Formal Methods in Dependable Systems Engineering: A Survey of Professionals from Europe and North America*

Preprint, compiled August 13, 2019

Mario Gleirscher[1] **and Diego Marmsoler**[2]

[1]Department of Computer Science, University of York,, Deramore Lane, Heslington, York YO10 5GH, United Kingdom
[2]Institut für Informatik, Technical University of Munich,, Boltzmannstraße 3, 85748 Garching, Germany

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 a more successful transfer into practice. *Objective:* We study the use of formal methods in mission-critical software domains, examining 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 the results also indicate a negatively perceived ease of use. Scalability, skills, and education seem to be among the key challenges to support this intent. *Conclusions:* We present the largest study of this kind so far ($N = 216$), and our observations provide valuable insights, highlighting directions for future theoretical and empirical research of formal methods.

***K*eywords** Formal methods · Empirical research · On-line survey · Usage · Usefulness · Practical challenges · Research transfer

1 Motivation and Challenges

Over the past decades, many software errors have been deployed in the field and some of these errors had a clearly intolerable impact.[2] This has been *the* motivation of formal methods (FM) as a first-class approach to error prevention, detection, and removal (Holloway, 1997).

In university courses on software engineering, we learned that FMs are among the best we have to design and assure correct systems. The question "Why are FMs not used more widely?" (Knight et al., 1997) is hence more than justified. With a Twitter poll,[3] emerged from our coffee spot discussions, we solicited opinions on a timely paraphrase of a statement argued by Holloway (1997): "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 1980s, FM researchers have started to examine this usefulness with the aim to respond to critical observations of practitioners (Jackson, 1987).

Hall (1990) and Bowen and Hinchey (1995a) 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 et al. (1997) examine reasons for the low adoption of FMs in practice. Barroca and McDermid (1992) ask: "To what extent should FMs form part of the [safety-critical SE] method?"

Glass (2002, pp. 148f, 165f) and Parnas (2010) observe that "many [SE] researchers advocate rather than investigate" by assuming the need for more methodologies. Glass summarises that FMs were supposed to help representing firm requirements concisely

[1]* This work is partly supported by the Deutsche Forschungsgemeinschaft (DFG) under the Grant no. 381212925.
Reference Format: Gleirscher, M., & Marmsoler, D. (2019). *Formal Methods in Dependable Systems Engineering: A Survey of Professionals from Europe and North America*. Technical report. Department of Computer Science, University of York, United Kingdom. arXiv: http://arxiv.org/abs/1812.08815 [cs.SE].

[2]See anecdotal evidence (grey literature, press articles) on software-related incidents, for example, by Kaner and Pels (1998, 2018), Charette (2018) and Neumann (2018).

[3]See https://twitter.com/MarioGleirscher/status/889737625178976256.

*correspondence: mario.gleirscher@york.ac.uk

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 (1992, p. 591), he notes that FMs have occasionally been sold as to make error removal complete, but there is no silver bullet (Glass, 2002, pp. 108f). Bad communication between theorists and practitioners sustains the issue that FMs are taught but rarely applied (Glass, 2002; Holloway and Butler, 1996, pp. 68ff). Parnas (2010) 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 et al. (2010) draw positive conclusions from recent applications of *model checking* and highlight lessons learned. In his keynote, O'Hearn (2018) conveys positive experiences in scaling FMs through adequate tool support for *continuous reasoning* in agile projects (see, e.g. Chudnov et al., 2018). Many researchers (see, e.g. Aichernig and Maibaum, 2003) have been working on the improvement of FMs towards their successful transfer. Boulanger (2012) and Gnesi and Margaria (2013) summarise 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. Pfleeger and Hatton, 1997; Sobel and Clarkson, 2002), often primarily conducted in research labs (e.g. Chudnov et al., 2018; Galloway et al., 1998 and many others). In late response to Holloway and Butler's request for empirical data (Holloway and Butler, 1996), Graydon (2015) still observes a lack of evidence for the effectiveness of FMs in assurance argumentation for safety-critical systems, suggesting 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 et al. (2015) identify questions and metrics for *FM productivity assessment*, supporting FM research transfer.

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

Overview. The next section introduces important terms. Section 3 relates our work to existing research. In Section 4, we explain our research design. We describe our data and answer our research questions in Section 5. In Section 6, we summarise and interpret our findings in the light of existing evidence and with respect to threats to validity. Section 7 highlights our conclusions and potential follow-up work.

2 BACKGROUND AND TERMINOLOGY

Formal Methods. By *formal methods*, we refer to *explicit* mathematical models and *sound* formal logical reasoning about *critical properties* (Rushby, 1994)—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 systems, including the use of notations (e.g. UML) and tools.

Tool and Method Evaluation. In the following, we give an overview of several evaluation approaches and explain in Section 4.2 which approach we take.

The widely used *technology acceptance model* (TAM; Davis, 1989) is a psychological test that allows the assessment of end-user IT based on the two constructs *perceived ease of use* (PEOU, i.e., positive and negative experiences while using an IT system) and *perceived usefulness* (PU, i.e., positive experiences of accomplishing a task using an IT system compared to not using this system for accomplishing the same task).

Complementary to TAM, Basili (1985) proposes the goal-question-metric (GQM) approach to method and tool evaluation. While GQM serves a as good basis for quantitative follow-up studies, we follow the user-focused TAM. Maturity models, such as CMMI (SEI, 2010), do not fit our purposes because they focus on engineering process improvement beyond particular development techniques. Poston and Sexton (1992) present tool survey guidelines based on technology-focused classification and selection criteria with a very limited view on tool usefulness and usability. Miyoshi and Azuma (1993) evaluate *ease of use* of development environments (i.e., specification and modelling tools) using metrics from the ISO/IEC 9126 quality model.

From comparing two models of predicting an individual's intention to use a tool, Mathieson (1991) supports TAM's validity and convenience but indicates its limits in providing enough information on users' opinions. For software methods and programming techniques, Murphy et al. (1999) show how surveys, case studies, and experiments can be used to compensate for this lack of information about usefulness and usability.

Because FMs are by definition based on a formal language and usually supported by tools, it is reasonable to adopt the TAM for the assessment of FMs. Unfortunately, the body of literature on the evaluation of FMs in TAM style is very small. However, Riemenschneider et al. (2002) apply TAM to methods (e.g. UML-based architecture design), concluding that "if a methodology is not regarded as useful by developers, its prospects for successful deployment may be severely undermined." Following their model, FM usage intentions would be driven by 1. an organisational mandate to use FMs, 2. the compatibility of FMs with how engineers perform their work, and 3. the opinions of developers' coworkers and supervisors toward using FMs. Overall, the application of TAM to FMs allows causal reasoning from FM user acceptance towards intention of FM use.

Specialising the approach in Riemenschneider et al. (2002), *ease of use* (EOU) of a FM characterises the type and amount of effort a user is likely to spend to learn, adopt, and apply this FM. *Usefulness* (U) determines how fit a FM is for its purpose, that is, how well 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)*. Together, PEOU and PU form the *user acceptance of a FM* and, by support of Mathieson (1991) and Riemenschneider et al. (2002), can predict the intention to use this FM.

Whereas TAM is a particular model based on the two user-focused constructs PEOU and PU, Kitchenham et al. (1997) propose a meta-evaluation approach called DESMET for tools and methods based on multiple performance indicators (e.g. with TAM as one of the indicators).

3 RELATED WORK

Table 1 shows a systematic map (Petersen et al., 2008) of 32 studies on FM research evaluation and transfer. For each study, we estimate the authors' attitude against or in favour of FMs, the motivation of the study, the approach followed, and the type of result obtained. Most of these works present personal experiences, opinions, case studies, or literature summaries. In contrast, the work presented in this paper focuses on the analysis of experience from a wide range of practitioners and experts. However, we found four similar studies.

Austin and Parkin (1993) tried to explain the low acceptance of FMs in industry around 1992. Using a questionnaire similar to ours with only open questions, they evaluated 111 responses 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 a second study in 2001, Snook and Harrison (2001) 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 et al. (2009) in 2009. They assess the state of the art of the application of FMs, using questionnaires to collect data on 62 industrial projects.

Liebel et al. (2016, pp. 102f) assess effects on and shortcomings of the adoption of model-based engineering (MBE) in embedded SE including a discussion of FM adoption. The authors observe a lack of tool support, bad reputation, and rigid development processes as obstacles to FM adoption. Their data suggests a need of FM adoption. 30% of the responses from industry declare the need for FMs as a reason to adopt MBE. Moreover, responses indicate that MBE adoption has a positive effect on FM adoption. One limitation of their study is the small number of responses from FM users.

