat	3.26	0.441
Case 4	3	1	1	1	0	Phase	2.52	0.652
						Threat	2.63	0.570

Phase: the per flight phase rating sheet, Threat: the per threat rating sheet

Discussion

The final objective of the CRM Skills measurement method is to answer two questions: What CRM Skills does a crew lack corresponding to each type of threat and error, and what training will the crew require to adequately manage threats and errors. For developing such a CRM Skills measurement method, attention needs to be focused on three issues: The CRM Skills rating sheet as a tractable tool for CRM Skills data acquisition; inter-rater reliability training to ensure that reliable data are obtained; and the content of LOFT scenarios. We discuss these issues using the results obtained from the experiment.

Validity of the Proposed Per Threat CRM Skills Rating Sheet

Judging from the results of average values of SD of rating for cases 1 (scenario 1) and 3 (scenario 3), the proposed per threat CRM Skills rating sheet assisted the raters in having a consistent viewpoint by assessing crew behaviors when managing or mismanaging threats / errors. However, individual differences between raters still remained in the results of cases 2 and 4 (scenario 2). Rater comments, mentioned in the Results section above, indicate that the variances of rating for the per flight phase rating sheet were greater than those for the per threat rating sheet because the scoring decisions were different. If raters observed both "effective behavior" and "ineffective behavior" during the same flight phase, one rater might score an average between these behaviors while another rater might score based only one of the behaviors. On the other hand, one rater's comments gave insight as to why individual rater differences existed even for the per threat rating sheet: the timing of when to evaluate a crew's threat management behavior is difficult because some threats persist for some time after appearing. Another reason identified is that each rater may focus on different CRM Skills items; e.g. Rater A might score 2 points for the Assertion skill for a given threat, while Rater B might not focus on Assertion and score blank for this skill but instead score 2 points for Leadership for the same threat. This is related to the issue of whether or not it is necessary to standardize the CRM Skills items on which raters should focus.

This study's analysis was based on the average value and variance of ratings. However, it is considered that analysis should not only use these numerical values but should take into account the raters' individual judgments of the relative importance of each CRM Skills item. The validity of the CRM Skills rating sheet will then be verified and inter-rater reliability training will be conducted by considering scores weighted by raters' judgment of importance. For example, if most raters assign a low score for a "Threat X", even if they focus on different CRM Skills items, the analyst should be feed back that the crew's management of "Threat X" is very weak, but might then indicate those CRM Skills on which each rater focused as points for improvement e.g. "Assertion", "Leadership". A CRM Skills rating sheet that uses a "weighted value" of ratings is likely to be discussed in future.

Inter-Rater Reliability Training

As the result of the discussions to introduce "True Scores", raters' scoring tendencies changed from making "ambiguous ratings" to "clear ratings" based on definite judgments; for example, before the discussions some raters hesitated before finally scoring 3 points for an item, but after the discussions their scoring changed from this central tendency to scoring clearly 2 points or 4 points. This is obvious from Fig. 5 and Table 3. As is apparent from Table 3, the proportions of 3-point scores by raters A and B were markedly lower in case 4 (re-rating scenario 2 after discussion). Rater D, who had check airman experience, commented that although it was difficult in case 2 to assign scores of below 3 points (when rating scenario 2 before discussions), in case 4 he was able to assign scores of below 3 points not from the viewpoint of pass or fail, but considering the need for retraining.

This change of rater D's scoring tendency is revealed from the increased proportion of 1-point scores in Table 3. Rater A's "central" scoring tendency, by which he tended to score average values, was also improved by the discussion as mentioned in the Results section. Rater A commented although he scored 3 points even for crew behavior which he could not observe in case 2, his rating method changed clearly in case 4 in that crew behaviors which he was not able to observe were scored blank. These findings show that the discussions between raters to introduce "True Scores" contribute to their changing interpretations of the rating scale and avoiding the "central tendency".

