Amsterdam University of Applied Sciences

RAAK PRO Project: Measuring Safety in Aviation

Deliverable: Concept for the Design of New Metrics

May 2017

Nektarios Karanikas, Steffen Kaspers, Alfred Roelen, Selma Piric, Robbert van Aalst, Robert J. de Boer

Project number: S10931

RAAK PRO Project: Measuring Safety in Aviation

Concept for the Design of New Metrics

Nektarios Karanikas[1], Steffen Kaspers[1], Alfred Roelen[1,2], Selma Piric[1], Robbert van Aalst[1], Robert J. de Boer[1]

[1]Aviation Academy, Amsterdam University of Applied Sciences, the Netherlands
[2]NLR, Amsterdam, the Netherlands

Contents

Executive Summary

Following the completion of the 1st phase of the RAAK PRO project Aviation Safety Metrics, during which the researchers mapped the current practice in safety metrics and explored the validity of monotonic relationships of SMS, activity and demographic metrics with safety outcomes, this report presents the concept for the design of new metrics. Those metrics will be based on the hypothesis that the greater the gap between Work-As-Imagined and Work-As-Done the lower the safety performance, and they correspond to a set of references from academic literature, challenges in professional practice, depiction of system structure, and consideration of "soft" organizational aspects. Along with the design of the alternative metrics, this report explains the respective concepts referred in the literature but excluded from the current research, as well as the process and possible difficulties in ensuring various validity types of the new metrics.

1. Introduction

In September 2015, the Aviation Academy of the Amsterdam University of Applied Sciences initiated the research project entitled "Measuring Safety in Aviation – Developing Metrics for Safety Management Systems". The project responds to specific needs of the aviation industry: Small and Medium Enterprises (SME) lack large amounts of safety related data in order to measure and demonstrate their safety performance proactively; large companies might obtain abundant data, but they need safety metrics of better quality. Therefore, the aim of the project is to identify ways to measure safety in scientifically rigorous, meaningful and practical ways without the benefit of large amounts of data (Aviation Academy, 2014). The project will last until August 2019, is co-funded by the Nationaal Regieorgaan Praktijkgericht Onderzoek SIA (SIA, 2015), and is executed by a team of researchers from the Aviation Academy in collaboration with a consortium of industry, academia and research institutions, as well as representatives from authorities.

During the first phase of the research, which lasted from September 2015 to November 2016, the following were achieved:

- The current views and practices on safety metrics were identified by reviewing state-of-art academic literature, (aviation) industry practice, and documentation published by regulatory and international aviation bodies (Kaspers et al., 2016a). Appendix A presents the conclusions from this literature review.
- Surveys to aviation companies were conducted with the aim to explore the extent to which the findings from the aforementioned literature review were reflected in the practice of the partner companies. It was examined (1) what, how and why certain safety metrics are used, and (2) whether monotonic relations of SMS process, activity and demographic data with safety outcomes metrics are evident (Kaspers et al, 2016b, 2016c). Appendix B presents the conclusions from the surveys conducted in 2016.

The findings to date indicate the need to develop metrics that will be more representative of SMS/operational processes and safety outcomes and will allow valid comparisons over time and across the industry. Based on the results of the previous research phase, the justification of the current project does not only stem from a need to improve scientific knowledge on the topic of aviation safety metrics, but is also supported by the concerns and needs of the industry, as stated in the first paragraph of this section and confirmed during the previous research phase.

The goal of the current research phase is, therefore, to design alternative safety metrics which are suitable for SME's and fulfil the quality criteria referred in literature and collectively listed by Kaspers et al. (2016a). This phase will result in a set of new metrics which during the next phase of the project will be applied to aviation companies as a means to assess their association with safety outcomes. The rest of this document describes the approach of the research team with regard to the concepts that will be embedded in the design of the new metrics along with the hypotheses linked to those metrics (to be tested during the next phase), the concepts that the researchers will not consider in this research, and the process for ensuring the validity of the metrics.

2. Safety Concepts and Approaches

Both academia (e.g., Dekker, 2011; Leveson, 2015) and industry (e.g., ICAO, 2013; FAA, 2006) contemplate that safety performance is negatively affected by the gap between what must be done (e.g., regulations, standards, assumptions during design and intentions of system operation, procedures, check-lists) and what is actually done (i.e. practices on the work floor or the functional level under study). To date, external and internal audits, inspections, observations and techniques such as the Line Operations Safety Audits (LOSA) are used to depict the gaps between standards/established procedures and actual deliverables (i.e. tasks and results), but there has been little empirical evidence about the relationship of the findings of such activities with safety outcomes (Kaspers et al, 2016a, 2016b). The aforementioned authors attributed this lack of relationship to the following factors:

a. the ambiguity in the classification of safety outcomes;
b. the fact that the metrics currently used by the industry do not meet the quality criteria proposed in literature;
c. the prevalent safety views, which imply deterministic cause-effect relationships and do not embrace systemic perspectives suggesting the consideration of dependencies and interactions amongst system components;
d. the diversity in which SMS is implemented across companies and over time that does not allow benchmarking across and within companies;
e. the non-uniformity in the maintenance of data related to SMS processes because the former is not always linked with monitoring activities.

