MATCHING RESEARCH GOALS AND METHODS IN SYSTEM SAFETY ENGINEERING

R.D. Alexander, A.J. Rae, M. Nicholson

University of York, UK, robert.alexander@cs.york.ac.uk, ajrae@cs.york.ac.uk, mark@cs.york.ac.uk

Keywords: research methods, epistemology, empirical, taxonomy, evaluation

Abstract

System safety research has a weakness in terms of evaluation. In particular, the field has a poor track record for applying appropriate research methods to the questions we wish to answer, and for evaluating the answers in a convincing way. Given changes in government and public attitudes to research funding, this will be an increasing concern in the near future. In this paper, we define a taxonomy of knowledge types which illustrates the problem and helps to solve it by matching types of questions with appropriate methods. We present prioritised suggestions for further action.

1 Introduction

It is our assertion that system safety research has problems looming.

First, research sponsors and funders are challenged by the increasingly important issue of “research governance”. Put simply, research governance is about checking research proposals before funding them, and assessing research results when it is completed. It is receiving greatly increased attention in the wake of the 2009 ClimateGate incident. A key part of this is asking researchers “Are you providing useful information?”, and a second part is “Is the information you are providing true?”. In the wake of that trails “are you providing value for [taxpayers, often as not] money?” Researchers can dodge the first two questions by not making any factual claims or practical recommendations, but then they de facto fail the value-for-money question.

Our concern is that as universities and research funding bodies tighten up on enforcement of their research governance responsibilities, system safety research is in a vulnerable position.

Secondly, research consumers are faced with competing voices and don’t know who to listen to. There is a large volume of safety advice available, but it is not clear what is good and what is not worthwhile.

Finally, researchers themselves face problems of credibility and impact. This challenges academic researchers most acutely. It is often hard for researchers to produce results that are credible to industry, and academic researchers are vulnerable to the accusation that they are out of touch. At the other extreme, academic researchers also find it difficult to get system safety research published in high-rated conferences and journals, which in turn makes it difficult for their institutions to support their activities.

We believe that the above concerns are genuine, but that there are solutions to these problems. Our argument for this has two parts. First, in [1], we use a survey of past IET proceedings to show that there are correctable deficiencies with the way research is reported which prevent us, as a research community, from producing strongly-evaluated research. Second, in this paper we show that there are methods of generating and evaluating knowledge which can be effective but are underutilised. We provide a mapping from knowledge types to methods, and point to exemplars of the use of these methods in safety or systems engineering.

Our intent is to provide a framework for assessing the breadth and rigour of safety engineering research as a whole, and to provide guidance to others who seek to improve the rigour of their work. This process is as much a learning experience for us as for anyone else, and it is something that we will draw on in our future research and teaching (particularly in our supervision of safety-critical systems Masters projects).

2 How We Got to Where We Are

2.1 The relative size of meaningful experiments and measurable effects

Research design for system safety faces challenges due to the nature of the questions asking. Key challenges include:

1. the nature of risk and uncertainty [2];
2. the difficulty of determining the risk presented by low frequency events;
3. the complexity of the systems on which system safety engineering operates; and
4. the large costs and programme risk associated with changes to established engineering practice.

The sum of these challenges is that it is seldom feasible to measure the beginning-to-end effectiveness of any suggested change in safety engineering practice. Imagine a research project requiring the design and operation of two nearly identical passenger aircraft, one as a control and the other using a new safety technique. After several billion in project costs, twenty years of regular flights, and the ethical concerns of allowing passengers to fly in the less safe aircraft, the results would still likely be inconclusive.

Similar problems have been recognised in the application domain of patient safety, where there is ongoing tension between the evidence required for new forms of medical intervention, and the limited evidence available for efficacy of patient safety initiatives (see [3-5]).

2.2 Commercial and Military Secrecy

Few organisations welcome exposure of their business practices to competitors or potential critics. This makes them unwilling to publish research about current practice, or to participate in evaluative research that compares the effectiveness of practices used by different organisations.

Large organisations have the ability to compare projects within the same organisation, but these internal studies are seldom published.

2.3 Immediate Funding Demands

There is tension between "practical advice", which is in a form suitable for immediate application, but lacks rigorous research and evaluation, and "pure research" which generates new information, but is not immediately applicable in an industrial context. Research sponsors in the field of system safety often phrase questions as requests for practical advice.