While these studies focus on the elicitation of the state of the art, the main focus of our study is to compare the current FM adoption or use with the intention to adopt and use FMs in the future. To the best of our knowledge, our study offers the largest set of data points investigating the use of FMs in SE, so far. See Section 6.3 for a further discussion of how our findings relate to the findings of these studies.

4 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 (2008) for self-administered surveys and use our experience from a previous more general survey (Gleirscher and Nyokabi, 2018).

4.1 Research Goal and Questions

The questions in Section 1 have led to this survey on the *use, usage intent, and challenges of FMs*. Our interest is 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*?

Table 1: Overview of related work on FM use and adoption

Study	A	Motivation	Support	E	C	R
Surveys						
Austin and Parkin, 1993	=	LoEv	Interviews		•	•
Snook and Harrison, 2001	=	LoEv	Interviews	•		
Oliveira, 2004	=	Edu./Train.		•		
Woodcock et al., 2009[a]	=	LoEv	Interviews		•	
Liebel et al., 2016	+	LoEv		•		
Ferrari et al., 2019	+	TechTx	Literature study	•		
Literature Studies and Summaries						
Wing, 1990	+	SotA	O/E	•	•	
Bloomfield et al., 1991	=	SotA		•		
Fraser et al., 1994	=	TechTx				•
Heitmeyer, 1998	=	TechTx				•
Expert Opinions and Experience Reports						
Jackson, 1987	=	TechTx			•	
Bjorner, 1987	=	TechTx			•	
Barroca and McDermid, 1992	=	SotA			•	
Bowen and Hinchey, 1995a	+	Hyp. Testing				•
Bowen and Hinchey, 1995b	+	TechTx				•
Hinchey and Bowen, 1996	–	TechTx			•	
Heisel, 1996	+	TechTx				•
Holloway and Butler, 1996	+	LoEv			•	
Lai, 1996	+	TechTx			•	
Bowen and Hinchey, 2005	+	Hyp. Testing	Literature study			•
Parnas, 2010	=	TechTx			•	•
Case Studies and Experiments						
Hall, 1990	+	Hyp. Testing	O/E			•
Craigen et al., 1995[b]	+	SotA	Multiple cases, O/E	•		
Knight et al., 1997	=	TechTx	Field experiment	•		
Pfleeger and Hatton, 1997	=	Hyp. Testing	Effect analysis	•		
Sobel and Clarkson, 2002	=	Hyp. Testing	Lab experiment	•		
Miller et al., 2010	=	TechTx	Multiple cases, O/E	•		
Klein et al., 2018	+	TechTx		•		
Chudnov et al., 2018	=	TechTx		•		

[a]See also Bicarregui et al., 2009, [b]see also Craigen, 1995; Craigen et al., 1993; (A)ttitude, (E)valuation/analysis, (C)hallenges, (R)ecommendations, +/=/– . . . positive/neutral/negative, LoEv . . . lack of empirical evidence, O/E . . . opinion/experience report, SotA . . . state of the art, TechTx . . . technology transfer

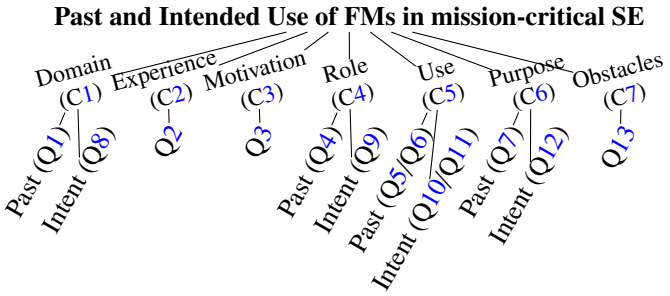


Figure 1: Construct

RQ2 Which *discrepancies* can we observe between FM users' *experience and intentions to use FMs*?

RQ3 How difficult do study participants perceive frequent FM *challenges*?

RQ4 What can we say about the *perceived ease of use* and the *perceived usefulness* of FMs?

4.2 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 2.

Id.	Description [Scale]
C1*	Application domains of FMs [MC among domains]
C2	Level of FM experience [duration ranges in years]
C3	Motivation to use FMs [degree per motivational factor]
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 freq. per purpose]
C7	Difficulty of obstacles to using FMs [degree per challenge]

Table 2: Concepts and scales for the construct in Figure 1

MC... multiple-choice, *... measured twice

Measuring Past and Intended Use. For RQ1 (UFM), we examine potential application *domains* for FMs (C1), *roles* when using FMs (C4), *motivations* and *purposes* of using FMs (C3, C6), and the extent of UFM at the general (C2) 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 the domain (C1), role (C4), FM class (C5), and purpose (C6). We measure UFM_i by relative frequency (Table 4) with respect to a participants' current situation, FM class, and purpose of use. This way, we reduce the burden on respondents to make comprehensive predictions of their professional future.

For RQ3, we measure the perception of difficulty of several obstacles (C7) known from the literature and from our experience.

Method Evaluation and TAM-style Interpretation. We follow DESMET (Kitchenham et al., 1997) and Murphy et al., 1999 insofar as we combine a *qualitative survey* (i.e., FM evaluation by SE practitioners and researchers) and a *qualitative effects analysis* based on the past and intent measurements for C6 (i.e., subjective assessment of effects of FMs by asking SE practitioners and researchers).

We assume UFM is roughly equivalent to the use of the tools automating the corresponding FMs. This assumption is justified inasmuch as for all FMs referred to in this survey, tools are available. In fact, in the past two decades (the period most survey respondents could have possibly used FMs), the development of a FM has mostly gone hand in hand with the development of its supporting tools.

For RQ4, we associate our findings from RQ2 and RQ3 with PEOU and PU. Whereas TAM predicts UFM_i of a specific tool by measuring PEOU and PU, we directly interrogate past (like in Mohagheghi et al., 2012, Fig. 2) and intended use of classes of FMs. We measure UFM_i (C1, C4, C5, C6) in more detail than TAM. Our approach relates to TAM for methods (Riemenschneider et al., 2002, Table 2) inasmuch as we collect data for PEOU through asking about expected obstacles to further use of FMs (C7) based on experience with past FM use (UFM_p). For this, respondents are asked to rate the difficulty of several known challenges to be tackled in typical FM applications. Furthermore, UFM_i is known to be correlated with PU. We then interpret the answers to RQ3 to examine the PEOU and, furthermore, interpret the answers to RQ2 to reason about PU. In Section 4.4, we discuss our questionnaire including the questions for measuring the sub-constructs.

4.3 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.

4.4 Survey Instrument: On-line Questionnaire

Table 3 summarises the questionnaire we use to measure UFM (Figure 1). The scales used for encoding the answers are described in Table 4.

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 15 minutes to complete the questionnaire.

Face and Content Validity. We derived answer options from the literature, our own experience, discussions with colleagues, pilot responses, and coding of open answers. 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

Table 3: Summary of questions from the questionnaire

Id.	Question or Question Template	Scale (see Table 4)	Sec.	Fig.
Q1	In which *application domains* (C1) in industry or academia have you mainly used FMs?	MC among domains	5.2	3
Q2	How many years of *FM experience* (including the study of FMs, C2) have you gained?	Duration range in years	5.2	4
Q3	Which have been your *motivations* (C3) to use FMs?	Degree per motivational factor	5.2	5
Q4	In which roles (C4) have you used FMs?	MC among roles	5.3	6
Q5	Describe your *level of experience* (C5) for ⟨*class of formal description techniques*⟩.	Level of experience per class	5.3	7
Q6	Describe your *level of experience* (C5) for ⟨*class of formal reasoning techniques*⟩.	Level of experience per class	5.3	8
Q7	I have mainly *used FMs for* (C6) ...	Absolute frequency per purpose	5.3	9
Q8	In which *domains* (C1) in industry or academia do you intend to use FMs?	MC among domains	5.4	10
Q9	In which *roles* (C4) would (or do) you intend to use FMs?	MC among roles	5.4	11
Q10	I (would) *intend to use* (C5) ⟨*class of formal description techniques*⟩ ⟨*this*⟩ often.	Relative frequency per class	5.4	12
Q11	I (would) *intend to use* (C5) ⟨*class of formal reasoning techniques*⟩ ⟨*this*⟩ often.	Relative frequency per class	5.4	13
Q12	I (would) intend to *use FMs for* (C6) ⟨*purpose*⟩.	Relative frequency per purpose	5.4	14
Q13	For any use of FMs in my future activities, I consider ⟨*obstacle*⟩ (C7) as ⟨*that*⟩ difficult.	Degree of difficulty per obstacle	5.5	15

MC. . . multiple-choice

Table 4: 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

bold. . . express lack of knowledge or indecision; (N)ominal, (O)rdinal, Ln . . . LIKERT-type scale with *n* values

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 4.6) supports us *with getting complete data points*, reducing the effort to deal with missing answers.

