

Measuring Safety in Aviation: Empirical Results about the Relation between Safety Outcomes and Safety Management System Processes, Operational Activities and Demographic Data

Steffen Kaspers, Nektarios Karanikas, Selma Piric,
Robbert van Aalst, Robert Jan de Boer
Aviation Academy
Amsterdam University of Applied Sciences
Amsterdam, The Netherlands
Email: s.e.kaspers@hva.nl, n.karanikas@hva.nl,
s.piric@hva.nl, r.j.aalst@hva.nl, rj.de.boer@hva.nl

Alfred Roelen
Safety Institute
Netherlands Aerospace Centre
Amsterdam, The Netherlands
Email: alfred.roelen@nlr.nl

Abstract— A literature review conducted as part of a research project named “Measuring Safety in Aviation – Developing Metrics for Safety Management Systems” revealed several challenges regarding the safety metrics used in aviation. One of the conclusions was that there is limited empirical evidence about the relationship between Safety Management System (SMS) processes and safety outcomes. In order to explore such a relationship, respective data from 7 European airlines was analyzed to explore whether there is a monotonic relation between safety outcome metrics and SMS processes, operational activity and demographic data widely used by the industry. Few, diverse, and occasionally contradictory associations were found, indicating that (1) there is a limited value of linear thinking followed by the industry, i.e., “the more you do with an SMS the higher the safety performance”, (2) the diversity in SMS implementation across companies renders the sole use of output metrics not sufficient for assessing the impact of SMS processes on safety levels, and (3) only flight hours seem as a valid denominator in safety performance indicators. At the next phase of the research project, we are going to explore what alternative metrics can reflect SMS/safety processes and safety performance in a more valid manner.

Keywords - Safety Metrics; Safety Management Systems; Safety Performance; Safety Outcomes.

I. INTRODUCTION

This paper presents part of an on-going 4-year research project “Measuring Safety in Aviation – Developing Metrics for Safety Management Systems” [1] executed by the Amsterdam University of Applied Sciences and co-funded by the Nationaal Regieorgaan Praktijkgericht Onderzoek SIA [2]. A literature review we conducted in 2016 identified several challenges concerning the measurement of safety [3]. Between February and June 2016, surveys were executed to explore (1) what, how and why, safety metrics are used and (2) whether a monotonic relation between Safety Management System (SMS) process metrics and safety outcomes could be established.

Safety outcomes are defined as accidents, (serious) incidents, occurrences and other safety related events [3]. SMS process metrics include indicators on safety staff,

improvements, training, communication, hazard identification, risk management and emergency response [3]. A full listing is given in appendix 2. The SMS process metrics can be applied at a system level but are usually more informative at the sub-system (department, activity type etc.) level. Safety outcomes on the other hand are emergent indicators representative of the whole system.

The results from the first part of the surveys were presented in the International Cross-industry Safety Conference 2016 and published in the proceedings [4], thus this paper focusses on the 2nd part of the surveys. The relation between safety related processes and outcomes can be claimed through two channels: empirical evidence or credible reasoning [5]. Since respective empirical evidence is scarce [6], we aimed at finding associations between SMS process and safety outcome metrics by using data collected from the partners of the project.

In section 2, the research problem and the hypotheses are introduced. Section 3 describes the methodology used and the results are presented in section 4. In section 5, the results are discussed. The paper finishes with the conclusion in section 6.

II. RESEARCH PROBLEM

In the literature review [3], it was concluded that the reasoning behind the relationship between SMS processes and safety performance lies principally on linear safety/accident models, where a direct cause-effect relation between safety management activities and safety events is implied. Thus, the relationship between SMS/safety processes and outcome metrics is seen as monotonic in practice under a “necessary but not sufficient” logic; a single failure or deviation from a SMS/safety process might not lead to an adverse outcome, but multiple failures (e.g., malfunctioning barriers) or deviations (e.g., incompliance with procedures) are likely to cause unwanted outcomes. Besides the linear accident models, few systemic models have been introduced in literature [7][8] but they haven’t been extensively applied to the industry.

The aforesaid thinking and industry practice are translated into two hypotheses:

H1: There are consistent and similar monotonic relations of SMS process data with safety outcomes across all companies.