The duration of the discussions, only three hours, was too short to achieve standardization of the raters. Although it was insufficient to achieve totally consistent scoring by the raters, however, some viewpoints such as the interpretation of the rating scale and examples of crew behaviors corresponding to each grade (1-point, 2-point, 3-point and 4-point) could be standardized. The trial discussions therefore helped to familiarize raters with the scoring method, and it is supposed that actual inter-rater reliability training will be conducted in future by repeating the scoring and discussion between raters. It is considered that by such training each rater will understand the rating errors that are easy to commit and be familiarized with the scoring method, and then agreement on the interpretation of the rating scale and CRM Skills items will be performed through an iterative process of scoring and discussions. However, we could not draw any conclusions as to the number of iterations that will be required.

Table 3 Change of Rating from Case 2 to Case 4 *(per flight phase rating sheet)*

		Rater A	Rater B	Rater C	Rater D
Average	2[*1]	2.77	2.61	2.67	2.67
	4[*2]	2.71	2.42	2.53	2.42
SD	2	0.505	0.495	0.586	0.595
	4	0.579	0.620	0.567	0.807
Number of 1 point	2	2	0	0	0
	4	2	1	1	5
Number of 2 point	2	8	15	14	13
	4	6	17	13	9
Number of 3 point	2	43	23	20	18
	4	26	12	18	16
Number of 4 point	2	0	0	2	2
	4	0	1	0	1
Number of Blank	2	3	18	20	23
	4	22	25	24	25

2: Case 2 (rated in scenario 2 before the discussion), 4: Case 4 (rated in scenario 2 after the discussion)

Development of LOFT Scenarios for Threat and Error Management Training

As mentioned in the Results section, it is possible that the variance of rating is affected by the type and timing of threats that appear in scenarios. Variance of rating was greatest for scenario 2, in which a technical event (an Aircraft threat) appeared continuously. Scenario 2 appears to have been aimed at operating procedures training, and it is possible that execution of CRM Skills was hardly observed for this scenario since crew behavior in technical events requires technical skills to execute a prescribed procedure rather than CRM skills in general. Some raters evaluated the fact that captain made decisions by himself without communication with other crew members, because some SOPs does not require much discussion between crew members since the procedures are clearly specified , while other raters evaluated the crew’s behavior based on only the captain’s “Leadership”. It is considered that such issues contributed to the higher variance of rating in scenario 2 than other scenarios.

If the purpose of a LOFT scenario is not procedures training but exercising CRM Skills for TEM, it is necessary to carefully investigate which types of threat are appropriate to be included in the scenario and their timing. A scenario which generates a large variance and low average value of rating between raters is considered inappropriate as a CRM Skills training scenario.

Conclusion

A CRM Skills Measurement Method which includes a Threat and Error Management concept was proposed, and its validity was verified by a CRM Skills measurement experiment.

The results of the experiment showed that the proposed per threat CRM Skills Rating sheet assisted raters to have consistent viewpoint by assessing crew behavior when managing or mismanaging threats / errors, but individual differences between raters still remained. Additionally, discussion between the raters to introduce “True Scores” prompted them to clarify their rationale for scoring. Analysis of the results indicates that factors contributing to individual scoring differences include not only the contents of the CRM Skills Rating sheets and the inter-rater reliability training method, but also the threat types included in simulated LOFT scenarios and the timing of their appearance.

References

Baker D. P., Mulqueen C. and Dismukes R. K. (1999). Training pilot instructors to assess CRM: The utility of frame-of-reference (FOR) training, International Aviation Training Symposium.

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

Klinect, J. R., Wilhelm, J. A. and Helmrech, R. L. (2001). LOSA Error Code Book 9.0, Proc. of 1st LOSA week.

Noda F., Iijima T., Tsuda H., Sudo K., Yamamori H. and Kobayashi H. (2005). Assessment of CRM Skill Indicators, 44th Aircraft Symposium, No. 3G11.

Tsuda H., Iijima T. and Noda F. (2006). Development of CRM Skills Measuring Method, ICAS2006.