4.5 Data Collection: Sampling Procedure

We could not find an open/non-commercial panel of engineers. Large-scale *panel services* are either commercial (e.g. Decision Analyst, 2018) or they do not allow the sampling of software engineers (e.g. Leiner, 2014). 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
3. ask these people to disseminate our survey.

To check how well our *sample represents our targeted population*, we examine C2, C1, C4, and C5 from Table 2 for balanced levels.

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

Table 5: Channels used for sampling

4.6 Data Evaluation and Analysis

For RQ1, we summarise the data and apply descriptive statistics for categorical and ordinal variables in Section 5.3. We answer RQ2 by comparison of the data for the past and future views regarding the domain (C1), role (C4), FM class (C5), and purpose (C6) in Section 5.4. 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, practised 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 5.5. 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 (Neuendorf, 2016). For MC questions, we eliminate the choice of "I do/have not…" options from the data if ordinary answer options where also checked.

Tooling. We use Google Forms (Google, 2018) for implementing our questionnaire and for data collection (Section 4.5) and storage. For statistical analysis and data visualisation (Section 4.6), we use GNU R (The R Project, 2018) (with the packages `likert`, `gplots`, and `ggplot2` and some helpers from the "Cookbook for R" and the "StackExchange Stats" community[4]). Content analysis and coding takes place in a spreadsheet application. Electronic supplementary material to this work including the questionnaire is available in Gleirscher and Marmsoler, 2018.

5 EXECUTION, RESULTS, AND ANALYSIS

We describe the sample, summarise the responses to Table 3, and answer our research questions (Section 4.1).

5.1 Survey Execution

For data collection, we 1. advertised our survey on the channels in Table 5 and 2. personally invited > 30 persons. The sampling period lasted *from August 2017 til March 2019*. In this period, we *repeated step 1 up to three times* to increase the number of participants. Figure 2 summarises the distribution of responses. The channels in Table 5 particularly cover the European and North American areas.

5.2 RQ 1: Description of the Sample

Assuming participants are, on average, member of at least three of the channels listed in Table 5, 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

[4] See http://www.cookbook-r.com and https://stats.stackexchange.com.

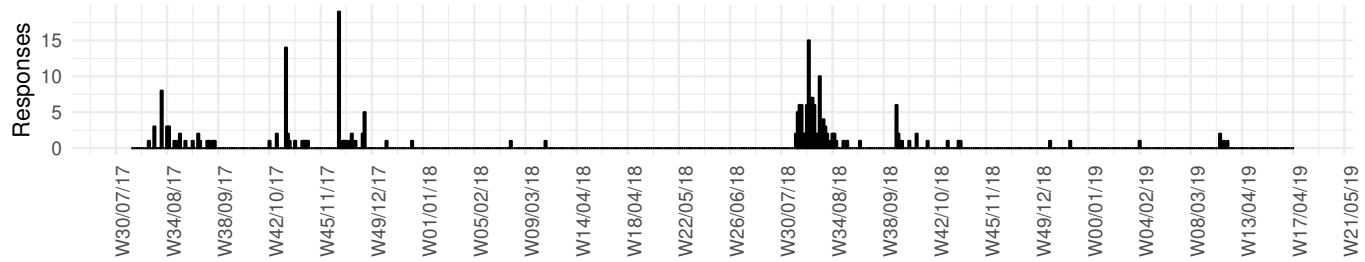


Figure 2: Distribution of responses over time

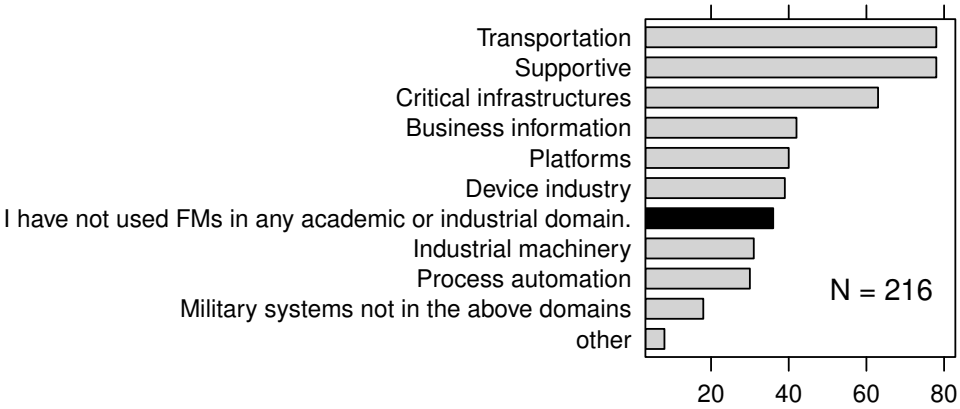


Figure 3: In which application domains in industry or academia have you mainly used FMs? (MC)

SE practitioners (Evans Data, 2018) and assuming that at least 1% are mission-critical SE practitioners, our *population* might comprise at least 230K persons.[5] We received $N = 216$ 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, the majority from the US, UK, Germany, France, and a sixth from other EU and non-EU countries.

Q1: Application Domain. For each domain, Figure 3 shows the number of participants having experience in that domain.[6] Note that 180 of the respondents do have experience with applying FM in different industrial contexts, while only 36 have not applied FMs to any application domain.

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 4.6, 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 (29 of those with even more than 25 years).

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 9 respondents, none of the given factors was motivating at all.

A further analysis of the study participants' experience profile is available from the supplementary material in Gleirscher and Marmsoler, 2018.

5.3 RQ 1: Facets of Formal Methods Use

Q4: Role. Figure 6 shows in which roles the respondents applied FMs. An analysis of the MC answers shows that 72% of the participants used FMs in an *academic environment*, as a researcher, lecturer, or student. 50% of the participants applied FMs in *practice*, as an engineer or consultant (see also Gleirscher and Marmsoler, 2018).

[5] An estimation in Gleirscher et al. (2019) suggests that about 5% of the overall ICT/IS developer population are embedded systems practitioners in critical and non-critical domains.

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

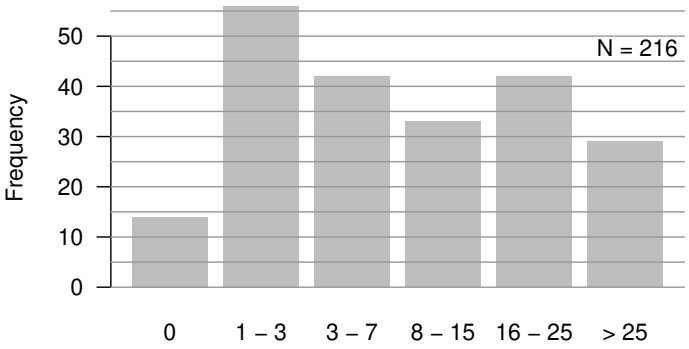


Figure 4: How many years of FM experience (including the study of FMs) have you gained?

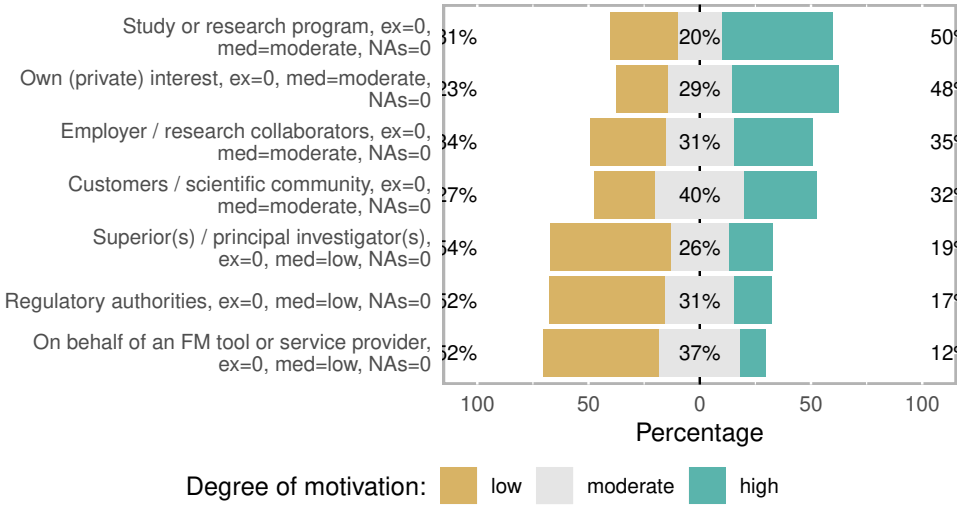


Figure 5: Which have been your motivations to use FMs?