Taking into account the limitations of current safety metrics, the research team contemplates that the effects of the gap between Work-As-Imagined (WaI) (i.e. the activities/work as prescribed in procedures and rules) and Work-As-Done (WaD) (i.e. the actual activities and work delivered) on safety outcomes has not yet been sufficiently and evidently examined. Therefore, the overarching hypothesis to be tested is that the greater the gap between WaI and WaD, the lower the safety performance in terms of adverse outcomes [i.e. increased number of (serious) incidents and accidents]. It is noticed that although traditionally WaI is seen as the reference with which the WaD must comply, WaI might also include unsafe situations; the latter become visible when prescribed tasks are implemented, and WaD sometimes corrects these situations and actually improves safety performance (e.g., Boelhouwer, 2016). Therefore, the primary focus of the research team will be the distance between WaI and WaD, under the suggestion that if those get closer the changes can be induced to both or either of them.

In order to test the overarching hypothesis, the researchers have initially reviewed relevant literature to identify how the gap WaI-WaD can be depicted and quantified. The following sections describe the safety concepts and approaches that were reviewed, referring to the ones selected for operationalization along with the respective hypotheses (section 2.1) and the ones not to be addressed by the researchers (section 2.2). The main criterion for the inclusion/exclusion of concepts in the research was their potential to meet the project goals, meaning that the metrics must be practical (i.e. not requiring vast amount of resources and data), scalable (i.e. applicable to small-medium enterprises) and comprehensible (i.e. not requiring deep knowledge of the underlying theoretical foundations). In applying these criteria, the available research resources were taken into account. The degree to which the new metrics will meet the project goals and quality criteria will be reassessed during the reviewing process and pilot studies to companies (see section 3).

2.1 Concepts included

The concepts explained below along with the respective literature include topics that are currently suggested by academia (i.e. WaI-WaD at the work floor; Safety space), driven by gaps in professional practice (i.e. Performance-based SMS assessment; Effectiveness of risk controls), depictive of system structure (i.e. Resource scarcity; System complexity/coupling), and reflective of the “soft” side of organizations (i.e. Safety culture). The corresponding hypotheses to be tested at the next research phase are also stated in order to provide to the reader an overall picture of the direction of the project.

2.1.1 WaI-WaD at the operational process level

Based on the generic concept discussed above about the distance between WaI and WaD, the aim is to create more insights in the differences between how work should be performed according to the rules and what people on the work floor actually do to achieve desired results. A system-theoretic model of a relatively constrained aviation process (e.g., turn-around, stabilized approach) will be constructed according to the operating procedures, as well as for the process as executed in practice.

Central to the method is the identification of control signals and feedback flowing between actors in the system. Each signal can be judged to be apt (= 1) or flawed (e.g., missing, delayed, distorted, etc.; = 0), but intermediate scales might be also considered. The resulting vectors will determine the "distance" of the two instantiations of the same process. The gap between documented procedures and how tasks are performed in reality will be quantified in a "distance vector", as, for example, applied by Chatzimichailidou, Karanikas and Dokas (2016). The hypothesis for the next research phase is:

H-1: The larger the gap between WaI and WaD at the operational process level, the lower the safety performance of that process

2.1.2 Safety Space

Rasmussen (1997) argued that economy, workload and safety constitute the principal constraints of complex systems, a concept that has been embraced by the industry under the term "safety space" (e.g., ICAO, 2013). However, according to the literature reviewed by Karanikas (2015b), there is limited empirical evidence about the relationship amongst the aforementioned constraints. The aforesaid author identified a relationship between wage fluctuations (i.e. aspect of economy), rates of safety events attributed to human error (i.e. aspect of safety) and task load (i.e. aspect of workload) within an organization, but no further research is known to the research team.

The researchers will explore what indicators can reflect the three constraints proposed by Rasmussen (1997) and encompassing the Efficiency-Thoroughness Trade-Offs (ETTO) principle (Hollnagel, 2009). The latter suggests that people during their tasks usually make a choice between being effective and being thorough, since it is rarely possible to be both at the same time. The ETTO principle is connected with the resources employees have to execute their activities and produce the desired outcome (Hollnagel, 2009); excluding the human and temporal element from the type of resources since those are included in the workload dimension suggested by Rasmussen (1997), the material resources remain to be depicted. Regarding the workload measurement, the research team will explore the use various tools referred in literature (e.g., NASA, 1986; Roscoe, 1984; Cooper-Harper, 1969; Reid, Potter & Bressler, 1989). The hypotheses to be tested after the design of the metric(s) are:

H-2: The lower the value of material assets per service/product output, the lower the safety performance

H-3: The higher the workload, the lower the safety performance

H-4: The higher the workload, the stronger the positive relationship between value of material assets per service/product output and safety performance (i.e. workload as moderating variable)

H-5: The lower the value of material assets per service/product output, the higher the workload, leading to a lower safety performance (i.e. workload as mediating variable).

