Formal Methods: Oversold? Underused? A Survey

Mario Gleirscher
Department of Computer Science, University of York
York, United Kingdom
Mario.Gleirscher@york.ac.uk

Diego Marmsoler
Institut für Informatik, Technical University of Munich
Garching, Germany
Diego.Marmsoler@tum.de

***Abstract*—Context: Formal methods (FM) have been around for a while, still being unclear how to leverage their benefits, overcome their challenges, and set new directions for their improvement towards more successful transfer into practice. Objective: We study the use of formal methods in mission-critical software domains, probing industrial and academic views. Method: We perform a cross-sectional on-line survey. Results: Our results indicate an increased intent to apply FMs in industry, suggesting a positively perceived usefulness. But we observe a negatively perceived ease of use. Scalability, skills, and education seem to be among the key challenges to support this intent. Limitations: Some difficulties in achieving a large sample at a good response rate lead to limited generalizability. Conclusions: However, we present the largest study of this kind so far ($N = 192$), and our observations provide valuable insights, highlighting directions for future theoretical and empirical research of formal methods.**

***Index Terms*—formal methods, empirical research, on-line survey, practical challenges, research transfer**

I. Motivation and Challenges

Software errors were deployed, some had intolerable impact.[1] This has been *the* motivation of *formal methods (FM, Section II) as a first-class approach to error prevention, detection, and removal* (e.g. [5]). From lectures, we heard FMs are among the best we have to design and assure correct systems. The question "Why are FMs not used more widely?" [6] is justified. With a Twitter poll, emerged from our coffee spot discussions, we solicited agreement on a timely paraphrase of a statement argued by Holloway [5]: "FMs should be a cornerstone of dependability and security of highly distributed and adaptive automation." What can a tiny opportunity sample of 22 respondents from our social network tell? Not much, well, • 55% *agree*s seem to attribute importance to this role of FMs, • 14% *disagree*s oppose that view, • 32% just *don't know*. Why should and how could FMs be a cornerstone?

Since the beginning of software engineering (SE) there has been a debate on the *usefulness of FMs* to improve SE. In the 1990s, FM researchers have started to examine this usefulness with the aim to respond to critical observations of practitioners.

Hall [7] and Bowen and Hinchey [8] illuminate 14 myths (e.g. "formal methods are unnecessary"), providing their insights on when FMs are best used and highlighting that FMs can be an overkill in some cases but highly recommendable in others. The transfer of FMs into SE practice is by far not straightforward. Knight, DeJong, Gibble, *et al.* [6] examine reasons for the low adoption of FMs in practice. Barroca and McDermid [9] ask: "To what extent should FMs form part of the [safety-critical SE] method?"

Glass [10, pp. 148f, 165f] and Parnas [11] observe that "many [SE] researchers advocate rather than investigate" by assuming the need for more methodologies. Glass summarizes that FMs were supposed to help representing firm requirements concisely and support rigorous inspections and testing. He observes that *changing requirements* have become an established practice even in critical domains, and inspections, even if based on FMs, are insufficient for complete error removal. In line with Barroca and McDermid [9, p. 591], he notes that FMs have occasionally been sold as to make error removal complete, but there is no silver bullet [10, pp. 108f]. Bad communication between theorists and practitioners sustains the issue that FMs are taught but rarely applied [10, pp. 68ff]. Parnas [11] compares alternative paradigms in FM research (e.g. axiomatic vs. relational calculi) and points to challenges of FM adoption (e.g. valid simple abstractions).

In contrast, Miller, Whalen, and Cofer [12] draw positive conclusions from recent applications of *model checking* and highlight lessons learned. In his keynote, O'Hearn [13] conveys positive experiences in scaling FMs through adequate tool support for *continuous reasoning* in agile projects (see also, e.g. [14]). Many researchers (see, e.g. [15]) have been working on the improvement of FMs towards their successful transfer. Boulanger [16] and Gnesi and Margaria [17] summarize promising industry-ready FMs and present larger case studies.

Have software errors been overlooked because of not having been detected as inconsistencies in a formalism? Are such errors a compelling argument for the use of FMs? Strong evidence for *the ease of use of FMs and their efficacy and usefulness* is scarce and largely anecdotal, rarely drawn from *comparative studies* (e.g. [18]), often primarily conducted in research labs (e.g. [14], [19] and many others). Graydon [20] observes a lack of evidence for the effectiveness of FMs in assurance argumentation for safety-critical systems, suggesting

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GL 915/1-1.

(Pr)eprint Reference Format: Gleirscher, M., & Marmsoler, D. (2018). *Formal Methods: Oversold? Underused? A Survey*. Unpublished working paper. Department of Computer Science, University of York, United Kingdom. eprint: arXiv reference.

[1] See anecdotal evidence (gray literature, press articles) on software-related incidents, e.g. by Kaner and Pels [1], [2], Neumann [3] and Charette [4].

empirical studies to examine hypotheses and collect evidence.

FMs have many potentials but SE research has reached a stage of maturity where strong empirical evidence is crucial for further *research progress and transfer*. Jeffery, Staples, Andronick, *et al.* [21] identify questions and metrics for *FM productivity assessment*, supporting FM research transfer.

Contributions: We contribute to SE and FM research by 1) presenting results of the largest cross-sectional survey of FM use among SE researchers and practitioners to this date, 2) answering research questions about the past and intended use of FMs and the perception of systematically mapped [22] FM challenges, 3) relating our findings to the perceived ease of use and usefulness of FMs according to the technology acceptance model (TAM) [23], and 4) providing a research design for repetitive (e.g. longitudinal) FM studies.

Overview: The next section introduces important terms. Section III relates our work to existing research. In Section IV, we explain our research design. We describe our data and answer our research questions in Section V. In Section VI, we summarize and interpret our findings in the light of existing evidence and with respect to threats to validity. Section VII highlights our conclusions and potential followup work.

II. Background and Terminology

By *formal methods*, we refer to *explicit* mathematical models and *sound* formal logical reasoning about *critical properties* [24]—such as reliability, safety, availability, security, and dependability and effectiveness in general—of electrical, electronic, and programmable electronic or software systems in mission- or property-critical application domains. Model checking, theorem proving, abstract interpretation, assertion checking, and formal contracts are examples of FMs. By *use of FMs*, we refer to their application to critical engineered systems, including the use of notations (e.g. UML) and tools.

The *technology acceptance model* [23] incorporates a model to assess end user IT technology using the two constructs *perceived ease of use* and *perceived usefulness*. Because FMs are often supported by IT tools, we find it reasonable to adopt the TAM for the assessment of engineering methods. *Ease of use* (EOU) of a FM characterizes the type and amount of effort a user is likely to spend to learn, adopt, and apply this FM. *Usefulness* (U) determines whether a specific FM is fit for purpose, that is, whether it supports the engineer to accomplish an appropriate task. If EOU and U are measured by a survey whose data points are user perceptions then we talk of *perceived ease of use (PEOU)* and *perceived usefulness (PU)* according to Davis [23]. Together, PEOU and PU form the *user acceptance of FMs* (and corresponding tools).

III. Related Work

For each reference in Table I, we list the authors' estimated attitude against or in favor of FMs, the motivation of the study, the followed approach, and the type of result obtained. Most of the work presents personal experience and opinion and single case studies or literature. In contrast, the work presented in this paper focuses on the analysis of experience from multiple experts. However, we found three similar studies.

Austin and Parkin [25] had the aim to explain the low acceptance of FMs in industry around 1992. Using a questionnaire similar to ours but open, they analysed 126 answers from a sample of size 444, using a sampling method similar to ours (then using different channels). Responses from FM users are distinguished from general responses. Their questions examine benefits, limitations, barriers, suggestions to overcome those barriers, personal reasons for or against the use of FMs, and ways of assessing FMs. In knowledge of FM benefits, we steered our half-open questionnaire towards a refined classification of responses, comparing past with intended use, and interrogating recently perceived obstacles. We received their paper report after finishing our study and conclude that our work can be seen as a near-replication.

In a second study in 2001, Snook and Harrison [26] conduct single interviews with representatives from five companies to discover the main issues involved in the use of FMs, in particular, the issues of understandability and the difficulty of creating and leveraging formal specifications.

A similar, though more comprehensive interview study was performed by Woodcock, Larsen, Bicarregui, *et al.* [27] in 2009. They assess the state of the art of the application of FMs, using questionnaires to collect data on 62 industrial projects.