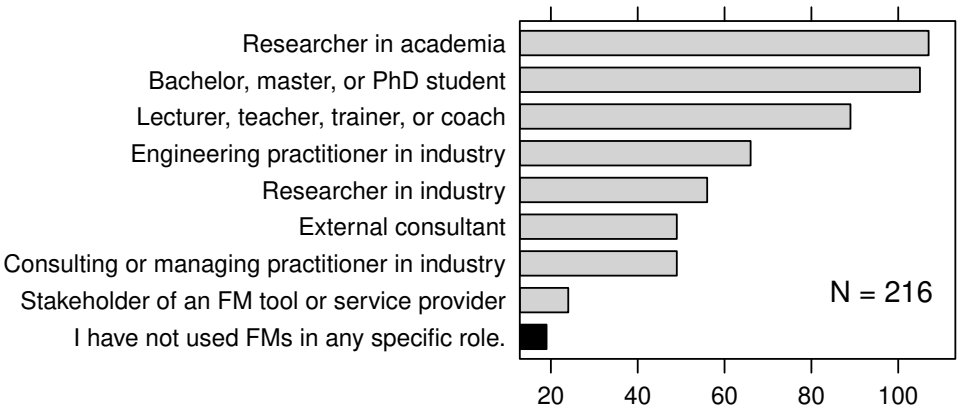


Figure 6: In which roles have you used FMs? (MC)

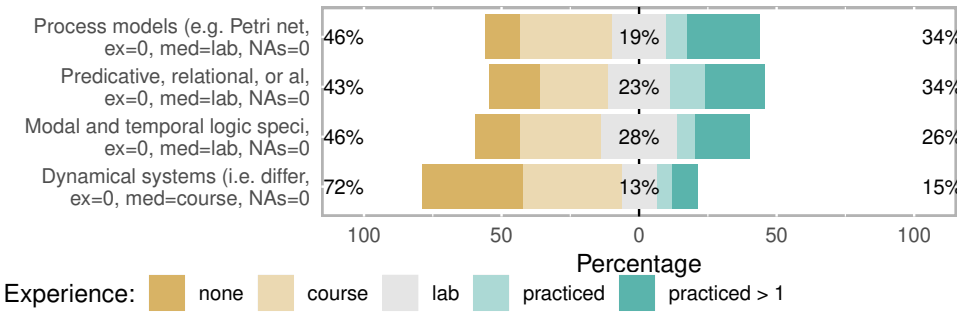


Figure 7: Describe your level of experience with each of the following classes of formal description techniques?

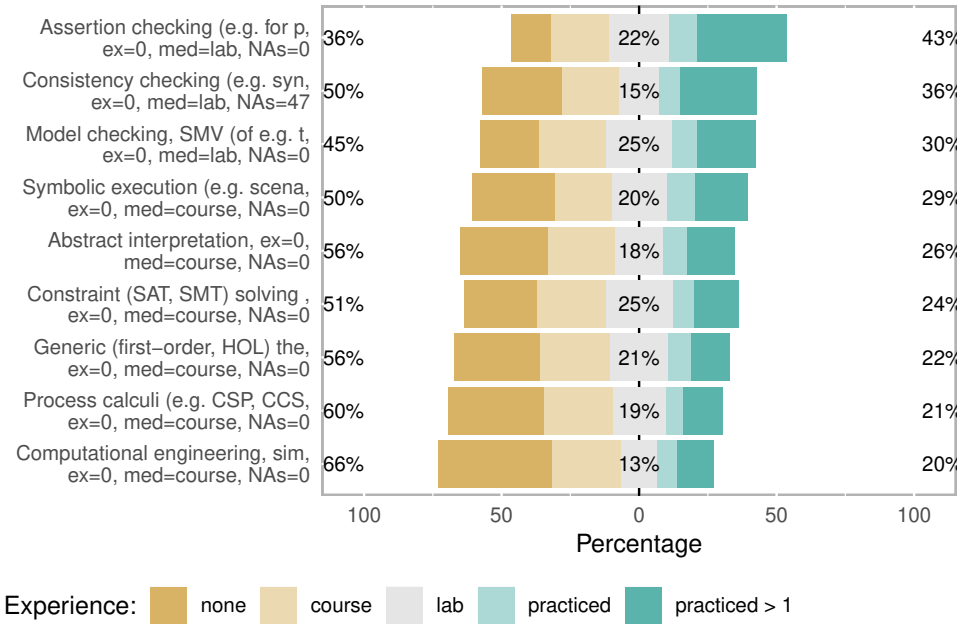


Figure 8: Describe your level of experience with each of the following classes of formal reasoning techniques?

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.

Q6: Use in Analysis. The use of FMs for analysis is depicted in Figure 8. Similar to specification techniques, we 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 5.3, the use of FMs for dynamical systems analysis, such as differential calculus, is again exceptionally low. One explanation for this is that our sample mainly comprises software and systems engineers who will work less intensively with this technique than, for example, mechanical or control engineers.

Q7: Purpose. Figure 9 depicts the participants' purposes to apply FMs. It seems that the respondents employ FMs mainly for assurance, specification, and inspection. Synthesis, on the other hand, to them seems to be only a subordinate purpose in the use of FMs.

5.4 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.

Application Domain. Figure 10 compares the respondents' current domain of FM application with their intended domain (see Q8). This figure reveals two insights into the participants' intentions to use FMs: (i) The number of participants

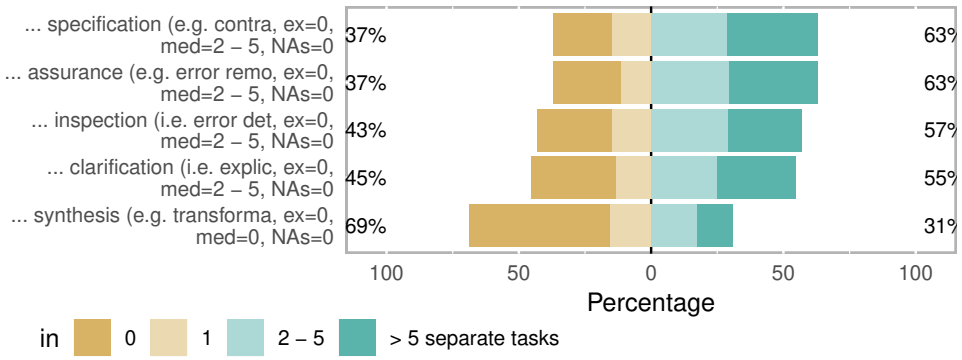


Figure 9: I have mainly used FMs for ...

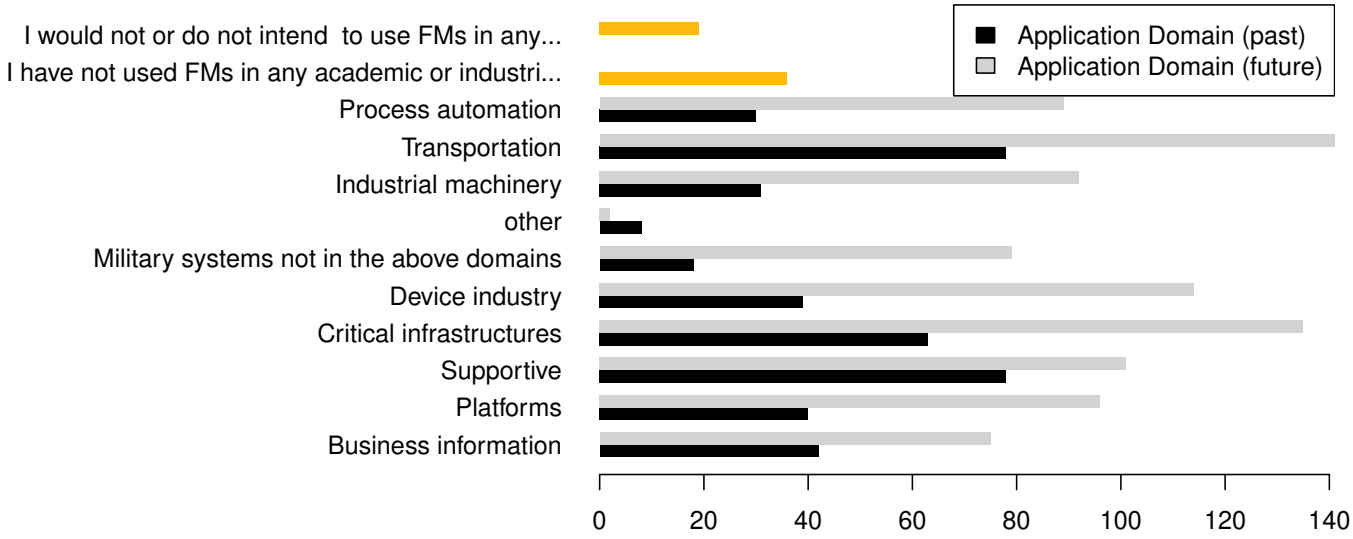


Figure 10: Number of respondents using FMs by domain (past vs. intent)

who have not yet used FMs is almost twice the number of participants who want to apply FMs in the future. Thus, some of the participants who have not used FMs so far, aim to apply them in the future. (ii) The intended application of FMs outperforms the current application of FMs across *all* domains. Hence, there is a tendency to increase the use of FMs across all application domains.