In order to judge whether there is a positive or negative effect of an SMS process on safety outcomes, the direction of the relationship, the scope and timeliness of the respective process must be considered. For example, in the cases of safety training and audits, a negative correlation is expected under the argument that more training or audits lead to fewer safety events. When considering other SMS processes, such as safety reporting and hazard identification, a positive correlation might be expected since those activities retrofit risk assessment with a goal to mitigate risks and improve safety performance; on the other hand, a negative correlation might also reflect that risk assessment does not succeed to increase safety performance, meaning to decrease adverse events.

H2: There are consistent and similar monotonic relations (i.e., regardless their positive or negative direction) of demographic and operational activity data with safety outcomes across all companies.

Correlations of operational activity or/and demographic data with safety outcomes (1) over time for each company and (2) across the whole sample when considering respective averages per company, indicate validity of the respective indicators used in the industry (e.g., accidents per passenger miles or flights).

III. METHODOLOGY

Thirteen companies who participated in the project were asked to provide data in the form of a data-sheet. The request was based on the types of metrics identified through the literature [3] and represented in appendices 1 and 2. The data sheet included 5 operational activity figures (e.g., departures), 12 demographic data fields (e.g., number of staff,), safety outcomes (e.g., number of occurrences of various severities) and 38 fields covering output and frequency of SMS activities (e.g., number of hazards identified, amount of SMS documentation updates) from up to 10 years. Specific instructions were not provided since the fields correspond to data that organizations are familiar with, but some clarifications were offered upon request from the partners.

Ten companies provided the data requested within the time frame set. Most of the large companies reported that they

TABLE 1: RESEARCH SAMPLE

	Size		Domain			
	Large (N=7)	Small (N=6)	Flight Ops (N=7)	ATC (N=2)	GS (N=1)	MRO (N=3)
Data-sheets	2	3	4	1		
Dashboards	2		1	1		
Insufficient data	1	2	1		1	1

needed considerable time and resources for retrieving the data from several databases since such data were not always directly linked to safety performance and maintained by the safety department. Two large companies sent their annual safety dashboards and the research team converted that data to the respective fields of the datasheet. Due to a recent implementation of a SMS in three out of the ten companies, the sheets received did not include enough data points for statistical analysis. Consequently, data sets from seven companies were used for statistical tests (Table 1).

After the collection of data, raw figures were converted to ratios in order to use comparable figures across years for each company; this resulted in an extensive list of measures. The researchers tested all available pairs (i.e., Operational Activities – Outcomes, Demographics – Outcomes and SMS processes – Outcomes) as a means to examine all relationships. Because of the limited sample size, all data were tested with non-parametric correlations. Spearman's coefficient was chosen to explore any monotonic relations of SMS/operational/demographic metrics with safety outcome ones. Spearman's coefficient indicates the presence of a monotonic relationship and not the strength of linear associations. The statistical significance was set to $p=0.05$.

IV. RESULTS

Table 2 shows the number of pairs tested for monotonic relations. The table is divided into three sections corresponding to operational activities, demographics and SMS processes tested for associations with safety outcomes.

TABLE 2: VALID PAIRS TESTED FOR MONOTONIC RELATIONS

Company	Operational Activities - Outcomes		Demographics - Outcomes		SMS - Outcomes	
	Valid pairs	Significant correlations	Valid pairs	Significant correlations	Valid pairs	Significant correlations
1	4	0, (0%)	0	0, (0%)	25	0, (0%)
2	30	6, (20%)	57	7, (12.3%)	165	19, (11.5%)
3	3	0, (0%)	0	0, (0%)	12	5, (41.7%)
4	36	10, (27.8%)	0	0, (0%)	116	27, (23.3%)
5	232	0, (0%)	188	6, (3.2%)	1292	82, (6.3%)
6	62	8, (12.9%)	48	20, (41.7%)	380	42, (11.1%)
7	72	57, (79.2%)	12	8, (66.7%)	12	8, (66.7%)
Total	439	81 (18.5%)	305	41 (13.4%)	2002	183 (9.1%)