While these studies focus on the elicitation of the state of the art, the main focus of our study is to compare the current state of the art with the desired state of the art. To the best of our knowledge, our study offers the largest set of data points investigating the use of FMs in SE, so far. For a discussion of how our findings relate to the findings of these studies, we refer to Section VI-C.

Table I: Overview of related work on FM use and adoption

Ref.	Att.	Motivation	Approach	Result
[25]	n/a	n/a	*Surv./Interv.*	n/a
[26]	=	LoEv	*Surv./Interv.*	Analysis / Eval.
[28]	=	Edu./Train.	*Survey*	Analysis / Eval.
[27], [29]	=	LoEv	*Surv./Interv.*	Challenges
[30]	=	SotA	Lit.	Analysis / Eval.
[31]	+/–	TechTx	Lit.	Recom.
[32]	=	TechTx	O/E	Challenges
[9]	+/–	SotA	O/E	Challenges
[8]	+	Hyp. Testing	O/E	Recom.
[33]	+	TechTx	O/E	Recom.
[34]	–	TechTx	O/E	Challenges
[35]	+	TechTx	O/E	Method
[36]	+	TechTx	O/E	Challenges
[37]	+	Hyp. Testing	O/E, Lit.	Recom.
[11]	=	TechTx	O/E	Chall., Recom.
[7]	+	Hyp. Testing	CS, O/E	Recom.
[38]–[40]	+	SotA	mult. CS, O/E	Analysis / Eval.
[6]	=	TechTx	CS, Field Exp.	Analysis / Eval.
[18]	=	Hyp. Testing	Lab Exp.	Analysis / Eval.
[12]	=	TechTx	mult. CS, O/E	Analysis / Eval.
[14]	=	TechTx	CS	Analysis / Eval.

Legend: +/=/– ... pos./neutral/neg., CS ... case study, Exp. ... experiment, Lit. ... literature survey, LoEv ... lack of empirical evidence, O/E ... opinion/experience report, Recom. ... recommendations, SotA ... state of the art, TechTx ... technology transfer

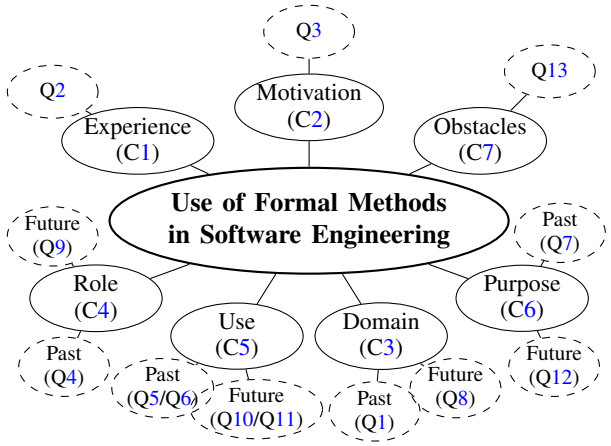

Figure 1: Constructs identified for analysis

IV. Research Method

In this section, we describe our research design, our survey instrument, and our procedure for data collection and analysis. For this research, we follow the guidelines of Kitchenham and Pfleeger [41] for self-administered surveys and use our experience from a previous more general survey [42].

A. Research Goal and Questions

The questions in Section I have led to this survey on the *use, usage intent, and challenges of FMs*. Our interest is particularly devoted to the following *research questions (RQ)*:

- RQ1 In which typical domains, for which purposes, in which roles, and to what extent have *FMs been used*?
- RQ2 Which *discrepancies* can we observe between FM users' *experience and intentions to use FMs*?
- RQ3 Who perceives which FM *challenges* to be how difficult?
- RQ4 What can we say about the *perceived ease of use* and the *perceived usefulness* of FMs?

B. Construct and Link to Research Questions

Figure 1 depicts the constituents of our construct called *use of FMs in mission-critical SE* (UFM). The *construct scales* are shown in Table II.

For RQ1 (UFM), we examine potential • application *domains* for FMs (C3), • *roles* when using FMs (C4), • *motivations* and *purposes* of using FMs (C2, C6), and • the extent of UFM as the general (C1) and specific (C5) experience level of our study participants when using FMs. For RQ2, we compare the *past* (UFM_p) and *intended use* (UFM_i) of FMs regarding domain (C3), role (C4), FM class (C5), and purpose (C6). For RQ3, we measure the perception of difficulty of known obstacles (C7) depending on UFM.

For RQ4, we associate our findings from RQ2 and RQ3 with the TAM. Because our study design does not allow to measure effort, we approximate EOU qualitatively by the challenges to overcome in typical FM applications. We then interpret the answers to RQ3 to examine and predict the PEOU and, furthermore, interpret the answers to RQ2 to reason about PU.

Table II: Elements of our construct according to Figure 1. **Legend:** MC... multiple-choice, *... measured twice.

Id.	Description [Scale(s)]
C1	Level of FM experience [duration ranges in years]
C2	Motivation to use FMs [degree per motivational factor]
C3*	Application domains of FMs [MC among domains]
C4*	Role in using FMs [MC among roles]
C5*	Use of FMs [experience level/relative frequency per FM class]
C6*	Purpose of using FMs [absolute/relative frequencies per purpose]
C7	Difficulty of obstacles to using FMs [degree per challenge]

In Section IV-D, we discuss our questionnaire including the questions for measuring the sub-constructs.

C. Study Participants and Population

Our target group for this survey includes persons with • an educational background in engineering and the sciences related to critical computer-based or software-intensive systems, preferably having gained their first degree, *or* • a practical engineering background in a reasonably critical systems or product domain involving software practice. We use *(study or survey) participant* and *respondent* as synonyms, we talk of *FM users* whenever appropriate.

D. Survey Instrument: On-line Questionnaire

Table III summarizes the questionnaire we use to measure the construct in Figure 1. The scales used for encoding the answers are described in Table IV. Most questions are half-open, allowing respondents to go beyond given answer options. We treat *degree* and *relative frequency* as 3-level Likert-type scales.

For each question, we provide *"do not know" (dnk)*-options to include participants without previous knowledge of FMs in any academic or practical context. If participants are not able to provide an answer they can choose, e.g. "do not know", "not yet used", "no experience", or "not at all", and proceed. This way, we reduce bias by forced responses. Below, we indicate *dnk*-answers whenever we *(ex)clude* them. Our questionnaire tool (Section IV-F) supports us *with getting complete data points*, reducing the effort to deal with missing answers.

Although we do not collect personal data, respondents can leave us their email address if they want to receive our results. We expect participants to spend about 8 to 10 minutes to complete the questionnaire.

E. Data Collection: Sampling Procedure

We could not find an open/non-commercial panel of engineers. Large-scale *panel services* are either commercial (e.g. [43]) or they do not allow the sampling of software engineers (e.g. [44]). Hence, we opt for a mixture of opportunity, volunteer, and cluster-based sampling. To draw a reasonably diverse sample of potential FM users, we

1) advertise our survey on various on-line discussion channels,
2) invite software practitioners and researchers from our social networks, and

Table III: Summary of questions from the questionnaire. **Legend:** MC... multiple-choice

Id.	Question or Question Template	Scale (see Table IV)	Sec.	Fig.
Q1	In which *application domains* (C3) in industry or academia have you mainly used FMs?	MC among domains	V-B1	3
Q2	How many years of *FM experience* (including the study of FMs, C1) have you gained?	duration range in years	V-B2	4
Q3	Which have been your *motivations* (C2) to use FMs?	degree per motivational factor	V-B3	5
Q4	In which roles (C4) have you used FMs?	MC among roles	V-C1	6
Q5	Describe your *level of experience* (C5) for ⟨*class of formal description techniques*⟩.	level of experience per class	V-C2	7
Q6	Describe your *level of experience* (C5) for ⟨*class of formal reasoning techniques*⟩.	level of experience per class	V-C3	8
Q7	I have mainly *used FMs for* (C6) ...	absolute frequency per purpose	V-C4	9
Q8	In which *domains* (C3) in industry or academia do you intend to use FMs?	MC among domains	V-D1	10
Q9	In which *roles* (C4) would (or do) you intend to use FMs?	MC among roles	V-D2	11
Q10	I (would) *intend to use* (C5) ⟨*class of formal description techniques*⟩ ⟨*this*⟩ often.	relative frequency per class	V-D3	12
Q11	I (would) *intend to use* (C5) ⟨*class of formal reasoning techniques*⟩ ⟨*this*⟩ often.	relative frequency per class	V-D4	13
Q12	I (would) intend to *use FMs for* (C6) ⟨*purpose*⟩.	relative frequency per purpose	V-D5	14
Q13	For any use of FMs in my future activities, I consider ⟨*obstacle*⟩ (C7) as ⟨*that*⟩ difficult.	degree of difficulty per obstacle	V-E1	15