2.4 Ethical Challenges

Controlled studies, especially randomised controlled studies, are an excellent source of data. There are many ethical objections, however, to randomised controlled trials when human lives are at stake.

2.5 Epistemological Difficulties

Mahdjoubi [6] and Vincenti [7] note that the epistemology of engineering in general has been little-studied; we can contrast this with the enormous effort that has gone into the epistemology of science.

Engineering is also rife with "wicked problems" that "cannot be formulated" [8].

2.6 The Need for an Appropriate Response to Research Difficulties

These problems described in this section aren't going to go away, but by designing research appropriately, and by adapting research methods from a range of other disciplines we can improve the way we evaluate research and hence face up to the external credibility challenges we face.

Patient safety researchers, faced with similar ethical problems to system safety, have successfully mounted the argument that expending resources (and risking lives) using unproven techniques is at least as unethical as performing trials with the potential to cause harm [9].

There are attempts to construct a meaningful epistemology for engineering (e.g. [10, 11]), and we can build on these. We can divide "wicked problems" into smaller well-defined questions. We don't have to "solve the whole problem" in order to be useful. And the fact that we can't solve all problems is not an excuse to treat every part of engineering as rigour-proof.

3 A Taxonomy of Knowledge Types

There are several types of knowledge that engineering research must deal with. There are infinitely many plausible category schemes but we will present one that is of interest to our concerns. Knowledge can be divided into:

1. Definitions of terms, or systems of definitions
2. Observations and measurements of the real world
3. Theories that explain observations and measurements
4. Procedures and instructions for doing things
5. Designs of artefacts
6. Logical conclusions built upon other knowledge

Most research will cover more than one of these categories. As examples, mathematics and theoretical computer science are concerned with 1 and 6, and astronomy covers 1, 2 and 3.

An engineering project might begin with 2 to form requirements, move on to 5 and then 4, and finish with 2 again to evaluate the success of the project.

4 Using the Taxonomy to Evaluate Research

Classifying research according to the categories of knowledge it produces allows us to apply two tests for the strength of the research.

4.1 Primary Test – contribution within a category

The primary test measures the knowledge according to its contribution within a category.

Terms and definitions or logical conclusions must, if they are to make a contribution:

- Be internally logically consistent
- Meaningfully relate concepts of interest

Observations and measurements should be evaluated by scientific standards or by the methods used by historians and the like; we should ask "Are the conclusions consistent with the evidence?" or "Does this study have the ability to give the results it claims?". Large, well controlled studies of credible hypotheses have increased evaluative strength [12].

Theories are evaluated by their ability to explain observations and measurements without unnecessary hypotheses, and by their ability to suggest future experiments.

Procedures and artefacts are evaluated on practical grounds. "Do they have the attributes claimed by their creators, or bring about the benefits claimed by their creators"?

4.2 Secondary Test – Interfaces Between Categories

The secondary test measures knowledge according to the quality of interfaces between categories. For example, if we are to evaluate procedures and instructions (i.e. to make observations and measurements of their effects), then those procedures and instructions must have the following properties:

- They must have a clear scope of application
- They must have enough detail of the method to allow reproduction by others
- They must have specific claims which can be tested

If there is insufficient detail of the method, it is not possible for others to use the procedures or methods; they are an advert for consultancy rather than a research contribution. If there are no specific claims, then we don’t have the prospective benefit that would make it worthwhile to apply or evaluate the research.

4.3 The Role of Quality Claims in Engineering Research Evaluation

When evaluating procedures, instructions or designs, we can study a range of different "goods". In system safety engineering, prescriptive knowledge may include claims about benefits in terms of:

- Risk perception (knowledge about the risks of a particular system or activity)
- Risk reduction
- Risk acceptability (which may be increased by belief that the risk was appropriately managed)
- Legal/regulatory compliance (related to risk acceptability, but a goal in itself)
- Facilitation of "better" systems through safety risk management
- Cost effectiveness

For any given procedure or design, we want to know what goods it delivers, and how much of those goods can be expected under various circumstances. Subsequent research can then support a claim that some procedure or design achieves a good (e.g. Avizienis makes claims about the benefits of N-version programming in [13]) or rebut it (e.g. Knight and Leveson challenge those claims in [14]).