TABLE 3: CORRELATION OF AVERAGES OF ACTIVITY/DEMOGRAPHIC DATA WITH SAFETY OUTCOMES

Demographic and Operational Activity Figures (Averages of Companies)	Safety outcomes			
	Serious Incidents	Incidents	Occurrences	All Events
Flight Hours	$r_s(6)=0.845$ p = .034		$r_s(5)=0.900$ p = .037	$r_s(6)=0.943$ p = .005
Full Time Equivalent of Contractors		$r_s(4)=-1.000$ p = .000		
Flight Hours per Pilot			$r_s(3)=1.000$ p = .000	$r_s(3)=1.000$ p = .000

Within each section, the number of valid pairs are mentioned and the significant correlations for those pairs of data [number, (percentage)].

Appendix 3 includes a sample of cases where significant correlations within companies were found; the whole set of results were published in a technical report [9]. The cells in the corresponding tables include the direction of each correlation (i.e., POS: Positive and NEG: negative) and the number of companies for which the data permitted the conduction of valid correlations per case (i.e., sample N). The cells where POS or NEG are followed by a number (i.e., x Number) indicate how many companies had the respective significant correlation.

In addition to the results within companies, Table 3 shows the significant correlations of the averages of safety outcomes of all severities with activity (e.g., departures) and demographic data (e.g., full time equivalent of company staff.). Tests for miles flown were not feasible due to limited data. Through those correlations, we explored the validity of using demographic or operational activity data as denominators of ratios of adverse safety events, since such ratios are used by the industry to measure safety performance.

The findings presented in Table 3 showed that:

- Increased flight hours' activity is associated with more occurrences, serious incidents and safety events in general.
- The more FTEs are spent by contractors, meaning the more the outsourcing of company activities, the fewer the incidents recorded by the company.
- The more the flight hours' load per pilot, the more the occurrences and events in general.

Taking into account that the flight hours was the main variable associated with some types of safety outcomes, we conducted further statistical tests as follows (table 4):

- Mann – Whitney test was used as a means to explore if the ratios of each event type by flight hours differ between large companies and SMEs. The calculations did not show significant differences.

TABLE 4: DIFFERENCES BETWEEN AND WITHIN LARGE COMPANIES AND SMES.

Event type / flight hours	Mann – Whitney test between large companies and SME	Kolmogorov – Smirnov tests between SMEs
Accident	p=0.690	p=0.001
Serious Incident	p=0.143	
Incident	p=0.095	p=0.049
Occurrence	p=0.800	
All events combined	p=0.133	

- Kolmogorov - Smirnov tests were conducted for the ratios of each event type by flight hours for SMEs; the sample size did not allow the conduction of those tests for large companies. The results showed significant differences between SMEs regarding accidents and incidents per flight hours.

V. DISCUSSION

According to the results, the following observations can be made:

1. The significant correlations regard only part of the SMS processes and safety outcomes and a small portion of the sample, and the distribution of associations is highly scattered. No strong evidence was found that the output and frequency of all SMS processes had an effect on safety outcomes; significant associations were found only for few of the participant companies.

2. The results suggest that just the operation of an SMS does not guarantee an effect on safety outcomes; therefore, that other factors, such as the quality of SMS processes, might play an important role. Also, an evaluation of the effectiveness of an SMS against high severity events seems unjustified in the frame of this survey. More specifically:

a. Most of the significant correlations were found for occurrences (i.e., the lowest severity category of safety events) as well as all safety outcomes regardless their severity.

b. Accidents, serious incidents and incidents and their ratios by activity and demographic figures were associated with a very few SMS processes.

c. Only at a few companies the outputs and frequency of SMS processes had a visible effect on low severity events, the latter reflecting safety performance at shorter intervals.