Table IV: Scales used in the questionnaire

Name	Values	Type
degree of motivation	**"no motivation"**, "moderate motivation", "strong motivation (or requirement)"	L3
degree of difficulty	"not as an issue.", "as a moderate challenge.", "as a tough challenge.", **"I don't know."**	L3
experience level (duration-based)	**"I do not have any knowledge of or experience in FMs."**, "less than 3 years", "3 to 7 years", "8 to 15 years", "16 to 25 years", "more than 25 years"	O
experience level (task-based)	**"no experience or no knowledge"**, "studied in (university) course", "applied in lab, experiments, case studies", "applied once in engineering practice", "applied several times in engineering practice"	O
frequency (absolute)	**"not at all."**, "once.", "in 2 to 5 separate tasks.", "in more than 5 separate tasks."	O
frequency (relative)	**"no more or not at all."**, "less often than in the past.", "as often as in the past.", "more often than in the past.", **"I don't know."**	L3
choice	single/multiple: (*ch*)ecked, (*un*)checked	N

Legend: In bold, options to express lack of knowledge or indecision. (N)ominal, (O)rdinal, Ln ... LIKERT-type scale with n values.

Table V: Channels used for sampling

Channel Type	Examples & References
General panels	SurveyCircle, www.surveycircle.com
LinkedIn groups	E.g. on ARP 4754, DO-178, FME, ISO 26262
Mailing lists	E.g. system safety (U Bielefeld, formerly U York)
Newsletters	BCS FACS; GI RE, SWT, TAV
Personal pages	E.g. Facebook, Twitter, LinkedIn, Xing
ResearchGate	Q&A forums on www.researchgate.net
Xing groups	E.g. Safety Engineering, RE

3) ask these people to disseminate our survey.

To check how well our *sample represents our targeted population*, we examine C1, C3, C4, and C5 from Table II for balanced levels.

F. Data Evaluation and Analysis

For RQ1, we summarize the data and apply descriptive statistics for categorical and ordinal variables in Section V-C. We answer RQ2 by comparison of the data for the past and future views regarding the domain (C3), role (C4), FM class (C5), and purpose (C6) in Section V-D. We answer RQ3 by

- describing the *challenge difficulty ratings* after associating one of 1) domain, 2) motivational factor, 3) role, 4) purpose, and 5) FM class with challenge (C7) and
- distinguishing 1) more experienced (ME, >3 years) from less experienced respondents (LE, ≤ 3 years), 2) practitioners (P, practiced at least once) from non-practitioners (NP, not used or only in course or lab), 3) motivated (M) from unmotivated respondents (U), 4) respondents' (P)ast and (F)uture views, and 5) respondents with increased (II) from ones with decreased usage intent (DI).

in Section V-E. We apply association analysis between these categorical and ordinal variables, using *pairs of matrices* (e.g. Figure 16). The cells represent combinations of the scales, each cell containing data about the *mode* and *(med)ian* of degree of difficulty ratings, their *proportion* of *tough* ratings, and the *actual numbers* of data points. We answer RQ4 by arguing from the results for RQ1, RQ2, and RQ3.

Half-Open and Open Questions: We code open answers in additional text fields as follows: If we can subsume an open answer into one of the given options, we add a corresponding rating (if necessary). If we cannot do this then we introduce a new category "Other" and estimate the rating. Finally, we cluster the added answers and split the "Other" category (if necessary). For Q13, we performed the latter step combined with independent coding to confirm our consistent understanding of the challenge categories [45].

Tooling: We use Google Forms [46] for implementing our questionnaire and for data collection (Section IV-E) and storage. For statistical analysis and data visualization (Section IV-F), we use GNU R [47] (with the `likert`, `gplots`, and `ggplot2` packages and some helpers from "Cookbook for R" and "Stackexchange R community"). Content analysis and coding takes place in a spreadsheet application. Electronic supplementary material to this work is available in [48].

V. Execution, Results, and Analysis

We describe the sample, summarize the responses to Table III, and answer our research questions (Section IV-A).

Figure 2: Distribution of responses over time

A. Survey Execution

For data collection, we 1) advertised our survey on the channels in Table V and 2) personally invited > 30 persons. The sampling period lasted *from August 2017 til August 2018*. In this period, we *repeated step 1 up to three times* to increase the number of participants. Figure 2 summarizes the distribution of responses. The channels in Table V particularly cover the European and North American areas.

B. RQ 1: Description of the Sample

Assuming participants are, on average, member of at least three of the channels listed in Table V, an estimate of 65K *channel memberships* indicates that we could have *reached* up to 20K *real persons*. Given a recent estimate of worldwide 23 million SE practitioners [49] and assuming that at least 1% are mission-critical SE practitioners, our *population* might comprise at least 230K persons. We received $N = 192$ responses resulting in a *response rate* of about 1% and a *population coverage* of at most 0.1%. About 40% of our respondents provided their email addresses, including many from the US, UK, Germany, France, and a sixth from other EU and non-EU countries.

1) Q1: Application Domain: For each domain, Figure 3 shows the number of participants having experience in that domain[2]. Note that 160 of the respondents do have experience with applying FM in different industrial contexts, while only 32 have not applied FMs to any application domain.

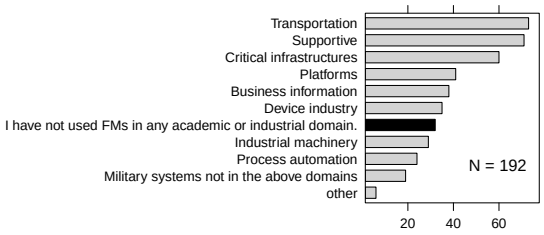

Figure 3: In which application domains in industry or academia have you mainly used FMs? (MC)

2) Q2: FM Experience: Figure 4 depicts participants' years of experience in using FMs, showing that the sample is *well-balanced* w.r.t. the experience levels. According to Section IV-F, one third of the participants can be considered LEs with up to three years of experience, and two thirds can be considered MEs with at least three years of experience (28 of those with even more than 25 years).

[2]MC entails that the sum of answers can exceed N.

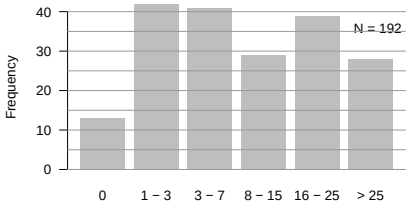

Figure 4: How many years of FM experience (including the study of FMs) have you gained?

3) Q3: Motivation: From Figure 5 it seems that *regulatory authorities* play only a subordinate role in the use of FMs. In contrast, *intrinsic motivation* (in terms of private interest) seems to be the major factor for using FMs. For twelve respondents, none of the given factors was motivating at all.

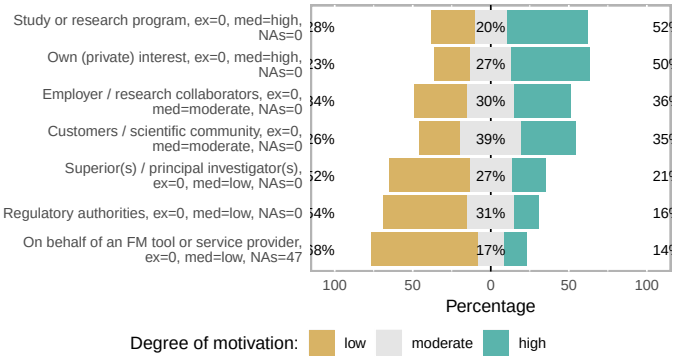

Figure 5: Which have been your motivations to use FMs?

C. RQ 1: Facets of Formal Methods Use

1) Q4: Role: Figure 6 shows in which roles the respondents applied FMs. An analysis of MC-answers shows that 72% of the participants used FMs in an *academic environment*, as a researcher, lecturer, or student. 52% of the participants applied FMs in *practice*, as an engineer or consultant.