Observations may identify correlations or causative relations between goods and particular context factors. Such studies may be designed to collect a set of relevant data, or be specifically designed to test a hypothesised relationship.

5 Selecting Methods Appropriate to Research Goals

A complete research cycle includes:

- Investigation (both of a problem and of the available solutions)
- Detailed formulation of a question to be answered
- Knowledge generation (theories or artefacts)
- Evaluation of the knowledge
- Reflection and looking forward

In system safety we have a very broad range of types of questions we would like to answer. The list below is far from complete, but we think it captures the most important types of questions. There is existing work that strives more for completeness within safety subfields; for example Holloway and Johnson [15] provide a large set of questions for software safety research, and we don’t attempt to duplicate that here.

Each question can be answered by one or more research methods or techniques, possibly by a combination of many. It is our belief that system safety researchers don’t fully exploit the range of techniques available.

There are a wealth of research methods from all branches of science and the humanities. In system safety we cannot rely on a comfortable subset drawn from a single discipline - we deal with the problems of all disciplines.

Perhaps the most important distinction is between methods that *produce* ideas (e.g. most qualitative/ethnographic methods) and those that *evaluate* ideas (e.g. classical scientific experiments). If we’re taking a mixed-methods approach (see Section 6.2), we could draw a distinction along the quantitative/qualitative (or fixed/flexible) line that is common in the social sciences (see, for example, Robson in [16]). This distinction is particularly relevant in the messy organisational domains that much system safety research deals in.

5.1 Goals Relating to Terms and Definitions

Examples of knowledge

- A category system that allows us to distinguish between accidents, hazards, risks and causal factors.

Examples of questions

- "What are the ultimate ambitions of our research?
- “What are the intermediate research questions that must be answered to make progress in the right direction?”

Knowledge production methods

Questions of this type can be addressed through logical reasoning and discussion. Hypothetical scenarios and thought-experiments can be used to filter candidate ideas and develop good ideas into full solutions.

Notes on Evaluation

The research could be evaluated by defining a set of properties that a classification system should have (e.g. repeatability, unique classification of each item, ability to classify all relevant items) and demonstrating by proof or experiment that the system had these properties.

Exemplar papers of good practice in this area

- Holloway and Johnson [15] gives a clear set of research questions that need to be answered in software safety. It mentions the properties that they would like their set of questions to have, but notes that the set is not yet ready for evaluation against these properties.
- Avizienis et al [17] provide a comprehensive taxonomy of concepts needed for talking about dependability

5.2 Goals Relating to Observation and Measurements

Examples of knowledge

- A record of the engineer hours spent on a set of distinct safety activities over a three month period during development of a new aircraft
- An account of personal experiences while working as a consultant to a small firm developing their first safety-related component

Examples of questions

- "What kinds of claims are made, and evaluation techniques used, in recent system safety research?"
- "How many years of safety experience does the typical MOD safety manager have?"
- "What happened when we tried to create safety-critical system X?"
- "How did major accident X happen?"

Knowledge production methods

This type of question can only be answered by sharing information across the research/practice boundary. Research techniques include survey of published materials, data analysis of records, ethnographic research, participant observation, reflective practice, and interviews and surveys of participants.

Notes on Evaluation

Raw data can only be evaluated in a negative sense, such as where there is a risk that the method of data collection has contaminated the results.

Summaries and conclusions can be evaluated either directly against the data collected, or against confirmatory or contradictory data from other research.

Direct evaluation includes judgements of whether the conclusions are necessary or valid given the nature of the data. Comparative evaluation includes judgements of whether there is a direct confirmation or contradiction, or some other explanation for why different research has produced different results.

Exemplar papers of good practice in this area

- Rae et al [1]
- The survey of industrial formal methods use in Woodcock et al [18]

5.3 Goals Relating to Theories

Examples of knowledge

- A quantitative model of how the attention level of a human supervising a process plant varies given various factors and stimuli
- A model which explains and predicts accidents

Examples of Questions

- "What is the relationship between the behaviour of a human observing a process and the design of the human-machine interface?"
- "Why do some organisations have better safety records than other organisations?"
- "Is the idea of a safety culture meaningful in a measurable and quantifiable fashion?"
- "What factors increase the risk that a particular social, organisational or technical situation will lead to an accident?"