3. There were 33 negative and 124 positive correlations between SMS process and safety outcomes. However, in 59 cases of all correlations the data regard a single company that provided respective data, so the results cannot be deemed as representative of the whole sample. Nevertheless:

a. The negative correlations sporadically regarded numbers or ratios related to staffing of the safety department, internal audits, safety training, safety surveys and hazard identification. Although due to the limited sample those associations cannot be generalized, the aforementioned areas of SMS processes were influential on safety outcomes of low severity mostly for a single company. It is noticed that a negative correlation between SMS processes and safety outcomes can be considered as a positive case only when outcomes decrease over time; in case that, under a negative

correlation, events increase over time, the SMS can be contemplated as insufficient.

b. Most of the positive correlations were found for the safety reporting and risk assessment processes, the interpretation of those associations being dependable on the timeliness of those processes. The SMS activities are performed continuously, so a distinction between a “positive reactiveness” (e.g., more risk assessments occur due to more outcomes) and “negative proactiveness” (e.g., more risk assessments lead to an increase of adverse events) is not directly evident. Contextual information is of paramount importance in order to interpret such results correctly.

Observation No 1 suggest that hypotheses H1 is partially rejected due to the limitations imposed by the sample size. Additionally, the diverse ways that SMS processes are implemented across the industry and over time, and the different interpretations of outcome thresholds [9] might have affected the results and the validity of comparisons within and amongst companies.

A. Correlations between operational activities and safety outcomes

The results presented in Appendix 3 do not suggest a consistent picture within companies. Some activity data related to departures, miles flown and flight hours were associated with all safety events, incidents and serious incidents, but in the majority of the cases those findings regarded only one company out of the whole sample. Only in seven cases the associations of flight hours related data with some types of safety outcomes were found for two companies. Interestingly, accidents were not represented in the significant correlations with operational activities, although annual reports published by regional and international bodies use rates of accidents as a means to depict safety performance (e.g., [10]); perhaps, the large sample that such reports include might render the use of accident ratios meaningful, but the results of our survey showed that those ratios might not be representative of safety performance at the company level. The latter is also supported by the fact that we did not observe any association between operational activity data and number of accidents when considering averages across the sample (Table 3).

Furthermore, in the case of flight hours, the correlations with outcomes were found interchangeably positive or negative depending on the denominator and the company, whereas in few cases the same correlation was found negative for one company and positive for another. This observation might once more reflect the dissimilarities in the interpretation of safety outcome definitions, or/and the differences regarding the effectiveness of safety management in those companies; a positive correlation between activity and outcome data indicates that safety management is not improving (i.e., as safety management activities increase, safety outcomes increase too and vice versa), whereas a negative correlation signals that safety management performs either as expected (i.e., when outcomes decrease over time) or poorly (i.e., when outcomes increase over time).

Monotonic relations were found across the companies regarding flight hours and flight hours per pilot with safety

outcomes, the accidents excluded, thus suggesting that the specific type of operational activity might be a more valid exposure measurement than departures and miles flown. By nature, departures do not reflect the total load imposed to company staff (e.g., time that pilots fly or maintenance requirements based on the hours that aircraft operate), and miles flown are not also directly related to the total load due to a variety of factors such as aircraft capabilities, flight plans and fuel efficiency policies (e.g., the same distance might be covered in shorter or longer time based on the air traffic and average flying speed). The findings of our study are aligned with [11], who showed a relation of task load expressed in total flight hours per employee with rates of events attributed to human error.

B. Correlations between demographics and safety outcomes

The picture is even more distorted regarding the relationship between demographic figures and safety outcomes. The correlations found were highly dependable on the denominators used in the safety outcomes; for example, the average aircraft age was positively correlated with number of occurrences and the ratio of occurrences by flight hours, but negatively correlated with the ratios of occurrences by miles flown and departures. Hence, under the expectation that the higher the age of the aircraft, the more the occurrences under increasingly complex operations, it seems that, in this case too, flight hours can act as a more representative denominator compared to miles flown and departures.

Furthermore, the number of company employees was positively correlated with occurrences, but negatively associated with incidents and all safety events regardless severity. Although those differences do not refer to the same company, they suggest that the use of raw demographic data alone cannot render respective indicators valid. In conjunction with the discussion of the results of the paragraph above, ratios of activity figures, and especially flight hours, by demographic data can be more valid representations of risk exposure in comparison with net numbers of operational activities or demographics.

Taking into account the overall picture and the limitation imposed by the sample size, the researchers claim that the hypothesis H2 is partially rejected. As in the discussion of the hypothesis H1, the different interpretations of outcome thresholds might have affected the results.