2.1.3 Performance-based evaluation of SMS

The industry has recognised the need to move from a compliance-driven assessment of SMS to a performance-based evaluation scheme (ICAO, 2013; EASA, 2014). Tools such as the Safety Management System Evaluation Tool developed by the Safety Management International Collaboration Group (SMICG, 2012), and the Effectiveness of Safety Management (EoSM) instrument, which was devised by the Eurocontrol (2012), have been introduced to support the aforesaid transition but they include subjective measurement scales, still emphasize on SMS implementation under the typical quality management cycle Plan-Do-Check-Act, and do not address the connections and dependencies of SMS processes (Karanikas, 2016). Therefore, although such tools introduce the transition from merely

checking the existence of SMS elements and processes to considering the sufficiency of their output and indicate necessary improvements, the interlinks between SMS activities are not yet addressed.

The Aviation Academy is developing an SMS evaluation tool based on the System Theoretic Process Analysis (STPA) technique (Leveson, 2011) with the goal to address both compliance (i.e. existence/operation of SMS components) and dependencies (i.e. necessary connections between SMS components). In this approach, we evaluate the design of the SMS from a control loop perspective, identifying whether the control hierarchy and the control actions are capable of ensuring that safety is effectively managed, meaning that it functions adequately. The tool will include specific assessment topics in order to guide auditors, and a scoring method which encompasses compliance to standards (i.e. documentation and implementation of SMS processes) and connectivity (i.e. dependencies amongst SMS elements). In a further step the actual implementation of the SMS can also be evaluated with similar means, by revealing causal factors/scenarios (i.e. why the system does not function according to the standards), so to support organizations in bringing SMS performance to the desired level. The hypothesis for the next research phase will be:

H-6: The lower the SMS score in terms of functionality, the lower the safety performance

2.1.4 Effectiveness of risk controls

ICAO (2013) encompasses the traditional risk management cycle and refers to the effectiveness of risk control measures. Currently, in the frame of a proactive safety management, the expected effects of risk controls on safety outcomes are subject to expert judgment, which is subject to biases and randomness. The researchers will explore whether properties of risk controls, such as their functionality (Hollnagel, 2004) and hierarchy (e.g., Leveson, 2011), combined with other parameters (e.g., frequency of failures of risk controls) can be blended in a metric for measuring their effectiveness. The hypothesis to be tested after the design of the metric(s) is:

H-7: The higher the effectiveness of risk controls, the higher the safety performance

2.1.5 Resource scarcity

The effects of resource scarcity on organizational drift was discussed by Dekker (2011) and it is connected with the ETTO concept presented by Hollnagel (2009), which was discussed above. The requirement for ensuring adequate resources for the implementation of the safety policy, is also stated by ICAO (2013). The researchers contemplate that the differences between the resources required by the task design, resources necessary according to the current task load and resources actually available can partially reflect the aforementioned concepts and also correspond to the WaI and WaD. The team will explore the design of a metric based on different types of resources (e.g., human, time, equipment, budget). The underlying hypotheses for the next research phase are:

H-8: The higher the positive distance between designed and available resources, the lower the safety performance

H-9: The higher the positive distance between necessary and designed resources, the lower the safety performance

H-10: The higher the positive distance between necessary and available resources, the lower the safety performance

2.1.6 System complexity and coupling

Modern systems become increasingly complex due to the interconnections and dependencies of system components, both human and technical ones. System complexity and coupling have been viewed as factors affecting safety performance, due to the limited ability to understand and control such systems and react to unforeseeable situations (Hollnagel, 2012; Leveson, 2011; Perrow, 1984). Considering that complexity is not just a reflection of the number of elements and their connections within a system, but also embraces the dynamic behaviours of such components and their interactions, the researchers will explore (1) how complexity/coupling can be measured for a system with given nodes and channels (i.e.

static factor) and (2) what demographic (e.g., work experience, age) and/or activity data (e.g., time available, nature of tasks) can be linked to the capacity/capability of system users to cope with complexity/coupling (i.e. dynamic factor).

The literature does not explicitly name the distinction between static and dynamic factors of complexity, but some indicative references for the operationalization and measurement of complexity/coupling distinguish between structural and dynamic aspects of complexity (e.g. Righi and Saurin, 2015; Frost and Mo, 2014; Rouse and Serban, 2011; Yadav and Khan, 2011; Schöttl and Lindemann, 2015; Butkiewicz, Madhyastha and Sekar, 2011; Simic and Babic, 2015). The hypotheses to be tested after designing the metrics are:

H-11: The higher the "static" complexity, the lower the safety performance

H-12: The tighter the coupling, the lower the safety performance

H-13: The lower the capability/capacity of systems to deal with "static" complexity, the stronger the negative relationship between "static" complexity and safety performance (i.e. capability/capacity as moderating variable)

H-14: The lower the capability/capacity of systems to deal with coupling, the stronger the negative relationship between coupling and safety performance (i.e. capability/capacity as moderating variable)

2.1.7 Safety Culture Development Plans

Safety culture has been for long a discussion topic in the academia and the industry, and, following also the expression of interest from the project partners, it comprises a topic of this research. Kaspers et al. (2016a) identified that there has been little consensus whether safety culture reflects the way an SMS is operated (i.e. as a safety performance metric), or the effects of SMS on safety performance (i.e. a safety outcome); at the same time safety culture is not consistently assessed within organizations (Kaspers et al., 2016b, 2016c).