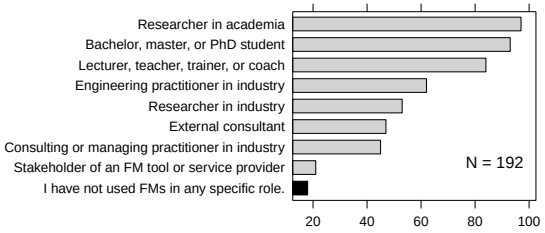

Figure 6: In which roles have you used FMs? (MC)

2) Q5: Use in Specification: The degree of usage of FMs for specification is depicted in Figure 7. There is an *almost balanced* proportion between theoretical and practical experience with the use of various specification techniques. Only the use of FMs for the description of dynamical systems seems to be remarkably low.

3) Q6: Use in Analysis: The use of FMs for analysis is depicted in Figure 8. Similar to specification techniques, we

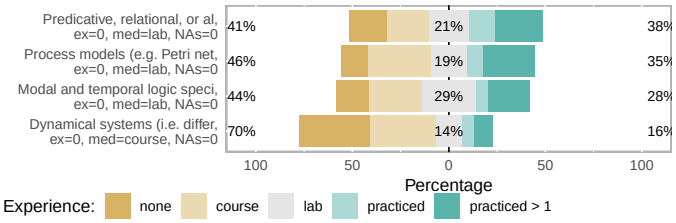

Figure 7: Describe your level of experience with each of the following classes of formal description techniques?

observe an *almost balanced* proportion between theoretical and practical experience with the usage of various analysis techniques. Outstanding is the use of assertion checking techniques, such as contracts. As expected from the observations of Section V-C2, the use of FMs for dynamical systems analysis, such as differential calculus, is again exceptionally low.

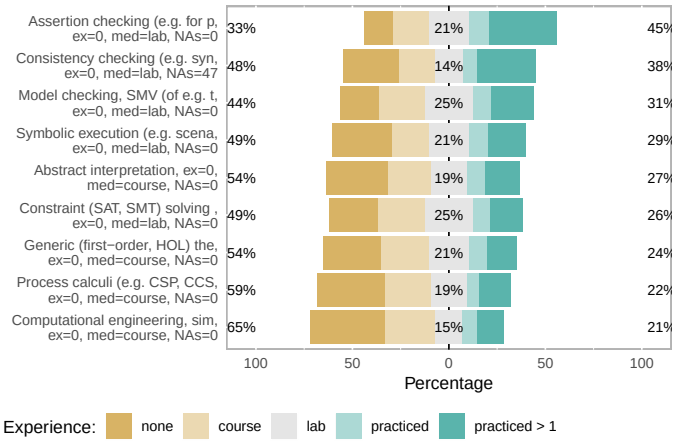

Figure 8: Describe your level of experience with each of the following classes of formal reasoning techniques?

4) Q7: Purpose: Figure 9 depicts the participants' purposes to apply FMs. It seems that they employ FMs mainly for specification, verification, and error detection. Synthesis, on the other hand, seems to be only a subordinate purpose in the use of FMs.

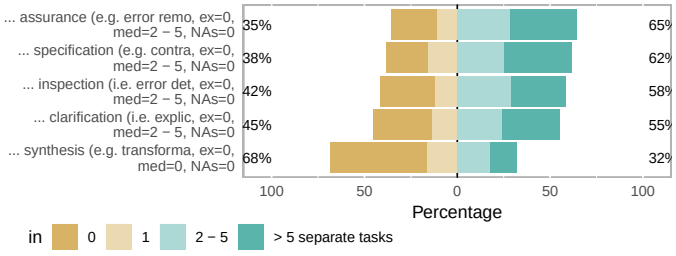

Figure 9: I have mainly used FMs for ...

D. RQ 2: Past Use versus Usage Intent

We investigate the usage intent of FMs across various domains and roles as well as the participants' intent to use various FMs and their intended purpose to use FMs.

1) Application Domain: Figure 10 compares the respondents' current application domain of FMs with their intended one (see Q8). It reveals two insights into the participants intention to use FMs: (i) The number of participants which did not yet use FMs is almost double the number of participants which want to apply FM in the future. Thus, some of the participants which did not use FMs, so far, aim to apply them in the future. (ii) The intended application of FMs clearly outperforms the current application of FMs across *all* domains. Thus, their is a clear tendency to increase the use of FMs across all application domains.

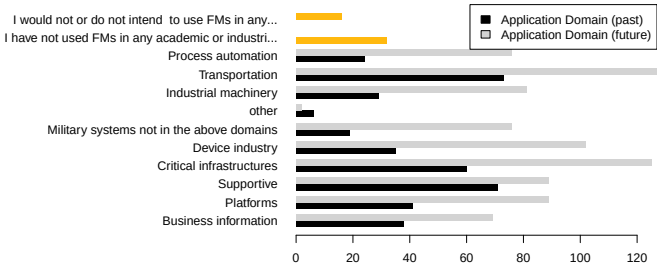

Figure 10: # of resp. using FMs by domain (past vs. intent)

2) Role: Figure 11 compares the participants' roles in which they applied FMs in the past with their intended role to apply FMs in the future (see Q9). Similar to the application domain, we can observe that some participants who have not applied FMs in any role so far, intend to apply such methods in the future. However, the comparison reveals another interesting observation: *Academic* disciplines (student, researcher, or lecturer) seem to be *saturated*, i.e., there is a decrease (or only small increase) in the number of participants who applied FMs to academic domains in the past and the number of participants who want to apply such methods to these domains in the future. In contrast, there is a *significant* increase in the number of participants aiming to apply FMs, across all *industrial* roles.

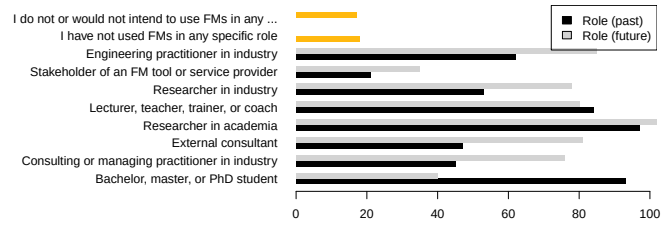

Figure 11: # of resp. applying FMs by role (past vs. intent)

3) Q10: Intended use for Specification: Figure 12 depicts the respondents' intended *future* use of applying various formal techniques for system specification. In general, the figure shows an *almost equal* amount of participants aiming to decrease and increase their use of FMs for specification. Only *dynamical* system models seem to be an exception again: more participants want to decrease their use of this technology, compared to participants who want to increase it.

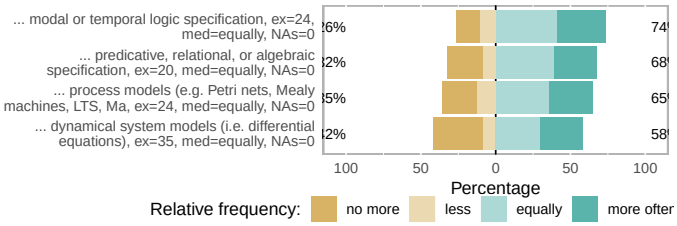

Figure 12: I (would) intend to use ...

4) Q11: Intended use for Analysis: The respondents' intended use of formal techniques for the analysis of specifications is depicted in Figure 13. The figure shows for every technique a clear tendency of the participants to *increase* their use of the technique in the future.

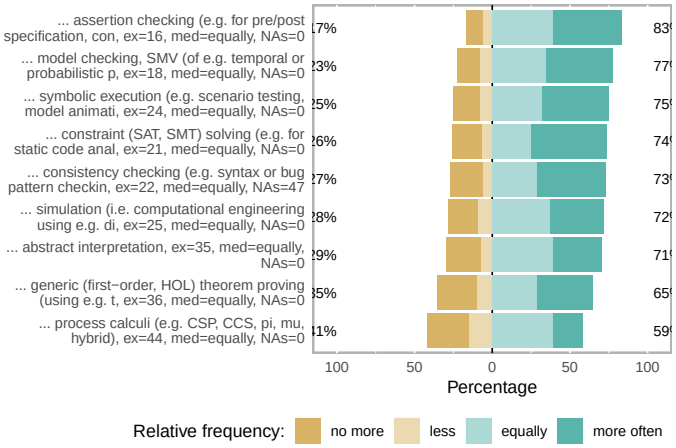

Figure 13: I (would) intend to use ...