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 results for the application domain, we 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 that *academic* disciplines (i.e., researcher and lecturer) seem to be *stable*. There is only a small difference between the number of participants who applied FMs in 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.

Furthermore, the diagram shows a strong contrast between past and indented use in the category "Bachelor, master, or PhD student." We can see several reasons for this difference. From the respondents who "used FMs as a student," many (i) might not be able to "use FMs as a student" anymore because of having graduated, (ii) did not find FMs or the way FMs were taught as helpful, or (iii) moved into a business domain with no foreseeable demand for the application of FMs.

Q10: Intended use for Specification. Figure 12 depicts the respondents' intended *future* use of various FMs for system specification (i.e., formal description techniques). The figure shows an *almost equal* amount of participants aiming to decrease (i.e., "no more" and "less") and increase (i.e., "more often") their use of FMs for specification. Only *dynamical* system models again seem to be an exception: more participants want to decrease their use of this technology, compared to participants who want to increase it.

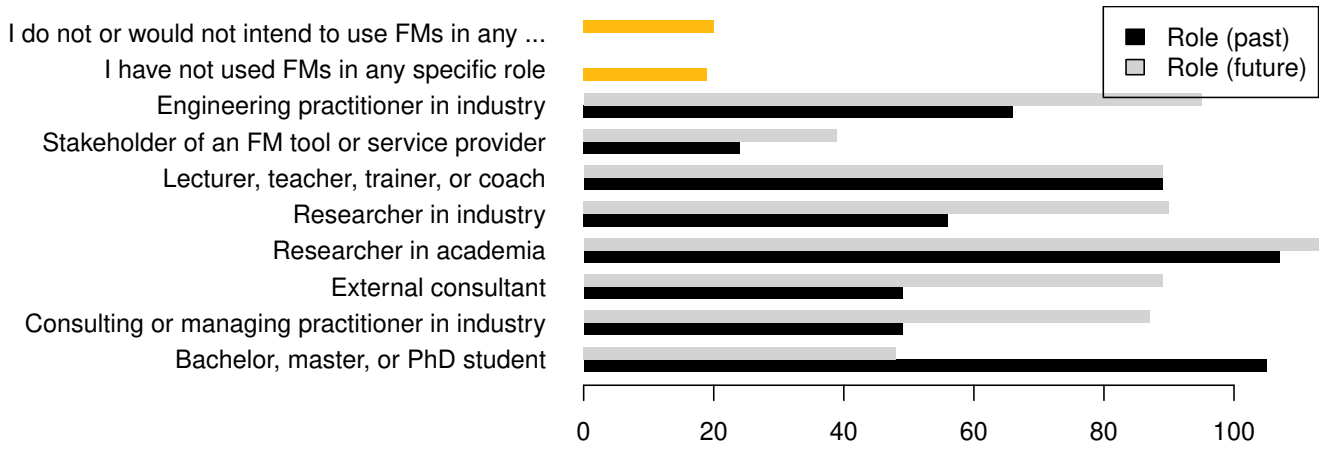


Figure 11: Number of respondents applying FMs by role (past vs. intent)

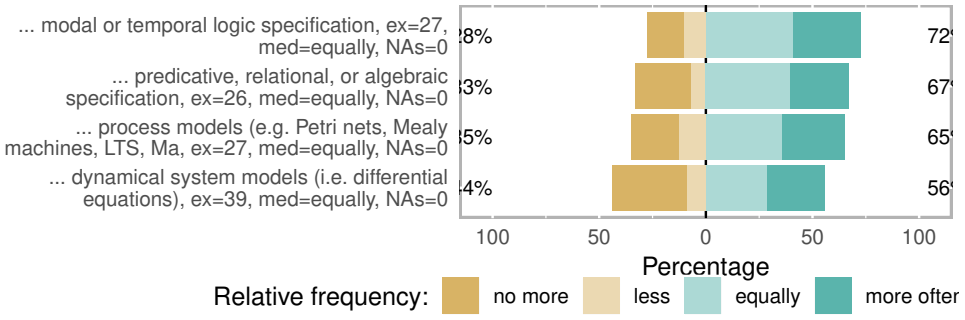


Figure 12: I (would) intend to use ...

Q11: Intended use for Analysis. The respondents' intended use of FMs for the analysis of specifications (i.e., formal reasoning techniques) is depicted in Figure 13. Except for process calculi, we observe a general tendency of the participants to *increase* their future FM use.

Q12: Intended Purpose. Figure 14 depicts the purpose why respondents intend to apply FMs. Again, there is a tendency of the participants to *increase* FM use across all listed purposes.

5.5 RQ 3: Perception of Challenges

Table 6 lists the FM challenges subject to discussion, their background, and literature referring to them. We apply the procedure described in Section 4.6.

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.

In the following, we compare specific groups of respondents by how they perceive the difficulty of the various challenges. We group respondents according to the criteria in Section 4.6 and according to the role, motivating factor, FM class, and purpose they specified.

Less Experienced (LE) versus More Experienced (ME) 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*.

Non-Practitioners (NP) versus Practitioners (P) 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*.

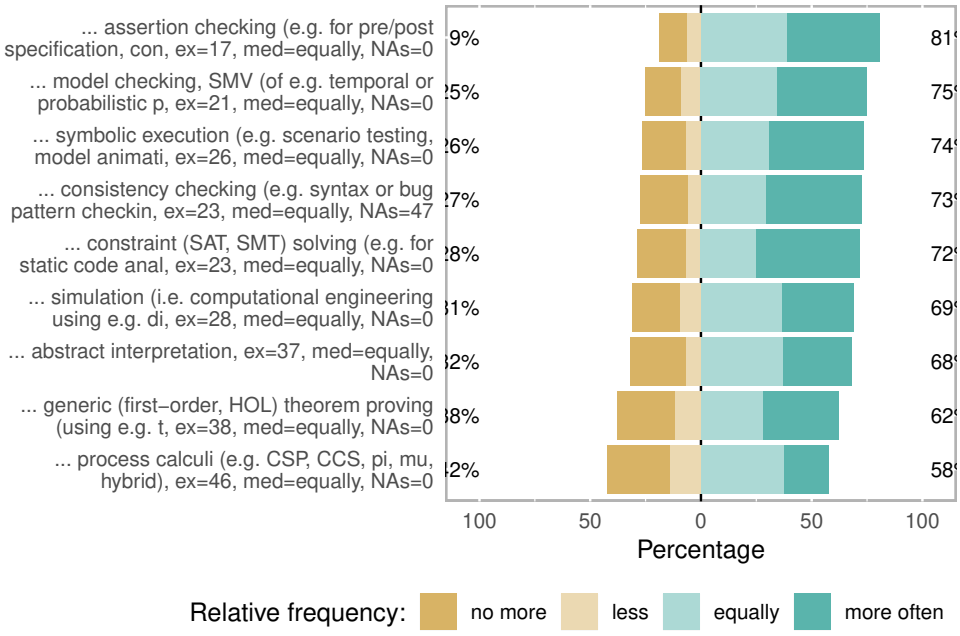


Figure 13: I (would) intend to use ...

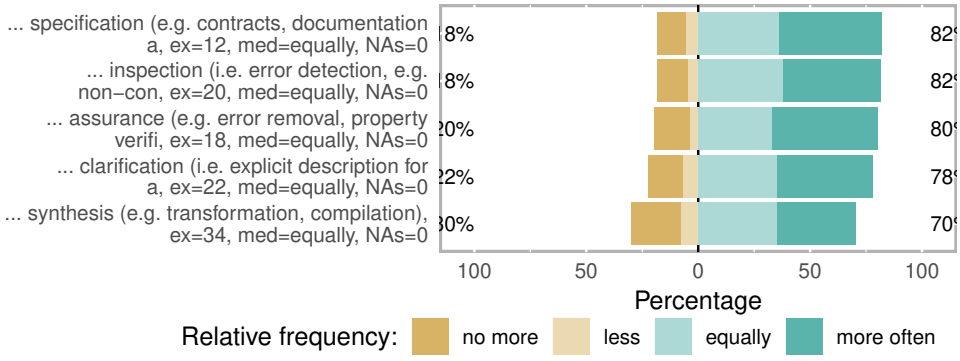


Figure 14: I (would) intend to use FMs for ...