VI. CONCLUSIONS

From the numerical analysis of the data sample, consistent correlations between operational activity figures, demographic data, SMS process data and safety outcomes could not be established. The correlations in the sample demonstrated a wide variety, and there were no correlations supported by all usable datasets. Only part of the datasets resulted to significant correlations for specific combinations of data, and in some cases, there were both positive and negative correlations for the same pair of variables in the sample.

Due to the limited sample size (i.e., number of participating companies and data points per company), we do

not claim external validity of the results and we could not fully reject the research hypotheses. However, since the latter cannot be fully confirmed, the current practices in safety performance measurement seem of limited validity. The partial rejection of hypotheses H1 and H2 is aligned with, and indirectly validated by, the concerns of the companies about the existing safety metrics and their needs for better / alternative ones [9]. Nevertheless, the diverse and, occasionally, contradictory findings from the quantitative analysis might be attributed to the (1) different interpretations of thresholds of safety outcomes, (2) implementation of SMS processes in various ways, due to which the data points of this study reflected different contexts of the companies and changes over time, and (3) limited value of the linear approach to safety, as suggested by the models widely used by the industry and the emergent behavior at the system level that constitutes safety. This latter consideration is exuberated by the mismatch of indicators at the sub-system (department, activity process) and system level.

In overall, the findings of this study indicate the need to move towards the development of metrics that will be more representative of SMS processes and safety outcomes and will allow valid comparisons over time and across the industry. Based on the results of this research phase, the justification of the overall project does not only stem from a need to improve scientific knowledge on the topic of aviation safety metrics, but it is also jointly supported by the concerns and needs of the industry and the findings of the analysis of numerical data collected in this research phase.

ACKNOWLEDGMENT

The research team would like to expresses their deep thanks to:

The companies which participated in the surveys: Helicentre, KLM (Royal Dutch Airlines), KLM Cityhopper, Life Line Aviation, LNVL (Air Traffic Control the Netherlands), MUAC (Maastricht Upper Area Control Centre), Olympus Airways, SAMCO, Sky Service Netherlands, Transavia.

The members of the knowledge experts group of the project, who reviewed and provided enlightening and valuable feedback (in alphabetical ascending order of partner organization): CAA NZ: C. Mills, EASA: J. Franklin, Kindynos: J. Stoop, KLM Cityhopper: E. Hiltermann, Klu /

MLA: R. Van Maurik, Purdue University: J. Keller, Team HF: G. Hofinger, TU Delft: A. Sharpanskykh.

REFERENCES

- [1] Aviation Academy. "Project Plan RAAK PRO: Measuring safety in aviation – developing metrics for Safety Management Systems", Hogeschool van Amsterdam, Aviation Academy, The Netherlands, 2014.
- [2] SIA. "Decision on request grant scheme RAAK PRO 2014 for the project Measuring Safety in Aviation - Developing Metrics for Safety Management Systems "(project number: 2014-01-11ePRO)" ("Besluit inzake aanvraag subsidie regeling RAAK-PRO 2014 voor het project Measuring Safety in Aviation – Developing Metrics for Safety Management Systems '(projectnummer:2014-01-11ePRO)"). Feature: 2015-456, Nationaal Regieorgaan Praktijkgericht Onderzoek SIA. The Netherlands, 2015.
- [3] S.E. Kaspers, N. Karanikas, A.L.C. Roelen, S. Piric, and R.J. de Boer, "Review of Existing Aviation Safety Metrics, RAAK PRO Project: Measuring Safety in Aviation," Project S10931, Aviation Academy, Amsterdam University of Applied Sciences, the Netherlands, 2016.
- [4] S.E. Kaspers et al., "Exploring the Diversity in Safety Measurement Practices: Empirical Results from Aviation". The First International Cross-industry Safety Conference, Amsterdam, 3-4 November 2016, Journal of Safety Studies, 2(2), pp. 18-29, 2016. DOI: 10.5296/jss.v2i2.10437
- [5] J. Wreathall, "Leading? Lagging? Whatever!". Safety Science, 47(4), 493 - 494. doi:10.1016/j.ssci.2008.07.031
- [6] T. Reiman, and E. Pietikäinen, "Leading indicators of system safety – monitoring and driving the organizational safety potentia l." Safety Science, 50(10), 1993- 2000, 2012. doi:10.1016/j.ssci.2011.07.015
- [7] E. Hollnagel, "FRAM, the functional resonance analysis method: modelling complex socio-technical systems," Ashgate Publishing, Ltd., 2012.
- [8] N. Leveson, "Engineering a safer world: Systems thinking applied to safety," Mit Press, 2011.
- [9] S.E. Kaspers, N. Karanikas, A.L.C. Roelen, S. Piric, and R.J. De Boer, "Results from Surveys about Existing Aviation Safety Metrics, RAAK PRO Project: Measuring Safety in Aviation," Project S10931, Aviation Academy, Amsterdam University of Applied Sciences, the Netherlands, 2016.
- [10] EASA, "Annual Safety Review." Cologne: European Aviation Safety Agency, 2016.
- [11] N. Karanikas, "Correlation of Changes in the Employment Costs and Average Task Load with Rates of Accidents Attributed to Human Error," Aviation Psychology and Applied Human Factors, 5(2), pp. 104-113, 2015.