Recently, a framework was suggested of the prerequisites which are necessary to develop a safety culture (Karanikas et al., 2016). This framework is based on academic literature and industry standards, and has since been developed into a tool (Grolleman, 2017). The objective of the tool is not to measure "safety culture" but to gain insights of what prerequisites (i.e. conditions) for building a positive culture are available and implemented within an organisation. The prerequisites are clustered in six categories following Reason's (1998) typology of safety culture (i.e. general organizational prerequisites and ones linked to the just, flexible, reporting, informative and learning sub-cultures). The hypotheses to be tested at the next research phase are:

H-15: The fewer the prerequisites planned, the lower the safety performance

H-16: The fewer the prerequisites implemented, the lower the safety performance

H-17: The less frequently the prerequisites are implemented, the lower the safety performance

2.2 Concepts excluded

Taking into account the goals of the project mentioned above (section 2) and the research resources available, the team contemplated that the following safety related concepts cannot be directly applied to this research, although they could be of interest for future studies:

- Views on human error and safety thinking: The Aviation Academy has developed a framework that includes the aspects of state-of-the-art human error views and safety thinking based, indicatively, on the work of Hollnagel (2004, 2014), Underwood & Waterson (2013), Leveson (2004, 2011), Catino (2008), Dekker (2006), and Rasmussen (1997). The framework has been applied to safety investigation reports of an organization (Karanikas et al., 2015a) and aviation authorities (Karanikas, 2015c), as a means to identify the extent to which investigators attempt to apply the respective

aspects. However, experience from this research showed that the analysis of each report requires considerable time and sufficient knowledge of the aspects and the results might be considerably affected by the subjectivity of the analyst. A way to use these as metrics was not immediately apparent.

- Dependency on initial conditions: Dekker (2011) and Leveson (2011, 2015) discuss the importance of the decisions and assumptions made during the design of systems and the effects of former on the performance of the latter over time during operations. However, based on the experience of the researchers, such decisions and assumptions are not always documented and/or known by companies and are difficult to be obtained from system designers.
- Decrementalism: The drift of organizational performance due to small changes, which are typically judged against the success of the most recent change and not the distance from the original design, has been illustrated by Dekker (2011). However, as discussed for the safety concept above, the research team contemplates that such information is not directly available to be collected.
- Unruly technology: Various authors (e.g., Sarter, Woods and Biilings, 1997; Dekker, 2011; Chow, Yortsos, and Meshkati, 2014) have pointed the fact that end-users do not always and fully obtain a sufficient understanding of how highly-automated systems function (e.g., type of data collected and analysed, algorithms used, connections of sub-systems and interdependencies), thus, system operators are not in place to deal successfully with automation surprises. Taking into account the variety of systems operated by aviation companies, the different levels of automation of the equipment and systems operated and the diverse set of skills and knowledge across end-users, the researchers believe that the operationalization of a relevant metric could not be feasible in the frame of this project.
- Resilience: The introduction of resilience engineering into the field of safety (e.g., Nemeth and Herrera, 2015; Costella, Saurin and de Macedo Guimaraes, 2009) has not yet been accompanied by research that quantifies the resilience abilities, namely anticipation, monitoring, response and learning. The Resilience Analysis Grid (Hollnagel, 2015) provides some guidance in the decomposition of the aforesaid abilities, but yet respective metrics have not been developed. Such an endeavour cannot be undertaken during this project since it requires long time and multiple studies. Nevertheless, the researchers believe that the metrics presented in the sections 2.1.1, 2.1.2, 2.1.5 and 2.1.6 above reflect the aforesaid capabilities and they can be connected with resilience.

3. Validity of Metrics

This section describes the validity criteria to be applied to the metrics of section 2.1 above. It is noted that, as it is explained in section 3.2 below, due to the implementation of metrics to different organizations once, the external and ecological validity along with the reliability of the metrics might be evaluated in future studies through their application to a larger sample of companies within and outside the aviation domain.

3.1 Accuracy, construct, content and face validity

The criteria against which accuracy, construct, content and face validity of the metrics will be assessed are the following [adapted from Kaspers et al. (2016a) and addressing the limitations of current metrics presented in section 1 above]:

- Reflective of the respective theoretical framework;
- Encompassing systemic views, where applicable;
- Valid (i.e. meaningful representation of what is measured);
- Fulfilment of laws, rules and other requirements, where applicable;
- Measurable, so to permit statistical calculations;
- Specific in what is measured;
- Availability or easiness of obtaining hard or/and soft data required including the quantification of the latter;
- Ability to set control limits for monitoring the calculated values;
- Manageable – practical (i.e. comprehension of metrics by the ones who will use them);
- Scalable/applicable to the context and area that the metric will be used (e.g., size of the company, type of activities such as air operations, maintenance, ground services, air traffic management);
- Cost-effective, by considering the required resources;
- Immune to manipulation;
- Sensitive to changes in conditions.