5) Q12: Intended Purpose: The purpose respondents aim to apply FMs to in the future is depicted in Figure 14. Again their is a clear tendency of the participants to *increase* their use of FMs for all purposes.

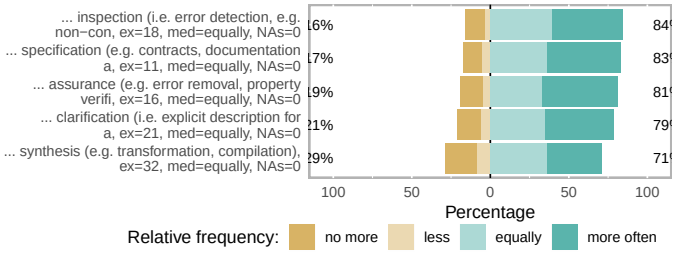

Figure 14: I (would) intend to use FMs for ...

E. RQ 3: Perception of Challenges

Table VI lists the FM challenges subject to discussion, their background, and literature referring to them. We apply the procedure described in Section IV-F.

1) General Ranking (Q13): Figure 15 shows the respondents' ratings of all challenges. Most of the participants believe that *scalability* will be the toughest challenge and *maintainability* is considered the least difficult of all rated obstacles. For *reuse of proof results*, *proper abstractions*, and *tool support*, our participants distribute more uniformly across moderate and high difficulty.

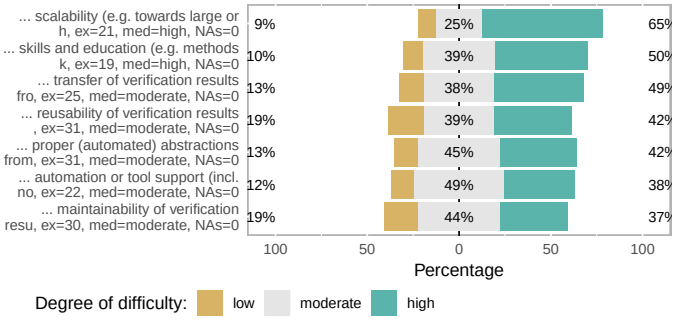

Figure 15: For any use of FMs in my future activities, I consider ⟨*obstacle*⟩ as [not an|a moderate|a tough] issue.

2) Less Experienced vs. More Experienced Respondents (Q2): The comparison of the difficulty ratings of LEs with the ratings of MEs shows that • LEs less often perceive the given challenges as tough, • MEs significantly more often rate *scalability* as tough, • both groups show the closest agreement on *transfer of verification results* and *skills and education*.

3) Non-Practitioners vs. Practitioners by Past Purpose (Q7): The perception of *skills and education* and *scalability* as the most difficult challenges is **largely independent of the purpose**, again **Ps attributing more significance** to *scalability*. The leadership of scalability in Figure 15 comes along with most tough-ratings from NPs in *synthesis* and from Ps in *assurance* and *clarification*.

4) Decreased vs. Increased Intent by Purpose (Q12): The comparison of the difficulty ratings of respondents with no or decreased intent to use FMs *for a specific purpose* and respondents with equal or increased intent shows: • The leadership of *scalability* and *skills and education* in Figure 15, particularly, comes along with most tough-ratings from IIs for *assurance* and *clarification* and from DIs for *synthesis* (54%). • The trend in Figure 15 is more clearly observable from IIs than from DIs, where *transfer of verification results* and *automation and tool support* seem to be tougher than *skills and education*.

5) Non-Practitioners vs. Practitioners by FM Class (Q5, Q6): Figure 16 shows • for NPs, the trend in Figure 15 is **largely independent of the FM class**, except for *consistency checking* leading with a *tough* proportion of 62%. • For Ps, difficulty ratings across FM classes vary more: The challenges leading in Figure 15 received the most *tough*-ratings from users of *process models*, *dynamical systems*, *process calculi*, *model checking*, and *theorem proving*. Difficulty ratings of users are often centered on moderate or tough, *proper abstraction* and *skills and education* show a comparably wide variety across FM classes. • **NPs' difficulty ratings vary less than Ps' rations**, being more independent from FM classes.

6) Decreased vs. Increased Intent by FM Class (Q10, Q11): • The trend in Figure 15 comes along with many

Table VI: Feedback on given and additional challenges. **Legend:** Q . . . in questionnaire, P . . . additionally raised by participants

Challenge Name & Description	Src.	Addressed/Examined in	Findings for RQ3 (Section V-E)
Scalability: Useful in handling large and technologically heterogeneous systems	Q	[7], [8], [12], [36], [38], [40], [50]	1st; by Ps more than by NPs; when using FMs for assurance and clarification; independent of FM class
Skills & Education: Methods known (little misconception), trained and experienced users available	Q	[7], [9], [26], [29], [32]–[36], [38], [40], [50]	2nd; agreed by LEs and MEs; largely independent of FM class; comparably small tough-proportions by Ms
Transfer of Proofs: Refinement between models and reality (e.g. code), handling incomplete specifications	Q	[7], [9], [11], [26], [30], [38], [40]	Agreed by LEs and MEs; top-rated by DIs and Us; largely independent of FM class
Reusability: Parametric proofs, reusable specifications and verification results	Q	[9], [33]	Top-rated by tool provider stakeholders and lectures
Abstraction: Useful and correct (automated) abstractions from irrelevant detail (for comprehension and validation)	Q	[6], [7], [9], [11], [12], [26], [31], [33], [35], [36]	Varies notably across FM classes
Tools & Automation: Useful notations and trustworthy tools (for manipulation, checking, collaboration, doc.)	Q	[6]–[8], [11], [13], [27], [29]–[34], [36]–[38], [40]	Top-rated by DIs; but comparably small tough-proportions from practitioners
Maintainability: Stable proofs, evolvable specifications and verification results	Q	[6], [9], [11]	Comparably small tough-proportions from practitioners
Resources: Sufficient resources, good cost-benefit ratio (despite adoption, training, licenses)	P (4)	[6]–[8], [27], [29], [30], [33], [35], [38], [40], [50]	No detailed data was collected: Because these challenges were mentioned several times each, we classify them to be at least of moderate difficulty.
Process Compatibility: Integration into existing process, method culture, standards, and regulations	P (6)	[6], [8], [13], [30]–[36], [40], [50]	
Practicality & Reputation: Benefit awareness and good empirical evidence for benefits	P (7)	[10], [11], [29], [36], [50]	

tough ratings for *transfer of verification results* from DIs of *consistency checking*. • However, DIs of *process calculi* provide comparably many tough-ratings for the generally low-ranked *automation and tool support*. • *Assertion checking* exhibits comparably low tough-proportions across all challenges whereas *process calculi* exhibit comparably high tough-ratings. • Mirroring the trend in Figure 15, **IIs show less variance than DIs across FM classes.**

7) Unmotivated vs. motivated by Motivating Factor (Q3): • Respondents with moderate to strong motivation to use FMs more likely identify given challenges as **moderate/tough, regardless of the motivating factor.** • The trend in Figure 15 comes along with many tough ratings from respondents **motivated by** *regulatory authorities* (70%) or **not motivated by** *tool providers* or *superiors/principal investigators*. • Us' *tough*-ratings are **notably lower than Ms'**.

8) Past and Future Views by Role (Q4, Q9): • Although participants show role-based discrepancies between their past and intended use of FMs (Figure 11), the **perception of difficulty** of the rated challenges seems to be **largely similar**, following the trend in Figure 15. • The high ranking of *scalability* (and *reusability of verification results*) comes along with many tough-ratings from **tool provider stakeholders** for the **past** view and many from **lecturers** for the **future** view.

9) Past and Future Views by Domain (Q1, Q8): The trend in Figure 15 comes along with highest tough-proportions for respondents from the *transportation*, *military systems*, *industrial machinery*, and *supportive* domains.

VI. Discussion

In this section, we discuss and interpret our findings, relate them to existing evidence, outline general feedback on the questionnaire, and critically assess the validity of our study.

A. Findings and their Interpretation

The following (F)indings are based on the data summarized and analysed in Sections V-B to V-D.

RQ 1: **(F1)** *Regulatory authorities* represent only a minor motivating factor to use FMs. *Intrinsic motivation* (maybe market-triggered) seems to be stronger.