APPENDIX 1 – SAFETY OUTCOME , ACTIVITY AND DEMOGRAPHIC METRICS

Activity data
Departures
Miles Flown
Flight Hours
Number of Company Staff
Ratio of Company Staff Turnover
Ground movements
Demographic data
Full Time Equivalent (Company)
Full Time Equivalent (Contractors)
Experience of Flight Crews (Flight Hours)
Hours Flown / Pilot
Experience of Ground Staff (Years)
Aircraft Fleet
Aircraft Age (Years)
Safety Outcomes
Number of All Safety Related Events
Number of Occurrences
Number of Incidents
Number of Serious Incidents
Number of Accidents

APPENDIX 2 – SMS PROCESS METRICS

SMS process - Safety Staff
Number of Safety Staff
Full Time Equivalent Safety Staff Spends on SMS
Number of Safety Staff Changed
SMS process - Improvements
SMS updates
SOPs, procedures, rules etc. updates
Number of External Audits
Findings from External Audits
Number of Internal Audits
Findings from Internal Audits
Number of Internal Safety Reviews / Meetings
Days for Implementing Decisions Internal Safety Reviews / Meetings
Number of Safety Meetings with External Organizations
Number of Safety Conferences, Workshops etc. Attended
Number of Safety Surveys
Ratio of Targeted Population Participated in Safety Surveys
Number of Safety Studies Accomplished (in addition to Safety Surveys)
SMS process - Safety Training & Education
Number of Safety Training Sessions Completed
Hours per Safety Training Session
Ratio of Staff Attending Safety Training
Ratio of Staff Passing Safety Training Exams on 1st Attempt
SMS process - Safety Communication
Number of Safety Bulletins, Notices etc.
Times of Safety Communication (each communication might include 1 or more safety messages, posters etc.)
SMS process - Hazard Identification
Number of Safety Reports Submitted by Company and Contractor Staff (e.g., Air Safety Reports, Hazard Reports)
Number of Safety Reports Followed-Up / Feedback Provided
Number of Hazards Identified from Sources Except Safety Reports (e.g., Safety Investigations, Safety Audits, Safety Observations)
SMS process - Safety Risk Assessment & Mitigation
Number of Total Risk Assessments Performed
Number of Risk Assessments Initially Rated as Low
Number of Risk Assessments Initially Rated as Medium
Number of Risk Assessments Initially Rated as High
Number of Risk Assessments Initially Rated as Unacceptable
Number of Low Risks in the Registry (after assessment & mitigation)
Number of Medium Risks in the Registry (after assessment & mitigation)
Number of High Risks in the Registry (after assessment & mitigation)
Days Between Hazard Identification and Risk Assessment
Days Between Risk Assessment & Implementation of Measures
SMS process - Emergency Response
Number of Emergency Response Exercises
Hours Spent on Each Emergency Response Exercise
Number of Emergency Response Planning Updates

APPENDIX 3