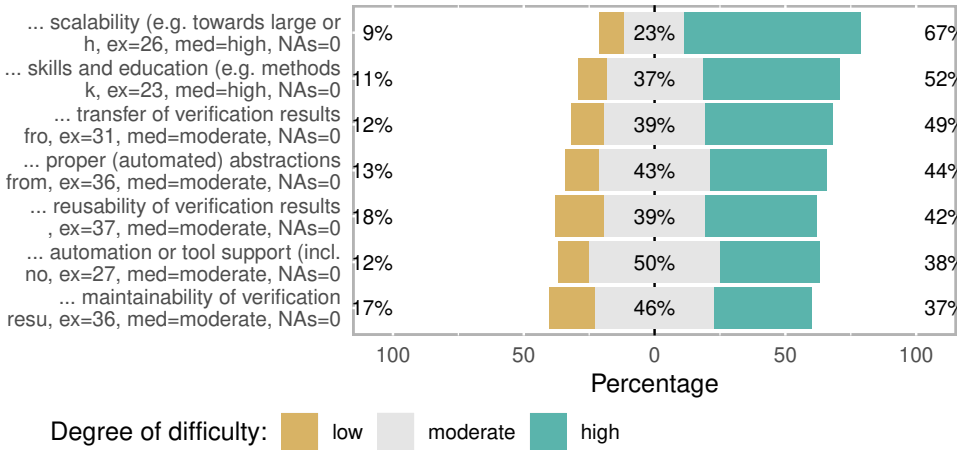


Figure 15: For any use of FMs in my future activities, I consider ⟨*obstacle*⟩ as [not an|a moderate|a tough] issue.

Table 6: Feedback on given and additional challenges

Challenge Name & Description	Src.	Supported by (oldest,newest)	Findings for RQ3 (Section 5.5)
Scalability: Useful in handling large and technologically heterogeneous systems	Q	7 studies, e.g. Hall, 1990; Miller et al., 2010	toughest in Figure 15; 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	12 studies, e.g. Bjorner, 1987, Bicarregui et al., 2009	2nd toughest; agreed by LEs and MEs; largely independent of FM class; comparatively small tough-proportions by Ms
Transfer of Proofs: Relation between models and reality (e.g. code), handling incomplete specifications	Q	8 studies, e.g. Jackson, 1987; Parnas, 2010	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	Barroca and McDermid, 1992; Bowen and Hinchey, 1995b	Top-rated by tool provider stakeholders and lectures
Abstraction: Useful and correct (automated) abstractions from irrelevant detail (for comprehension and validation)	Q	11 studies, e.g. Jackson, 1987; Miller et al., 2010; Parnas, 2010	Varies notably across FM classes
Tools & Automation: Useful notations and trustworthy tools (for manipulation, checking, collaboration, doc.)	Q	16 studies, e.g. Bjorner, 1987; O'Hearn, 2018	Top-rated by DIs; but comparatively small tough-proportions from practitioners
Maintainability: Stable proofs, easily modifiable specifications, and adaptable verification results	Q	Barroca and McDermid, 1992; Knight et al., 1997; Parnas, 2010	Comparatively small tough-proportions from practitioners
Resources: Sufficient resources, good cost-benefit ratio (despite adoption, training, licenses)	P (4)	11 studies, e.g. Hall, 1990; Woodcock et al., 2009	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)	12 studies, e.g. Bjorner, 1987; O'Hearn, 2018	
Practicality & Reputation: Benefit awareness and sufficient empirical evidence for benefits	P (7)	5 studies, e.g. Lai and Leung, 1995; Parnas, 2010	

Src.. . . source, Q . . . in questionnaire, P (*n*) . . . additionally raised by *n* participants

The leadership of scalability in Figure 15 comes along with most tough-ratings from NPs in *synthesis* and from Ps in *assurance* and *clarification*.

Decreased (DI) versus Increased Intent (II) by Purpose (Q12). The comparison of the difficulty ratings of respondents with no or decreased intent to use FMs *for a specific purpose* and of respondents with equal or increased intent shows: • The leadership of *scalability* and *skills and education* in Figure 15 comes along with most tough-ratings from IIs for *assurance* (67%) and *inspection* (66%) and from DIs for *synthesis* (53%). • 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*.

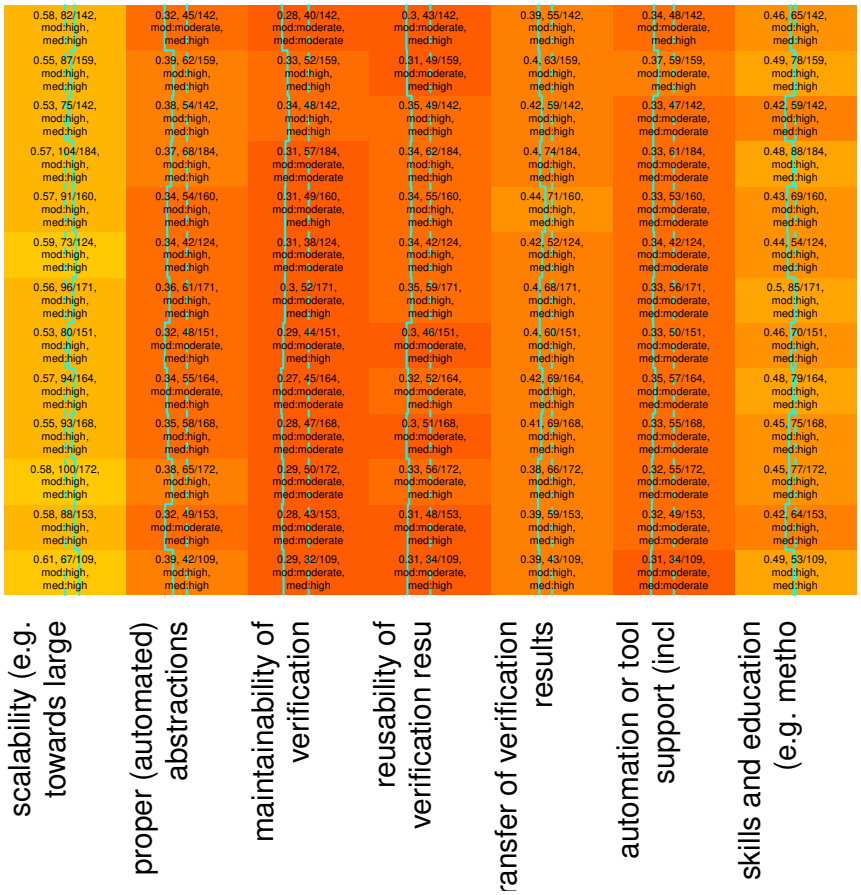
Non-Practitioners (NP) versus Practitioners (P) by FM Class (Q5, Q6). Figure 16 (top) shows for NPs, the trend in Figure 15 is **largely independent of the FM class**, except for *consistency checking* and *logic* leading with *tough* proportions of 49%. See Section 4.6 for a legend to the data in the diagram cells.

Figure 16 (bottom) shows 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 centred on moderate or tough, *proper abstraction* and *skills and education* show a comparatively wide variety across FM classes.

The histogram boxes in the lower right corners in Figure 16 indicate that • NPs' difficulty ratings vary less than Ps' ratings, • NPs' ratings are more independent from the FM classes, and • NPs' difficulty ratings are lower on average than Ps' ratings.