The process to evaluate the fulfilment of the above criteria will be the following:

a. Draft design of metrics
b. Review of metrics within the research team
c. 1st refinement of metrics
d. Review of refined metrics by knowledge experts
e. 2nd refinement of metrics
f. Review of refined metrics through pilot tests at least two company partners per metric (ideally, 1 large company and 1 SME).
g. 3rd refinement of metrics
h. Review of metrics within the research team
i. Final refinement of metrics before application and test of internal validity (see section 3.2 below).

3.2 Internal Validity of Metrics

Following the achievement of the validity types mentioned in section 3.1 above, the next steps will be the application of new metrics to different companies, collection and analysis of data and examination of the associations between those metrics and safety outcomes. During this process, the researchers will examine the criterion, predictive, statistical conclusion validity of metrics, which will ultimately lead to the evaluation of their internal validity.

In regard to the application of the metrics to companies, the researchers anticipate specific challenges, some of which must be addressed until the end of the current phase. More specifically, the data to be collected during the surveys must refer to both new and existing metrics, including safety outcomes, and the researchers must consider the limitations encountered during the previous research phase (Kaspers et al., 2016b, 2016c). In particular:

- Due to the diverse interpretations of thresholds of safety outcomes, especially concerning the boundaries between incidents and serious incidents, the research team considers to collect data about all classes of safety events recorded by companies in the frame of their mandatory occurrence reporting systems.
- Due to time limitations and taking into account that the concepts on which most of the new metrics are based reflect changes occurred and practices developed over long time periods rather than acutely, the new metrics will be applied once per company (e.g., it is hypothesized that the states of SMS, culture development plans, complexity/coupling etc. do not change dramatically in a few weeks' time). On the side of safety outcomes, the researchers will collect data referring up to 10 years back in time in order to explore associations of metrics with outcomes across the companies with correspondence to different periods (e.g., occurrences recorded the previous 1, 2, 3, etc. years).
- Taking into consideration that actual severity is not always representative of safety performance since the same causes can lead to different types and intensities of results (e.g., Geller, 2001; McSween, 2003; ICAO, 2013), the research team will explore whether it can adapt new approaches to classification of events, such as their controllability (Karanikas, 2015a). This is expected to require additional resources for the analysis of voluntary and mandatory reports maintained by the companies, but such an activity can be performed by graduate students or staff from within the organizations that hold insights of the context. Those studies might start and finish at time points other than the surveys since it will refer to the past, and ideally to records up to 10 years.
- Regarding existing safety metrics, the team will collect respective data during the on-field surveys depending on what the companies currently use to measure their safety levels/performance, and not what they could use according to the overall industry practice.
- The surveys must be performed to many companies (i.e. ideally more than 30) so to allow the conduction of valid statistical tests. Hence, the current number of 13 company partners, assuming that all of those will participate in the surveys, is deemed limited. Also, this number is expected to be lower per metric due to particularities of companies (e.g., not all aviation companies have an SMS in place). The research team has taken initiatives to communicate the research project across its network and during conferences and other types of events in order to attract more partners, but this needs to be also support by all partners.
- The metrics and their associated instruments will be developed in the English language. This might discourage companies from participating in the research and reduce the reliability of the data to be collected due to possible misinterpretations of the terms used. The research team will focus on the

simplicity of the language used in the metrics/tools to be tested at companies. The translation of those into local languages remains an option, but this will require much time including the verification of the translated texts. Nevertheless, the researchers plan to be present during the surveys and provide explanations for terms that seem difficult to understand.

4. Conclusions

The metrics to be designed and tested during the 2nd phase of the research will be based on a spectrum of suggestions from academia and industry that have not been yet consistently or entirely operationalised. The researchers will develop the metrics by ensuring that those will meet the quality criteria mentioned in literature and the objectives of the project, the latter referring mainly to the need of SMEs for safety metrics that do not require vast amount of operational data. Since such metrics have not been previously introduced and implemented, there is a risk of not detecting relationships between the SMS/process/organizational metrics and outcome ones. However, the research team contemplates that the inclusion of the seven types of metrics mentioned in section 2.1 above, corresponding to 17 hypotheses to be tested, decreases this risk and will lead to findings that indicate the internal validity for at least part of the metrics. Nevertheless, until the design of the new metrics is accomplished, the participation of an adequate number of companies in the project must be ensured in order to test the metrics and derive conclusive results.