RQ 2: **(F2)** It seems that in *all given domains* (Figure 10, except for *other*) respondents intend to *increase* their future use of FMs. Moreover, it seems that this tendency is *independent* of the *concrete technology* (except process calculi) or *purpose*. The data also suggest that the use of FMs in research is saturated, while there is an increased intention to apply FMs in *industrial contexts* in the future. **(F3)** From the data it seemed that experience in using a certain technology indeed impacts the intend to use this technology in the future. To investigate this suspicion, we analyzed the intended use of FM technologies based on the experience of participants in using this technology (also by association analysis, Section IV-F) [48]. Thereby, we observed that the *more experience* one has with using a specific FM technology, the *more likely she/he will apply it in the future*. No experience at all, results in an *exceptionally high resistance* against a specific FM technology and only little experience with a certain FM technology *significantly increases the willingness to apply it in the future*. Similar observations can be made for the use of FM in general for a specific purpose.

RQ 3: **(F4)** *Scalability* and *skills and education* lead the challenge ranking, independent of the domain, FM class, motivating factor, and purpose. Practitioners see scalability

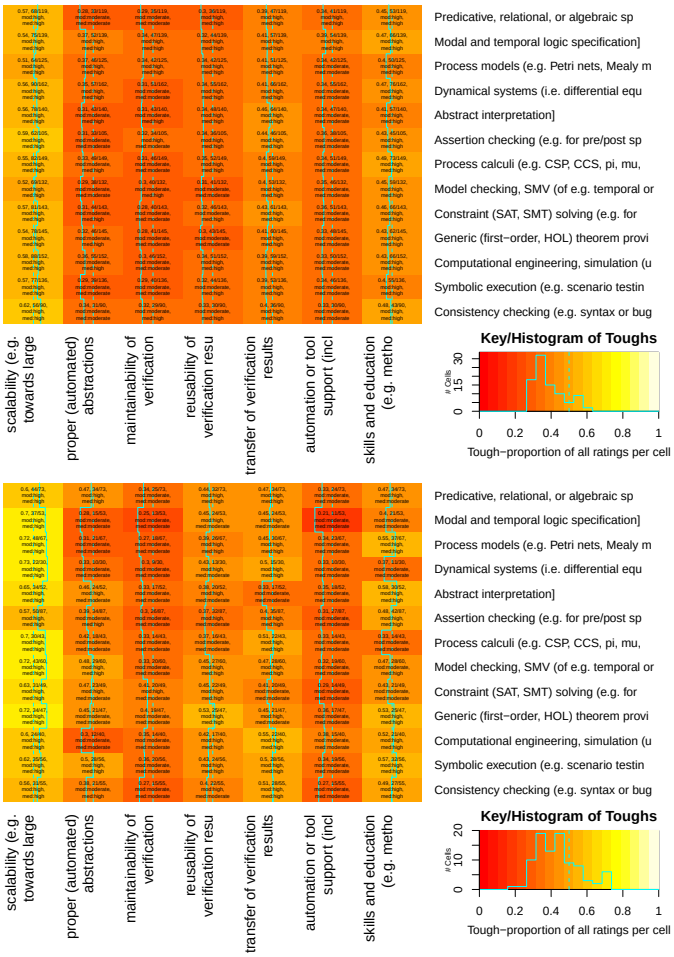

Figure 16: Difficulty of challenges (cols): NPs (top) compared to Ps (bottom) by type of used FM (rows)

as more problematic than non-practitioners, whereas non-practitioners perceive *skills and education* as more problematic than practitioners. **(F5)** *Maintainability of proof results* was found to be the least difficult challenge. **(F6)** *Reusability of proof results* was rated as tough by several practitioner groups. **(F7)** Furthermore, our respondents raised three additional challenges which we cross-validated with the literature (see highlighted rows in Table VI). **(F8)** Challenges are perceived *as moderate or tough*, largely similar between the pairs of groups we distinguish in Section IV-F. **(F9)** Process models were rated toughest for *scalability* which contrasts their high reputation as compositional methods. This might have been induced by the difficulty of scalability of model checking (Figure 16) as a frequent verification technique of process models.

B. RQ 4: Relationship to TAM (PEOU and PU)

Analogous to the reasoning in [23], an increased positive experience with practically applying FMs forms a high degree of PU. Davis [23, p. 329] observed that actual usage is strongly positively correlated with PU. According to Davis's definition of PU (Section II), we assume there is a similar (surely weaker) association between usage intent and PU. In fact, F2 suggests an increased intent to use FMs in the future. **(F10) Hence, it seems that our respondents perceive the *usefulness of FMs* more positive than negative.**

Furthermore, all challenges we discuss represent coarse substrata [23, p. 325] of the EOU for FMs because solutions to these challenges contribute to an increase in EOU. Hence, we represent an increased positive user experience with FMs by a high degree of PEOU. However, from F8, we observe that respondents rate most challenges as moderate to tough, largely independent of other variables (F4). **(F11) Hence, it seems that our respondents perceive the *ease of use of FMs* more negative than positive.**

C. Relationship to Existing Evidence

Our systematic map shows that our list of challenges is completely backed by substantial literature (see Table VI) raising and discussing these challenges. **(F12)** However, the fact that maintainability and reusability were least covered by our literature is, on the one hand, in line with F5 but, on the other, not with F6 and typical cultures of reuse in practice.

F3 is in line with other observations in [27], [29] that the repeated use of a FM results in lower overheads (i.e., an experienced effort or cost reduction and improved error removal), up to an order of magnitude less than its first use [12].

D. Threats to Validity

We assess our research design with regard to four common criteria [51], [52]. Per threat (↯), we estimate its criticality (min, maj), describe it, and discuss our mitigation (✓).

1) Construct Validity: Why would the construct (Section IV-A) appropriately represent the phenomenon?

maj ↯: Wrong or omitted questions / To support *face validity*, we applied our own experience from FM use to iteratively develop a meaningful set of questions. Because this questionnaire forms a novel instrument, we use feedback from colleagues, from respondents we personally know, and from the general feedback on the survey to improve and support *content validity*. ✓

min ↯: Questionnaire not suited for rich measurements of PEOU (e.g. per FM class) and PU / We avoid deriving conclusions specific to a FM class from our data. ✓

min ↯: Bias by omitted scale values (e.g. FM class, domain, purpose) / Respondents are encouraged to provide open answers to all questions, helping us to check scale completeness. Our systematic map confirms that we have not listed unknown challenges in Q13. We identified three additional challenges via open answers and the literature, however, unable to collect measurements for. We believe to have achieved good *criterion validity* through questions and scales for distinguishing important sub-groups (see Section IV-F) of our population. ✓

min ↯: No question about educational background / We approximate what we need to know by using data from Q1, Q3, Q4, and Q5. ✓

2) Internal Validity: Why would the procedure in Section IV lead to reasonable and justified results?

min ↯: Incomplete data points / Feedback from colleagues and first respondents made us extend Q3 with the option "on behalf of FM tool provider" and Q6 and Q11 with "consistency checking" after our 47th response. The enhancement of 145 complete data points to 192 maintained all trends. ✓

min ↯: Duplicate & invalid answers / To identify intentional misconduct, we checked for timestamp anomalies and for duplicate or meaningless phrases in open answers. Voluntarily provided email addresses (79/192) indicate only 3 unintentional double participants. Google Forms includes data points only if all mandatory questions are answered and the submit button is pressed. ✓

3) External Validity: Why would the procedure in Section IV lead to similar results with more general populations?

maj ↯: Low response rate / We believe our estimates in Section V-B to be sensible. We tried to • improve targeting by repetitively advertising on multiple appropriate channels, • spot unreliable contact information, • provide incentive (results by email), • keep the questionnaire short and comprehensible, • avoid forced answers, and • allow lack of topic knowledge. Yet there are further uncertainties such as lack of sympathy, personal motivation, and interest, or strong loyalty, and high expectations in the outcome, or intentional bias. ✓

maj ↯: Bias towards specific groups [51, p. 181] / We distributed our questionnaire on general SE channels. Our sample includes 82% of practitioners according to Section IV-F, ≈ 18% of NPs (incl. laypersons), and only ≈ 28% of pure academics. A bias towards FM experts (Figure 4) does not harm our PEOU discussion led by practitioners but shapes our PU discussion. Regarding application domains, our conclusions cannot be generalized to, e.g. finance and election sectors. ✓

min ↯: Lack of FM knowledge / 10 to 20% of our respondents did not know specific challenges (Figure 15). dnk-data points are (ex)cluded for parts of RQ1 and included in the analyses of RQ2 and RQ3 with no relevant influence. ✓

min ↯: Geographical background missing / Respondents were not required to own a Google account to avoid tracking and to increase anonymity and the response rate. The limited geographical knowledge about our sample constrains the generalizability of our conclusions, e.g. to ecosystems such as China, India, or Brazil. ✓

4) Reliability: Why would a repetition of the procedure in Section IV with different samples from the same population lead to the same results?

maj ↯: Change of proportions / The small sample and the low response rate make it hard to mitigate this risk. However, we compared the first (2017, $N_1 = 90$) and second (2018, $N_2 = 102$) half of our sample to simulate a repetition of our survey. A two-sided Mann-Whitney U test does not show a significant difference between these two groups (e.g. for Q13 and Q4), only for Q3 we recognise marginal differences. ✓

VII. Conclusions

We conducted an on-line survey of mission-critical SE practitioners to examine how FMs are used and how challenges in using FMs are perceived. Our aim was to contribute to the body of knowledge of the SE and FM communities.