Decreased (DI) versus Increased Intent (II) by FM Class (Q10, Q11). The trend in Figure 15 comes along with many tough ratings (48%) for *transfer of verification results* from DIs of *consistency checking*. However, DIs of *process calculi* provide

Comparison of Challenge Difficulty across FMs (users not practicing FMs, past)



scalability (e.g. towards large

proper (automated) abstractions

maintainability of verification

reusability of verification resu

ransfer of verification results

automation or tool support (incl

skills and education (e.g. metho

Predicative, relational, or algebraic sp

Modal and temporal logic specification]

Process models (e.g. Petri nets, Mealy m

Dynamical systems (i.e. differential equ

Abstract interpretation]

Assertion checking (e.g. for pre/post sp

Process calculi (e.g. CSP, CCS, pi, mu,

Model checking, SMV (of e.g. temporal or

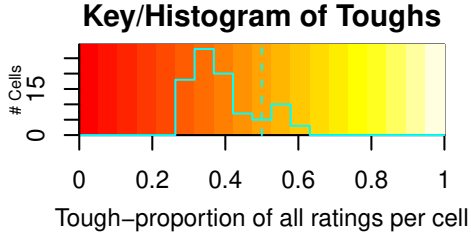
Constraint (SAT, SMT) solving (e.g. for

Generic (first–order, HOL) theorem provi

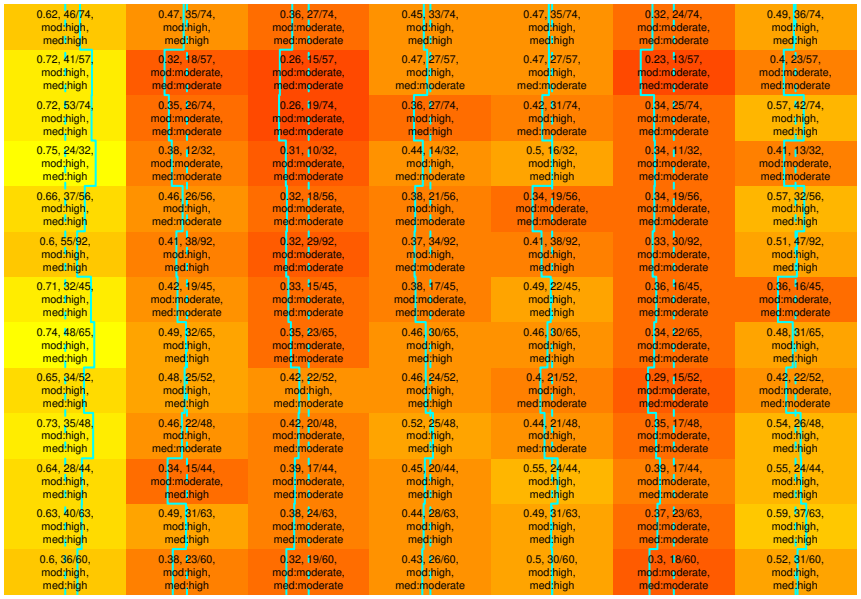
Computational engineering, simulation (u

Symbolic execution (e.g. scenario testin

Consistency checking (e.g. syntax or bug



Comparison of Challenge Difficulty across FMs (users practicing FMs, past)



scalability (e.g. towards large

roper (automated) abstractions

maintainability of verification

reusability of verification resu

nsfer of verification results

automation or tool support (incl

kills and education (e.g. metho

Predicative, relational, or algebraic sp

Modal and temporal logic specification]

Process models (e.g. Petri nets, Mealy m

Dynamical systems (i.e. differential equ

Abstract interpretation]

Assertion checking (e.g. for pre/post sp

Process calculi (e.g. CSP, CCS, pi, mu,

Model checking, SMV (of e.g. temporal or

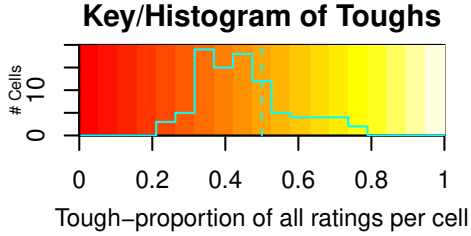
Constraint (SAT, SMT) solving (e.g. for

Generic (first–order, HOL) theorem provi

Computational engineering, simulation (u

Symbolic execution (e.g. scenario testin

Consistency checking (e.g. syntax or bug



comparatively many tough-ratings (39%) for the generally low-ranked *automation and tool support*. *Assertion checking* exhibits comparatively low tough-proportions across all challenges whereas *process calculi* exhibit comparatively high tough-ratings. Mirroring the trend in Figure 15, IIs show less variance than DIs across all FM classes.

Unmotivated (U) versus Motivated (M) respondents by Motivating Factor (Q3). Respondents with moderate to strong motivation to use FMs more likely identify given challenges as moderate or tough, **regardless of the motivating factor.** The trend in Figure 15 comes along with many tough ratings from respondents motivated by *regulatory authorities* (69%) or not motivated by *tool providers* (56%) or *superiors/principal investigators* (56%). Us' *tough*-ratings are **notably lower than Ms'**.

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. Respondents not having used FMs or not planning to use FMs exhibit the lowest tough-ratings but also the highest fractions of *dnk*-answers.

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.

6 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.

6.1 Findings and their Interpretation

The following (F)indings are based on the data summarised and analysed in the Sections 5.2 to 5.4.

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, we observe that this tendency is *independent* of the particular *FM class* (except process calculi) or *purpose*. The data also suggest that the use of FMs in research is saturated, while there is an increased intent to apply FMs in *industrial contexts* in the future. **(F3)** Our data suggest that experience in using a certain FM class is positively associated with the intent to use this FM class in the future. To investigate this suspicion, we analysed the intended use of a FM class based on the experience of participants in using this class (also by association analysis, Section 4.6, Gleirscher and Marmsoler, 2018). We observed that the *more experience* one has with using a specific FM class, 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 class and only little experience with a certain FM class *significantly increases the willingness to apply it in the future*. Similar observations can be made for the use of FMs 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 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)** FM users with decreased usage intent rate *tool deficiencies* as their top obstacle to FM adoption. **(F8)** Furthermore, our respondents raised three additional challenges which we cross-validated with the literature (see highlighted rows in Table 6). The fact that these obstacles were mentioned several times in addition to the given obstacles justifies them to be highly relevant and at least moderate. However, our data does not allow to rank them more precisely. **(F9)** Challenges are perceived *as moderate or tough*, largely similar between the pairs of groups we distinguish in Section 4.6. **(F10)** Process models (e.g. Petri nets, Mealy machines, LTS, Markov 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.

6.2 RQ 4: Relationship to TAM for Methods

In analogy to the reasoning in Davis (1989), an increased positive experience with practically applying FMs forms a high degree of PU (Section 2). Davis (1989, pp. 329,331) observed that current and intended usage are significantly correlated with PU, less with PEOU. In fact, F2 suggests an increased intent to use FMs in the future. **(F11)** Because the use of FMs is not mandatory for most respondents, a likely explanation for an increased intent (UFM_i) is that our respondents perceive the *usefulness of FMs* more positive than negative.

Table 7: Summary of findings per research question

RQ1: In which typical domains, for which purposes, in which roles, and to what extent have *FMs been used*?	
F1	Intrinsic motivation to use FMs is stronger than regulatory authorities.
RQ2: Which *discrepancies* can we observe between FM users' *experience and intentions to use FMs*?	
F2	Increased intent to use FMs observable across all application domains.
F3	Amount of experience is positively associated with the strength of usage intent.
RQ3: How difficult do study participants perceive frequent FM *challenges*?	
F4	Scalability and skills & education lead the challenge difficulty ranking.
F5	Maintainability of proof results is found to be the least worrying challenge.
F6	Reusability of proof results is rated as tough by several practitioner groups.
F7	FM users with decreased usage intent rate *tool deficiencies* as their top obstacle.
F8	Respondents identified resources, process compatibility, and reputation as further obstacles.
F9	All challenges are generally perceived as moderate or tough.
F10	Among all FM classes, process models are most strongly positively associated with scalability.
RQ4: What can we say about the *perceived ease of use* and the *perceived usefulness* of FMs?	
F11	Respondents perceive the usefulness of FMs as mainly positive and intend to increase their use.
F12	Respondents perceive the ease of use of FMs as mainly negative.
Relationship to Existing Evidence (from the literature):	
F13	Proof maintainability and reusability are least covered by the literature.
F14	We repeat Austin and Parkin (1993), excluding benefit analysis but with broader sample and more detailed questions.

We justify the validity of the FM-specific scale used for EOU and U using our literature review. From the literature, we know that tackling the challenges mentioned in Table 6 contributes to an increased EOU and U. First, the studies suggest that FMs are *easier to use* if users have sufficient skills and education, if the methods scale to large systems, if mature tool and automation is available, and if proofs are easily maintainable and reusable. Second, the studies suggest that FMs are *more useful* if they are compatible with the process, if their cost-benefit ratio is low, if their abstractions are correct and expressive, and if proofs can be correctly transferred to reality. Hence, these challenges represent FM-specific substrata (Davis, 1989, p. 325) of EOU and U for FMs. Moreover, a high degree of PEOU corresponds to an increased positive user experience with FMs which translates to a low proportion of tough ratings for the obstacles measured in Q13. However, from F9, we observe that respondents rate most challenges as moderate to tough, largely independent of other variables (F4). **(F12)** Thus, it seems that our respondents perceive the *ease of use of FMs* more negative than positive.

6.3 Relationship to Existing Evidence

Our systematic map shows that our list of challenges is completely backed by substantial literature (see Table 6) raising and discussing these challenges. **(F13)** 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 hand, not with F6 and typical cultures of reuse in practice.