Knowledge Production Methods

Theory production appears to be a process of insight and elimination – generating candidate hypotheses and discarding ones that prove unsuitable – although this remains controversial. Researchers may use analogies with other fields of study, with art or with the natural world to assist with the generation of candidate hypotheses.

Evaluation Methods

A theory may be challenged, either by its proponent or by others, by finding observations that are inconsistent with the theory. Passing such tests does not confirm that the theory has value – a theory may be consistent with observation, yet have no explanatory power. To be supported, a theory must make interesting predictions, and have these predictions confirmed.

5.4 Questions about Procedures and Instructions

Examples of knowledge

- A method for finding hazards and their causes in chemical process plants
- A method for building models of systems that capture the behaviours of humans that are required

- A set of competency criteria for selecting individuals for various safety engineering roles

Examples of questions

All research questions answered by procedures and instructions have the same form: “How should I X?”. This can be extended to the slightly longer form (“How can I X, while maintaining Y and avoiding Z?”). The answer, in both cases, is a set of instructions for performing X.

Exemplar papers of diverse good practice in this area

- Salewski and Kowalewski [19] evaluate a specific claim about N-version programming, using an experiment with two control methods (plausible alternatives to N-version methods). Note the negative result – their experiments don’t support the claim.
- Spool [20] evaluates the implicit claim of published web usability guidelines, that “a site which follows the provided guideline will be easier to use”. He turns the claim into a testable hypothesis, and tests it with published websites. Note, however, that the conclusions drawn are an excessive leap from those supported by the evidence.

5.5 Questions about Designs

Examples of knowledge

- A software architecture that provides partitioning between mixed-criticality processes running on the same hardware
- A nuclear reactor design with inherent, robust fail-safe features

Examples of questions

As with the previous category, questions about designs are of the form “How can I X (while maintaining Y and avoiding)?”, where the answer is a design rather than a procedure.

Exemplar papers of good practice in this area

- The Tokeneer project (see [21]) gives a design of a specific product which has been rigorously evaluated. The evaluation of the product is used as a proxy evaluation of the development techniques.

5.6 Questions about Logical Conclusions

Examples of knowledge

- A set of ethical principles that allows us to decide between two courses of action, or two sets of consequences, that seem to be equivalent in pragmatic terms
- A set of principles for deciding whether a research claim should be accepted

Examples of questions

- “How do we assess how much risk we have in a system?”
- “How do we know whether to believe the claims of a research paper?”
- *“What is necessary for a system in use to be considered to have its safety effectively demonstrated? Is passage of some period of time without any unacceptable accidents or losses sufficient? Or is something additional needed?”* [15]

Exemplar papers of good practice in this area

- Hannson [22] is an example of logical reasoning from axioms about risk

5.7 Answering Composite Questions

We can note that we’ve focussed here on very precise questions that could be tackled in a single paper or research study. In contrast, the questions asked by Holloway and Johnson [15] are broader in nature (for example *“How should differences in evaluations of safety be reconciled? For example, consider a software-intensive medical device, which is considered safe by the appropriate regulatory authority, but which has occasionally failed in such a way as to lead to successful lawsuits against its manufacturer. What should be done in this case? What evidence is needed to permit an informed decision to be made by the regulatory authority?”*). In essence, Holloway and Johnson are starting with “What do we (as an engineering discipline) need to know?”, while we are concerned with “What could one individual or group set out to study in a single defined research project?”. Both approaches are important; ideally, the two will meet in the middle, with the specific research questions providing pieces of the puzzles that constitute the larger questions.

Holloway and Johnson note that the software safety community is *“is trying to answer the broad questions, without first refining those questions into more foundational questions”*. In this paper, we are concerned with level even below their “foundational questions”. We are also trying to subdivide types of knowledge, whereas many of their questions cross types.

For example, *“How do system developers obtain adequate knowledge about the intended operational environment for the system?*” spans most of the categories. **Terms and Definitions**: “what is adequate knowledge?”, **Observations and Measurements:** “what forms is the knowledge available in, and how could it be used?”, **Procedures and instructions:** (process for obtaining knowledge), **Observations and Measurements (again):** “when applied, does that process achieve the goal?”.