Overall Findings: From the evidence we gathered for the use of FMs, we make the following observations:

- *Intrinsic motivation* is stronger than the regulatory one.
- Past experience is *correlated* with usage intent.
- Despite the challenges, our respondents show an *increased intent* to use FMs in industry.
- All challenges were rated *either moderately or highly* difficult, with scalability, skills, and education leading. Experienced respondents rate challenges as highly difficult more often than less experienced respondents.
- From the literature and the responses, we identified three additional challenges: *sufficient resources*, *process compatibility*, *good practicality/reputation*.
- Our data suggests that the *ease of use of FMs* is perceived more negative than positive.
- Gaining experience and confidence in the application of a FM seems to play a role in developing a *positive perceived usefulness of this FM*.

Hence, we believe **FMs are much more underused than oversold** in the sense of [9]. However, FMs still need to be improved and their benefits need to be better examined.

General Feedback on the Survey: The questionnaire seems to be well-received by the participants. One of them found it an "interesting set of questions." This impression is confirmed by another participant:

> "Well chosen questions which do not leave me guessing. Relevant to future FM research and practice."

Another respondent noted:

> "Thank you very much for this survey. It is very constructive and important. It handles most of the issues encountered by any practitioner and user of FMs."

Only one participant found it difficult for FM beginners.

Implications towards a Research Agenda: In the spirit of Jeffery, Staples, Andronick, *et al.* [21], we like to make another step in setting out an agenda for future FM research:

FM Improvement: To address *controllable abstractions*, we need semantics workbenches for underpinning domain-specific languages with formal semantics. We believe that further steps in *theory unification* have good potential to improve proof hierarchies, *reusability*, and *transferability*.

FM Transfer: To address *scalability*, we need more research on how compositional methods can be better leveraged in practical settings. To address *process compatibility*, we need more research in *continuous reasoning* (e.g. [13], [14]) and in cost-savings analyses of FM applications (e.g. [21]). This implies strong empirical designs (i.e., controlled field experiments) to collect strong evidence for successful transfers. To address *skills and education*, we need an enhanced *FM body of knowledge (FMBoK)* [53] with revised recommendations for lecture material [28], e.g. the teaching of modeling, composition, and refinement in practice. To address *reputation*, we

need to provide more incentives for practitioners to revive FMs and take recent progress in FM research into account when changing current software processes, policies, regulations, and standards. This includes convincing practitioners to invest in the support of large-scale studies for monitoring FM use in industry.

Future Work: Our survey is another important step in the research of effectively applying FM-based technologies in practice. To put it with the words of one of our participants: "[A] closed questionnaire is just a start." In a next survey, we like to ask about typical FM benefits, pose more specific questions on scalability and useful abstraction, and the geographical and educational background. We also like to change from 3-level to 5-level LIKERT-type scales to receive fine-granular responses. Our research design accounts for repeatability, hence, allowing us to go for a longitudinal study.

Acknowledgments: It is our pleasure to thank all survey participants for their time spent and their valuable responses, and all channel moderators for forwarding our postings. We are much obliged to Jim Woodcock, who has led previous studies in our direction, and helped us to critically reflect our work and relate it to existing evidence. We would also like to spend sincere gratitude to Krzysztof Brzezinski, Louis Brabant, and Emmanuel Eze for pointing us to valuable related work.

REFERENCES

[1] C. Kaner and D. Pels, *Bad Software*. Wiley, Aug. 1998, ISBN: 978-0471318262.

[2] ——, (Aug. 2018). Bad Software: Website, [Online]. Available: http://badsoftware.com.

[3] P. G. Neumann, "Risks to the public", *ACM SIGSOFT Software Engineering Notes*, vol. 43, no. 2, pp. 8–11, May 2018. DOI: 10.1145/3203094.3203102.

[4] R. N. Charette. (Jun. 2018). Fiat Chrysler is being sued over a software flaw, IEEE, [Online]. Available: https://spectrum.ieee.org/riskfactor/computing/software/court-allows-lawsuit-to-proceed-against-fiat-chrysler-over-software-flaw.

[5] C. M. Holloway, "Why engineers should consider formal methods", in *16th DASC. AIAA/IEEE Digital Avionics Systems Conference. Reflections to the Future. Proceedings*, vol. 1, Oct. 1997, pp. 16–22. DOI: 10.1109/DASC.1997.635021.

[6] J. C. Knight, C. L. DeJong, M. S. Gibble, and L. G. Nakano, "Why are formal methods not used more widely?", in *Fourth NASA Formal Methods Workshop*, 1997, pp. 1–12.

[7] A. Hall, "Seven myths of formal methods", *IEEE Software*, vol. 7, no. 5, pp. 11–19, 1990. DOI: 10.1109/52.57887.

[8] J. P. Bowen and M. G. Hinchey, "Seven more myths of formal methods", *IEEE Software*, vol. 12, no. 4, pp. 34–41, Jul. 1995, ISSN: 0740-7459. DOI: 10.1109/52.391826.

[9] L. M. Barroca and J. A. McDermid, "Formal methods: Use and relevance for the development of safety-critical systems", *Comp. J.*, vol. 35, no. 6, pp. 579–99, 1992. DOI: 10.1093/comjnl/35.6.579.

[10] R. L. Glass, *Facts and fallacies of software engineering*. Pearson Education (US), Oct. 28, 2002, ISBN: 978-0321117427.

[11] D. L. Parnas, "Really Rethinking 'Formal Methods'", *IEEE Computer*, vol. 43, no. 1, pp. 28–34, 2010. DOI: 10.1109/mc.2010.22.

[12] S. P. Miller, M. W. Whalen, and D. D. Cofer, "Software model checking takes off", *Communications of the ACM*, vol. 53, no. 2, pp. 58–64, Feb. 2010. DOI: 10.1145/1646353.1646372.

[13] P. W. O'Hearn, "Continuous reasoning", in *Proceedings of the 33rd Annual ACM/IEEE Symposium on Logic in Computer Science - LICS'18*, ACM Press, 2018. DOI: 10.1145/3209108.3209109.

[14] A. Chudnov, N. Collins, B. Cook, J. Dodds, B. Huffman, C. MacCárthaigh, S. Magill, E. Mertens, E. Mullen, S. Tasiran, A. Tomb, and E. Westbrook, "Continuous formal verification of Amazon s2n", in *Computer Aided Verification*, Springer International Publishing, 2018, pp. 430–446. DOI: 10.1007/978-3-319-96142-2_26.

[15] B. K. Aichernig and T. Maibaum, Eds., *Formal methods at the crossroads. From panacea to foundational support*. Springer Berlin Heidelberg, Nov. 18, 2003, ISBN: 3-540-20527-6.

[16] J.-L. Boulanger, *Industrial Use of Formal Methods: Formal Verification*. Wiley-ISTE, Jul. 11, 2012, 298 pp., ISBN: 9781848213630.

[17] S. Gnesi and T. Margaria, *Formal methods for industrial critical systems: A survey of applications*. Wiley-IEEE Press, 2013, ISBN: 9781118459898.

[18] A. Sobel and M. Clarkson, "Formal methods application: An empirical tale of software development", *IEEE Transactions on Software Engineering*, vol. 28, no. 3, pp. 308–320, Mar. 2002. DOI: 10.1109/32.991322.