Acknowledgments

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

- The companies Helicentre and Transavia which participated in the preliminary pilot studies of the safety culture development tool.
- The Aviation Academy graduates Glenn Grolleman, Mohamed Abrini and Joel Eduards who worked on the preliminary design of the safety culture development and SMS evaluation tools.
- The knowledge experts who provided feedback in the preliminary concepts of the safety culture development and SMS evaluation tools (in alphabetical ascending order of organization): Dimitrios Karabelias (Athens International Airport), Robert J. de Boer (HUFAG), Ilias Panagopoulos (NATO Airlift Management Programme), Gesine Hofinger (Team-HF), Anastasios Plioutsias (Technical University of Athens) and Frank Guldenmund (TU Delft).
- The members of the knowledge experts group of the project, who reviewed the draft version of this report and provided enlightening and valuable feedback (in alphabetical ascending order of partner organization): Charlotte Mills (CAA NZ), Sidney Dekker (Griffith University), John Stoop (Kindunos), Ewout Hiltermann (KLM Cityhopper), Ruud Van Maurik (Klu/MLA), Julius Jud Keller (Purdue University) and Gesine Hofinger (Team HF).

References

Aviation Academy. (2014) “Project Plan RAAK PRO: Measuring safety in aviation – developing metrics for Safety Management Systems”, Hogeschool van Amsterdam, Aviation Academy, The Netherlands.

Boelhouwer, D. (2016). Het uitbreiden van STAMP met de work-as-done: een studie uitgevoerd bij de treindienstleiding van NedTrain / the ectensionof STAMP with work-as-done: a study executed at the rail traffic control centre of NedTrain. Unpublished BSc thesis, Aviation Academy, Amsterdam University of Applied Sciences, the Netherlands.

Butkiewicz, M., Madhyastha, H. V., Sekar, V. (2011). Understanding Website Complexity: Measurements, Metrics, and Implications. Internet Measurement Conference, 2-4 November 2011, Berlin, Germany.

Catino, M. (2008). A review of literature: individual blame vs. organizational function logics in accident analysis. Journal of Contingencies and Crisis Management, 16(1), 53-62.

Chatzimihailidou, M. M., Karanikas, N. & Dokas, I. (2016). Measuring Safety Through the Distance Between System States with the RiskSOAP Indicator. Proceedings of the 1st International Cross-

industry Safety Conference, Amsterdam, 3-4 November 2016, Journal of Safety Studies, 2(2), pp. 5-17

Chow, S., Yortsos, S. and Meshkati, N. (2014). Asiana Airlines Flight 214: Investigating Cockpit Automation and Culture Issues in Aviation Safety. Aviation Psychology and Applied Human Factors, 4(2), pp. 113–121

Cooper, G. E. and Harper, R. P. (1969). The Use of Pilot Rating in the Evaluation of Aircraft Handling Qualities. NASA Technical Note D-5153, National Aeronautics and Space Administration, Washington D.C.

Costella, M. F., Saurin, T. A., de Macedo Guimarães, L. B. (2009). A method for assessing health and safety management systems from the resilience engineering perspective. Safety Science, 47, pp. 1056-1067.

Dekker, S. (2006). The field guide to understanding human error. Bedford, UK.

Dekker, S. (2011) Drift into Failure: From Hunting Broken Components to Understanding Complex Systems. Farnham, UK: Ashgate Publishing.

EASA. (2014). A Harmonised European Approach to a Performance Based Environment. Cologne: European Aviation Safety Agency.

Eurocontrol. (2012). Effectiveness of Safety Management. Brussels: Eurocontrol.

FAA. (2006). Introduction to Safety Management Systems for Air Operators. Advisory Circular 120-92. USA: Federal Aviation Administration.

Frost, B., Mo, J. P. T. (2014). System Hazard Analysis of a Complex Socio-Technical System: The Functional Resonance Analysis Method in Hazard Identification. Australian System Safety Conference, 28 — 30 May 2014, Melbourne, Australia.

Geller, E.S., 2001. The Psychology of Safety Handbook, 2nd ed. Lewis Publishers, UK.

Grolleman, G. (2017). Aviation Academy Tool for Assessing Safety Culture Development. Unpublished Bachelor Thesis. Aviation Academy, Amsterdam University of Applied Sciences, the Netherlands.

Hollnagel, E. (2004). Barriers and accident prevention. Aldershot: Ashgate.

Hollnagel, E. (2009). The ETTO Principle: Efficiency-Thoroughness Trade Off. Farnham: Ashgate

Hollnagel, E. (2012). FRAM: The Functional Resonance Analysis Method. Modelling Complex Socio- technical Systems. Ashgate: Farnham Surrey UK

Hollnagel, E. (2014). Safety-I and Safety-II: The Past and Future of Safety Management. Ashgate: Farnham Surrey UK.

Hollnagel, E. (2015). RAG – Resilience Analysis Grid. Downloaded from http://erikhollnagel.com/onewebmedia/RAG%20Outline%20V2.pdf in 20th February 2017.

ICAO (2013). Doc 9859, Safety Management Manual (SMM) (3rd Ed.) International Civil Aviation Organization. Montréal, Canada.

Karanikas, N. (2015a), An Introduction of Accidents' Classification Based on their Outcome Control, Safety Science, 72, pp. 182-189.

Karanikas, N. (2015b). 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.

Karanikas, N. (2015c), Human Error Views: A Framework for Benchmarking Organizations and Measuring the Distance between Academia and Industry, Proceedings of the 49th ESReDA Seminar, 29-30 October 2015, Brussels, Belgium.