6 The Diversity of Possible Methods

6.1 The Flexibility of Experimental Methods

One easy criticism is that doing experiments in engineering is too hard. This can be countered by observing that there are many types of experiment, developed for a wide range of purposes. Models include the controlled (parallel control group), uncontrolled (before and after measures on one group) and longitudinal (change over time). Variants include the "wedged experiment", where an intervention is progressively introduced to a population, and quasi-experiments where naturally-occurring changes are noted and the results studied. The patient safety literature is a strong source of such methods (e.g. see Brown et al [5]).

6.2 Mixed-Methods Research

Some methods are not single "methods" but are in fact meta-methods, which tell you how to use particular methods. We can draw a useful distinction here between *meta-methods* and *data sources (*these are our terms– the research methods literature does not have a helpful consensus on these terms; each textbook has its own set). Each meta-method will allow you to use several data sources, as appropriate.

An example of a general meta-method is "ethnography". In essence it means "observing people in their normal environment", but it does not specify techniques beyond that; when you carry out an ethnographic study, you use a range of techniques including interviews and structured observation.

Any specific meta-method that combines several data sources is a "mixed-methods approach". A simple example of a mixed-methods is given by Valerdi in [23] for systems engineering research. He proposes pilot interviews, leading to point sampling, leading in turn to structured surveys and interviews, followed by data analysis of the results.

7 The "Dark Side" of Rigour

If we are careless, and obsessed with methodological rigour at the expense of all else, we will throw away knowledge that we could have had. This is not necessarily worse than drowning in unfiltered ideas (after all, we cannot make decisions in that case either), but there's no reason to think it is better.

We need to be careful when evaluating prescriptive knowledge that we don't throw away good techniques because they don't work in all situations. Pawson and Tilley in [24] decry this in social program evaluation – they describe it as the *"nothing works"* reflex. The effectiveness of an engineering technique depends on the situation in which we apply it; if we dismiss techniques when they don't work in all situations, we may end up dismissing all techniques.

8 Key Areas for Further Work

In safety engineering research, we want for good descriptive work about what really goes on in safety engineering practice. There is no shortage of anecdotes and case studies, but they are unsystematic and often not attributable. We could really benefit from some systematic qualitative studies, such as an ethnographic study of a safety team at work on a new project (perhaps done as participant observation by an experienced consultant). There would be IP and confidentially problems, and many organisations would be wary of publishing such research, but the benefits would be immense. Such research would be a significant help in bridging the gap between academic researchers and practitioners.

There are some good examples around, mostly in the form of accident reports. The Nimrod report [25] is a good example in that it gives an idea of the time-starved conditions under which the Nimrod safety case was created, and of the way that the standard model of MOD safety cases led towards a default of "the system is safe". We could do with more like that.

We can note that engineering lifecycles have a descriptive form (as in "what is the typical lifecycle of a project of type X?"), not just a normative (prescriptive) form. This is much overlooked in engineering, where the tendency is to view a lifecycle as something aspirational that is created by a guru or expert, rather than something derived from knowledge of actual practice. Good descriptive lifecycles are essential guides for people developing the techniques and methods that will have to fit into them.

In safety, we have lots of descriptive knowledge in the form of accident reports, but these are often limited in scope (The Nimrod report is unusual, for example, in its detailed coverage of safety *engineering* activities rather than just operations) and are point studies selected by a dubious criteria (that the system had a major accident – that does not necessarily mean that it was the most risky system prior to that, or that it is the most illuminating system to study in order to get knowledge about the future). More systematic descriptive studies could put us on a better footing. In particular, they could dispel doubts about whether or not an accident situation was a particularly bad example. For example, is the safety case practice described in the Nimrod Review a dire example, or is it typical across the industry?

9 Conclusions

If we are right about the research evaluation problems in the field [1], and if we don't address them, we'll face a growing credibility problem. Those of us who take public funds, or who work in publicly-funded institutions, will suffer the worst from this. There is potential, however, to improve the range of research methods *across the community* and thereby reduce this problem.

There are many people who stand to benefit from work in this area. Anyone who wants academic credibility outside the discipline will benefit from a toolbox of techniques that provides increased rigour. Anyone who consumes research results (e.g. funding bodies, corporate safety managers) will benefit from guidance on what claims deserve to be believed. Anyone who is the subject of research governance scrutiny (academics may receive this from their institution, anyone may receive this from an external funding body) will benefit from good justifications of the research methods they use.