[19] A. J. Galloway, T. J. Cockram, and J. A. McDermid, "Experiences with the application of discrete formal methods to the development of engine control software", *IFAC Proceedings Volumes*, vol. 31, no. 32, pp. 49–56, Sep. 1998. DOI: 10.1016/S1474-6670(17)36335-8.

[20] P. Graydon, "Formal assurance arguments: A solution in search of a problem?", in *Dependable Systems and Networks (DSN), 2015 45th Annual IEEE/IFIP International Conference on*, Jun. 2015, pp. 517–528. DOI: 10.1109/DSN.2015.28.

[21] R. Jeffery, M. Staples, J. Andronick, G. Klein, and T. Murray, "An empirical research agenda for understanding formal methods productivity", *Information and Software Technology*, vol. 60, pp. 102–112, Apr. 2015. DOI: 10.1016/j.infsof.2014.11.005.

[22] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic Mapping Studies in Software Engineering", in *12th International Conference on Evaluation and Assessment in Software Engineering, EASE 2008, University of Bari, Italy, 26-27 June 2008*, 2008. [Online]. Available: http://ewic.bcs.org/content/ConWebDoc/19543.

[23] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology", *MIS Quarterly*, vol. 13, no. 3, pp. 319–40, Sep. 1989.

[24] J. Rushby, "Critical system properties: Survey and taxonomy", *Reliability Engineering & System Safety*, vol. 43, no. 2, pp. 189–219, 1994. DOI: 10.1016/0951-8320(94)90065-5.

[25] S. Austin and G. Parkin, "Formal methods: A survey", National Physical Laboratory, Teddington, Middlesex, UK, Tech. Rep., Mar. 1993.

[26] C Snook and R Harrison, "Practitioners' views on the use of formal methods: An industrial survey by structured interview", *Information and Software Technology*, vol. 43, no. 4, pp. 275 –283, 2001, ISSN: 0950-5849. DOI: 10.1016/S0950-5849(00)00166-X.

[27] J. Woodcock, P. G. Larsen, J. Bicarregui, and J. Fitzgerald, "Formal methods: Practice and experience", *ACM Comput. Surv.*, vol. 41, no. 4, 19:1–19:36, Oct. 2009, ISSN: 0360-0300. DOI: 10.1145/1592434.1592436.

[28] J. N. Oliveira, "A survey of formal methods courses in European higher education", in *Teaching Formal Methods*, Springer Berlin Heidelberg, 2004, pp. 235–248. DOI: 10.1007/978-3-540-30472-2_16.

[29] J. C. Bicarregui, J. S. Fitzgerald, P. G. Larsen, and J. C. P. Woodcock, "Industrial practice in formal methods: A review", in *FM 2009: Formal Methods*, A. Cavalcanti and D. R. Dams, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 810–813, ISBN: 978-3-642-05089-3.

[30] R. Bloomfield, P. Froome, and B. Monahan, "Formal methods in the production and assessment of safety critical software", *Reliability Engineering & System Safety*, vol. 32, no. 1-2, pp. 51–66, 1991.

[31] C. L. Heitmeyer, "On the Need for 'Practical' Formal Methods", in *Proceedings of the 5th International Symposium on Formal Techniques in Real-Time Fault Tolerant Systems (FTRTFT)*, vol. LICS 1486, Lyngby, DenmarkLyngby, Denmark, 1998, pp. 18–26.

[32] D. Bjorner, "On the use of formal methods in software development", in *Proceedings of the 9th International Conference on Software Engineering*, ser. ICSE '87, Monterey, California, USA: IEEE Computer Society Press, 1987, pp. 17–29, ISBN: 0-89791-216-0. [Online]. Available: http://dl.acm.org/citation.cfm?id=41765.41768.

[33] J. P. Bowen and M. G. Hinchey, "Ten commandments of formal methods", *Computer*, vol. 28, no. 4, pp. 56–63, Apr. 1995, ISSN: 0018-9162. DOI: 10.1109/2.375178.

[34] M. G. Hinchey and J. P. Bowen, "To formalize or not to formalize?", *IEEE Computer*, vol. 29, no. 4, pp. 18–19, Apr. 1996.

[35] M. Heisel, "A pragmatic approach to formal specification", in *Object-Oriented Behavioral Specifications*, Springer, Jan. 1, 1996, ISBN: 978-0-7923-9778-6. DOI: 10.1007/978-0-585-27524-6_4.

[36] R. Lai, "How could research on testing of communicating systems become more industrially relevant?", Springer, Jan. 1, 1996, pp. 3–13. DOI: 10.1007/978-0-387-35062-2_1.

[37] J. P. Bowen and M. G. Hinchey, "Ten commandments revisited: A ten-year perspective on the industrial application of formal methods", in *Proceedings of the 10th International Workshop on Formal Methods for Industrial Critical Systems*, ser. FMICS '05, Lisbon, Portugal: ACM, 2005, pp. 8–16, ISBN: 1-59593-148-1. DOI: 10.1145/1081180.1081183.

[38] D. Craigen, S. Gerhart, and T. Ralston, "An international survey of industrial applications of formal methods", in *Z User Workshop, London 1992: Proceedings of the Seventh Annual Z User Meeting, London 14–15 December 1992*, J. P. Bowen and J. E. Nicholls, Eds. London: Springer London, 1993, pp. 1–5, ISBN: 978-1-4471-3556-2. DOI: 10.1007/978-1-4471-3556-2_1.

[39] D. Craigen, "Formal methods technology transfer: Impediments and innovation (abstract)", in *CONCUR '95: Concurrency Theory: 6th International Conference Philadelphia, PA, USA, August 21–24, 1995 Proceedings*, I. Lee and S. A. Smolka, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995, pp. 328–332, ISBN: 978-3-540-44738-2. DOI: 10.1007/3-540-60218-6_24.

[40] D. Craigen, S. Gerhart, and T. Ralston, "Formal methods reality check: Industrial usage", *IEEE Transactions on Software Engineering*, vol. 21, no. 2, pp. 90–98, Feb. 1995, ISSN: 0098-5589. DOI: 10.1109/32.345825.

[41] B. A. Kitchenham and S. L. Pfleeger, "Guide to Advanced Empirical Software Engineering", in. Springer, 2008, ch. Personal Opinion Surveys, pp. 63–92.

[42] M. Gleirscher and A. Nyokabi, "Safety practice and its practitioners: Exploring and probing a diverse profession", *Information and Software Technology*, 2018, Under review. eprint: arxiv.

[43] Decision Analyst. (Aug. 2018). Technology Advisory Board, Decision Analyst, Inc., [Online]. Available: https://www.decisionanalyst.com/online/acop/.

[44] D. J. Leiner, "SoSci Survey", Tech. Rep., 2014. [Online]. Available: https://www.soscisurvey.de.

[45] K. A. Neuendorf, *The content analysis guidebook*, 2nd. Sage, Aug. 2016, ISBN: 9781412979474.

[46] Google. (Aug. 2018). Google Forms Service, Google, Inc., [Online]. Available: http://forms.google.com.

[47] The R Project. (Aug. 2018). R, The R Project, [Online]. Available: https://www.r-project.org.

[48] M. Gleirscher and D. Marmsoler, *Electronic supplementary material for "Formal methods: Oversold? Underused? A survey"*, Zenodo, Nov. 14, 2018. DOI: 10.5281/zenodo.1487596.

[49] Evans Data, "Global developer population and demographic study", Evans Data Corporation, Tech. Rep. Volume 1, 2018. [Online]. Available: https://evansdata.com/reports/viewRelease.php?reportID=9.

[50] R. Lai and W. Leung, "Industrial and academic protocol testing: The gap and the means of convergence", *Computer Networks and ISDN Systems*, vol. 27, no. 4, pp. 537–547, 1995. DOI: 10.1016/0169-7552(93)E0110-Z.

[51] F. Shull, J. Singer, and D. I. K. Sjøberg, Eds., *Guide to advanced empirical software engineering*. London: Springer, Oct. 2008.

[52] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in software engineering*. Springer, Jun. 2012, ISBN: 9783642290435.

[53] J. N. Oliveira, T. Aoki, M. Hinchey, J. Gibbons, and K. Taguchi. (Aug. 2018). Formal Methods Body of Knowledge (FMBoK), [Online]. Available: http://formalmethods.wikia.com/wiki/FMBoK.