Karanikas, N. (2016). Critical Review of Safety Performance Metrics, International Journal of Business Performance Management, 17(3), pp. 266-285.

Karanikas, N., Soltani, P., de Boer R. J., & Roelen, A. (2015a), Evaluating Advancements in Accident Investigations Using a Novel Framework, Proceedings of the 5th Air Transport and Operations Symposium (ATOS), 20-22 July 2015, Delft University of Technology, Netherlands

Karanikas, N., Soltani, P., de Boer, R. J., Roelen, A., & Dekker, S. (2015b), Prerequisites for Safety Culture Development, Technical Report X10910/B, Amsterdam University of Applied Sciences, Aviation Academy, Netherlands

Karanikas, N., Soltani, P., de Boer R. J. & Roelen A.L.C. (2016). Safety Culture Development: The Gap Between Industry Guidelines and Literature, and the Differences Amongst Industry Sectors, in Arezes, P. (ed.), Advances in Safety Management and Human Factors, Proceedings of the AHFE 2016 International Conference on Safety Management and Human Factors, July 27-31, 2016, Walt Disney World®, Florida, USA, Springer

Kaspers, S., Karanikas, N., Roelen, A.L.C., Piric, S., & de Boer, R. J. (2016a). Review of Existing Aviation Safety Metrics, RAAK PRO Project: Measuring Safety in Aviation, Project S10931, Aviation Academy, Amsterdam University of Applied Sciences, the Netherlands.

Kaspers, S., Karanikas, N., Roelen, A.L.C., Piric, S., van Aalst, R. & de Boer, R. J. (2016b). 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.

Kaspers, S., Karanikas, N., Roelen, A., Piric, S., van Aalst, R. & de Boer, R. J. (2016c). Exploring the Diversity in Safety Measurement Practices: Empirical Results from Aviation. Proceedings of the 1st International Cross-industry Safety Conference, Amsterdam, 3-4 November 2016, Journal of Safety Studies, 2(2), pp. 18-29

Leveson, N. (2004). A new accident model for engineering safer systems. Safety science, 42(4), 237-270.

Leveson, N. (2011). Engineering a safer world: Systems thinking applied to safety. Boston, Mass: MIT Press.

Leveson, N. (2015). A systems approach to risk management through leading safety indicators. Reliab Eng Syst Saf, 136, pp. 17–34

McSween, T.E., 2003. Value-Based Safety Process. John Wiley & Sons, NJ.

NASA (1986). Task Load Index-TLX, Human Performance Research Group, NASA Ames Research Centre, California.

Nemeth C. P., Herrera, I. (2015). Building change: Resilience Engineering after ten years. Reliability Engineering & System Safety, 141, pp. 1–4

Perrow, C. (1984). Normal Accidents: Living with High-Risk Technologies. Basic Books : USA

Rasmussen, J. (1997): Risk Management in a Dynamic Society: A Modelling Problem. Safety Science, 27(2/3):183-213.

Reason, J. (1998). Achieving a safe culture: theory and practice. Work & Stress, 12(3), 293-306.

Reid, G. B., Potter, S. S. and Bressler, J. R. (1989). Subjective Workload Assessment Technique (SWAT): A User's Guide, Harry Armstrong Aerospace Medical Research Laboratory, Human Systems Division, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio

Righi, A. W., Saurin, T. A. (2015). Complex socio-technical systems: Characterization and management. Applied Ergonomics, 50, pp. 19-30.

Roscoe, A.H. (1984). Assessing pilot workload in flight. In AGARD Conference Proceedings Flight Test Techniques, Paris.

Rouse, W. B., Serban, N. (2011). Understanding change in complex socio-technical systems. Information Knowledge Systems Management, 10, pp.25–49

Sarter, N.B., Woods D. D., and Billings, C.E. (1997). Automation Surprises, in Handbook of Human Factors & Ergonomics, second edition, G. Salvendy (Ed.), Wiley.

Schöttl, F., Lindemann, U. (2015). Quantifying the Complexity of Socio-Technical Systems – A Generic, Interdisciplinary Approach. Procedia Computer Science, 44, pp. 1-10.

SIA. (2015). 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). Kenmerk: 2015-456, Nationaal Regieorgaan Praktijkgericht Onderzoek SIA. The Netherlands.

Simic, T. K., Babic, O. (2015). Airport traffic complexity and environment efficiency metrics for evaluation of ATM measures. Journal of Air Transport Management, 42, pp.260-271.

SMICG. (2012). Safety Management System Evaluation Tool. Safety Management International Collaboration Group.

Underwood, P., & Waterson, P. (2013). Accident analysis models and methods: guidance for safety professionals.

Yadav, A., Khan, R. A. (2011). Coupling Complexity Normalization Metric- an Object Oriented Perspective. International Journal of Information Technology and Knowledge Management, 4(2), pp. 501-509.

Appendix A: Conclusions from Literature Review (Kaspers et al., 2016a).