We intend, ourselves, to reflect on how our future work matches our well-formedness properties, and we encourage others to do likewise as both producers and consumers.

Acknowledgements

The authors would like to thank Oleg Lisagor for his helpful comments on drafts of this paper.

References

1. Rae, A.J., R.D. Alexander, and M. Nicholson, *The State of Practice in System Safety Research Evaluation*, in *Proceedings of the 5th IET System Safety Conference*. 2010.
2. Wrenn, C.B., *Epistemology as Engineering.* Theoria, 2006. **72**(1): p. 60-79.
3. Shojania, K.G., et al., *Safe but Sound: Patient Safety Meets Evidence-Based Medicine.* JAMA, 2002. **288**(4).
4. Leape, L.L., D.M. Berwick, and D.W. Bates, *Evidence-Based Medicine Meets Patient Safety: What Practices Will Most Improve Safety?* JAMA, 2002. **288**(4): p. 501-507.
5. Brown, C., et al., *An epistemology of patient safety research: a framework for study design and interpretation, parts 1 to 4.* Quality and Safety in Health Care, 2008. **17**(3): p. 158-174.
6. Mahdjoubi, D., *Epistemology of Design*, in *Proceedings of the Seventh World Conference on Integrated Design and Process technology*. 2003.
7. Vincenti, W.G., *What Engineers Know and How They Know It: Analytical Studies from Aeronautical History*. 1993: John Hopkins University Press.
8. Webber, H.R.M., *Dilemmas in a general theory of planning.* Policy Sciences, 1973: p. 155-169.
9. Miller, F.G., *'Sham Surgery: An Ethical Analysis.* The American Journal of Bioethics, 2003. **3**(4): p. 41-48.
10. Figueiredo, A.D.d. *Toward an Epistemology of Engineering*. in *Proceedings of the Workshop on Philosophy & Engineering (WPE 2008)*. 2008. Royal Engineering Academy, London.
11. Cross, N., *Designerly Ways of Knowing: Design Discipline versus Design Science.* Design Issues, 2001. **17**(3): p. 49-55.
12. Ioannidis, J.P., *Why most published research findings are false.* PLoS Med, 2005. **2**(8).
13. Avizienis, A., *The N-Version Approach to Fault-Tolerant Software. IEEE Transactions on Software Engineering*, 1985. **11**(12): p. 1491-1501.
14. Knight, J.C. and N.G. Leveson, *A Large Scale Experiment In N-Version Programming*, in *Proceedings of the Fifteenth International Symposium on Fault-Tolerant Computing*. 1985: Ann Arbor, MI. p. 135-139.
15. Holloway, M. and C. Johnson, *Towards a Comprehensive Consideration of Epistemic Questions in Software System Safety*, in *Proceedings of the 4th IET System Safety Conference*. 2009.
16. Robson, C., *Real World Research*. 2002: Blackwell.
17. Algirdas Avizienis, J.-C.L., Brian Randell, *Dependability and its Threats: a Taxonomy*, in *Proceedings of the 18th IFIP World Computer Congress*. 2004.
18. Woodcock, J., et al., *Formal methods: Practice and experience.* ACM Computing Surveys, 2009. **41**(4).
19. Salewski, F. and S. Kowalewski, *Achieving Highly Reliable Embedded Software: An Empirical Evaluation of Different Approaches*, in *Computer Safety, Reliability, and Security*. 2007. p. 270-275.
20. Spool, J.M. *Evolution Trumps Usability Guidelines*. User Interface Engineering 2002; Available from: http://www.uie.com/articles/evolution_trumps_usability/.
21. *Tokeneer ID Station - EAL5 Demonstrator: Summary Report*. 2008, Praxis High Integrity Systems
22. Hansson, S.O., *Philosophical Perspectives on Risk.* Techné: Research in Philosophy and Technology, 2004. **8**(1).
23. Valerdi, R. and H.L. Davidz, *Empirical research in systems engineering: challenges and opportunities of a new frontier.* Systems Engineering, 2009. **12**(2): p. 169-181.
24. Pawson, R. and N. Tilley, *Realistic Evaluation* 1997: Sage Publications.
25. Haddon-Cave, C., *The Nimrod Review*. 2009.