1. Safety is widely seen as avoidance of failures and is managed through the typical risk management cycle which includes the stages of hazard identification, risk assessment, risk mitigation and risk monitoring. Under this concept:
    a. Hazards are identified through a spectrum of sources such as mandatory and voluntary reports, internal and external audits, safety investigation reports, and management of change.
    b. Risk assessment is predominately based on probabilistic approaches, which employ estimations of likelihood and severity. Although it is recognised that past performance does not guarantee future performance, likelihoods and severities are estimated with the use of historical data and/or expert judgement, the latter being subject to cognitive biases. In addition, the classification of likelihood and severity classes in risk matrices is not standardised and direct comparisons of risk levels across companies are not feasible.
    c. Risk mitigation or elimination is achieved through barriers of various types (e.g., procedures, technology, training), depending on the available resources and the degree of desired control of the risk.
    d. Risks are actually monitored through the same sources that hazards are identified.
2. Safety metrics can be, conventionally, split in two groups: safety process and outcome metrics.
    a. Safety process metrics are linked with operational, organizational and Safety Management System (SMS) activities. The premise is that better and adequate SMS/safety processes lead to improvement of safety outcomes.
    b. Outcomes are occurrences of any severity category (i.e. accident, serious incident, incident) and they are used by the industry to develop respective indicators (e.g., number of occurrences per aircraft departure) for measuring safety performance. However, the thresholds for incidents and serious incidents are not clearly defined; thus, safety outcomes cannot be directly compared across organizations, and the current taxonomy is differently interpreted. Furthermore, the units of exposure (e.g., departures, miles flown, number of staff) used to develop indicators are not uniform across the industry, and companies choose the ones that confirm their expectations (e.g., correlations between numbers of safety events and operational activity figures). In addition, accidents and incidents are infrequent when considering the amount of operational activities, therefore they cannot be seen as a useful indication of current safety level.
3. There is a lack of standardization across the aviation industry for the development of safety metrics and there is no explicit reference to quality criteria regarding the design of such metrics. Companies are asked to develop their own safety metrics, a practice that offers flexibility and opportunities for customization. However, this deprives the aviation sector from establishing a common language about safety metrics and perform benchmarks.
4. Safety culture is seen as either an outcome indicator (i.e. a result of safety management) or process indicator (i.e. a reflection and indication of safety management performance). Therefore, there is a lack of consensus whether safety culture needs to be influenced in order to improve safety performance or whether the former is a sort of measurement of the latter.
5. There is limited empirical evidence about the relationship between SMS/safety process and outcome metrics and the link between those often relies on credible reasoning. Such reasoning is principally based on linear safety/accident models, where a cause-effect relation between safety management and safety outcomes is implied. Thus, the relationship between SMS/safety processes and outcome metrics is seen as monotonic in practice and follows 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 but they haven’t been extensively applied to the industry.
6. Standards have mandated the transition from compliance-based to performance-based evaluations of safety, a concept that is supported by the industry but is not yet backed with specific tools and techniques.

Appendix B: Conclusion from Surveys (Kaspers et al., 2016b).

The results of the analysis of qualitative data partially verified the findings from the literature review performed before the surveys (Appendix A). On one hand, it was confirmed that:

- Safety is managed through the risk management cycle described in standards, and companies acknowledge the limitations of current risk assessment techniques.
- The safety data collected by the companies retrofit the risk assessment and safety assurance processes.
- Safety outcomes are used as measurement of safety performance, but the definitions of their severities are ambiguous.
- Accidents and incidents are infrequent events, especially at small companies, and cannot constitute reliable measurements of safety performance.
- Companies do not use predefined quality criteria for the design of their safety metrics; each company uses metrics that are specifically tailored to their organisation in terms of type of operations and availability of data.
- Traditional approaches are used for safety management, and most of the companies follow linear models such as the Swiss cheese model and bowties. Few companies already explore newer methods and approaches to safety based on systemic models.
- Companies recognise that better indicators are necessary in the future, and there are also concerns about the feasibility of establishing metrics of high quality in the future.
- Safety culture is seen as important part of safety management.

On the other hand, the research, contradictory to the expectations raised from the literature review, revealed that:

- Current safety metrics are not grounded on sound theoretical frameworks and, in general, do not fulfil the quality criteria as proposed in literature.
- Safety culture is not a consistent part of safety metrics and, therefore, not assessed.
- The companies collect data related to their SMS processes, but such data are not associated with SMS metrics; hence, some of the processes are performed but not measured.
- The data used differ across companies depending on own perceptions, safety models used implicitly or explicitly, and available resources.
- SMS assessment is yet based on a compliance-based approach, whereas standards require the transition to a performance-based evaluation.
- Few, diverse and occasionally contradictory correlations were found between SMS process and outcome metrics. This picture might be attributed to a combination of factors, which are linked to the limitations of a linear approach and the different ways SMS processes are implemented and safety outcomes are classified.

Especially regarding the results from the causal research, those led to the partial rejection of hypotheses regarding the consistency and similarity of monotonic relationships of SMS process, operational activities and demographics with safety outcomes. Due to the limited sample size (i.e. number of participating companies and data points per company) the researchers did not claim external validity of the results and could not fully reject those hypotheses.