Beyond the general findings about FM benefits in Austin and Parkin (1993), we steered our half-open questionnaire towards a refined classification of responses, comparing past with intended use, and interrogating recently perceived obstacles among a methodologically and geographically more diverse sample. Their sample mainly covers Z and VDM users in the UK. Our questionnaire has less focus on representation and methodology and excludes questions on benefits, suggestions, for example, improve education and standardisation and to perform case studies and define effectiveness metrics.

(F14) We received the paper report of Austin and Parkin (1993) after conducting this survey and conclude that our work forms a near-replication. The data from Figure 15 and table 6 confirms that many of the obstacles (i.e., limitations and barriers) they identified back in 1991/2 remain (e.g. understanding the notation and the underlying mathematics, resistance to process changes), some have been lightly addressed (e.g. lack of cost/benefit evidence) and some have been more strongly addressed (e.g. lack of expressiveness, lack of appropriate tools). Not mentioned in Austin and Parkin (1993) is scalability, rated by our respondents to be the toughest obstacle.

F3 is in line with other observations in Bicarregui et al. (2009) and Woodcock et al. (2009) 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 (Miller et al., 2010).

6.4 Threats to Validity

We assess our research design with regard to four common criteria (Shull et al., 2008; Wohlin et al., 2012). Per threat (↯), we estimate its criticality (minor or major), describe it, and discuss its mitigation (✓).

Construct Validity. Why would the construct (Section 4.1) appropriately represent the phenomenon?

maj ↯: Wrong or omitted questions / To support *face validity*, we applied our own experience from FM use to develop a core set of questions. For the design of our questionnaire, we use feedback from colleagues, from respondents we personally know, and from the general feedback on the survey to improve and support *content validity*. A positive comparison with the questionnaire in Austin and Parkin (1993) finally confirms the appropriateness of our questions. ✓

min ↯: Questionnaire limited for measurement of PEOU (e.g. per FM class) and PU / We avoid deriving conclusions specific to a FM or FM tool 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. Between 8% and 40% of the respondents made use of the text field "Other." Our systematic map confirms that we have not listed unknown challenges in Q13. We identified three additional challenges via open answers and the literature. We believe to have achieved good *criterion validity* through questions and scales for distinguishing important sub-groups (see Section 4.6) of our population. ✓

min ↯: Educational background asked indirectly / We approximate what we need to know by using data from Q1, Q3, Q4, and Q5. ✓

Internal Validity. Why would the procedure in Section 4 lead to reasonable and justified results?

min ↯: Incomplete data points / After the 47th response, feedback from colleagues and respondents resulted in an extension of Q3 with the option "on behalf of FM tool provider" (Figure 5) and of Q6 and Q11 with the option "consistency checking" (Figure 8). The enhancement of 169 complete data points to 216 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 (90/220) indicate only 4 double participants. We remove these 4 data points from our data set. Google Forms includes data points only if all mandatory questions are answered and the submit button is pressed. We also performed a consistency check of MC questions and corrected 5 data points where "I do/have not…" was combined with other checked options. ✓

External Validity. Why would the procedure in Section 4 lead to similar results with more general populations?

maj ↯: Low response rate / We believe our estimates in Section 5.2 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. Some uncertainties remain, for example, lack of sympathy, personal motivation, and interest, or strong loyalty, and high expectations in the outcome, or intentional bias. ✓

maj ↯: Bias towards specific groups (Shull et al., 2008, p. 181) / We distributed our questionnaire over general SE channels. We mix opportunity (only 5 to 10% chain referral), volunteer, and cluster-based sampling. Reduced social visibility and wrong control of referrals might, therefore, not affect selection bias. Our sample includes 50% of practitioners according to Section 4.6, ≈ 21% of NPs (incl. laypersons), and ≈ 31% 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 generalised to, e.g. finance and election sectors. ✓

min ↯: Non-response / We decided not to enforce responses or provide incentives. Our data suggests that our advertisement nevertheless stimulated responses from FM-critical minds. ✓

min ↯: Lack of FM knowledge / 11 to 18% 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 generalisability of our conclusions, e.g. to ecosystems such as China, India, or Brazil. ✓

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

maj ↯: Internal Consistency / All 7 items for the concept "obstacle to FM effectiveness" (C7) show good internal consistency for our sample with a CRONBACH $\alpha = 0.84$, the PEOU-part of C7 consisting of 5 items shows an $\alpha = 0.79$ (Shull et al., 2008). The other concepts are not measured with multiple items. ✓

maj ↯: Change of proportions / The limited sample and the low response rate make it hard to mitigate this risk. However, we compared the first (til 4.8.2018, $N_1 = 114$) and second (from 5.8.2018, $N_2 = 102$) half of our sample to simulate a repetition of our survey with the same questionnaire. A two-sided Mann-Whitney U test for difference does not show a significant difference

between these two groups (e.g. for Q13 and Q4). Only for the Q3 item "On behalf of FM tool provider," a $p = 0.07$ indicates a potential difference. This difference might stem from the fact that we added this answer option only after 47th respondent. ✓

7 CONCLUSIONS

We conducted an on-line survey of mission-critical software engineering practitioners and researchers to examine how formal methods have been used, how these professionals intent to use them, and how they perceive challenges in using them. This study aims to contribute to the body of knowledge of the software engineering and formal methods communities.

Overall Findings. From the evidence we gathered for the use of formal methods, we make the following observations:

- *Intrinsic motivation* is stronger than the regulatory one.
- Despite the challenges, our respondents show an *increased intent* to use FMs in industrial contexts.
- Past experience is *correlated* with usage intent.
- 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*.
- The negative responses to the questions about obstacles to FM effectiveness suggest 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 perception of usefulness of this FM*.

In response to Barroca and McDermid (1992), we believe that formal methods are more underused than oversold. However, our data suggests that these methods need improvement and support in several areas in order for their benefits to be better utilised.

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 et al. (2015), we want to make another step in setting out an agenda for future FM research.

To address *controllable abstractions*, we need semantics workbenches for underpinning domain-specific languages with formal semantics. We believe that further steps in *theory integration and unification* (Gleirscher et al., 2019) can help establish proof hierarchies and, hence, *reusability* and *proof 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. Chudnov et al., 2018; O'Hearn, 2018) and in cost-savings analyses of FM applications (e.g. Jeffery et al., 2015). 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; Oliveira et al., 2018) with revised recommendations for lecture material (Oliveira, 2004), for example, the teaching of modelling, 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." Hence, we aim at a follow-up study • to measure key indicators for successful use of FMs, • to measure what techniques work well where, • to identify management techniques needed to accommodate the changes in working practices, and, finally, • to provide guidance to future projects wishing to adopt FMs.

In a next survey, we like to ask about typical FM benefits, pose more specific questions on scalability and useful abstraction, and the geographical[7] 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.

Acknowledgements. 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 supported us to critically reflect our work and relate it to existing evidence. We are also grateful to John Fitzgerald and John McDermid for helpful feedback and for encouraging us to do further research in this direction. We would like to spend sincere gratitude to Krzysztof Brzezinski, Louis Brabant, and Emmanuel Eze for pointing us to several related works.

REFERENCES

Aichernig, Bernhard K. and Tom Maibaum, eds. (2003). *Formal Methods at the Crossroads. From Panacea to Foundational Support*. Springer Berlin Heidelberg. ISBN: 3-540-20527-6.

Austin, Stephen and Graeme Parkin (1993). *Formal methods: A survey*. Tech. rep. Teddington, Middlesex, UK: National Physical Laboratory.

Barroca, Leonor M. and John A. McDermid (1992). "Formal methods: Use and relevance for the development of safety-critical systems". In: *Comp. J.* 35.6, pp. 579–99. DOI: 10.1093/comjnl/35.6.579.

Basili, Victor R. (1985). *Quantitative evaluation of software methodology*. Tech. rep. URL: https://drum.lib.umd.edu/bitstream/handle/1903/7520/Quantitative+Evaluation.pdf?sequence=1 (visited on 05/30/2019).

Bicarregui, J. C. et al. (2009). "Industrial Practice in Formal Methods: A Review". In: *FM 2009: Formal Methods*. Ed. by Ana Cavalcanti and Dennis R. Dams. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 810–813. ISBN: 978-3-642-05089-3.

Bjorner, D. (1987). "On the Use of Formal Methods in Software Development". In: *Proceedings of the 9th International Conference on Software Engineering*. ICSE '87. Monterey, California, USA: IEEE Computer Society Press, pp. 17–29. ISBN: 0-89791-216-0. URL: http://dl.acm.org/citation.cfm?id=41765.41768.

Bloomfield, RE, PKD Froome, and BQ Monahan (1991). "Formal methods in the production and assessment of safety critical software". In: *Reliability Engineering & System Safety* 32.1-2, pp. 51–66.

Boulanger, Jean-Louis (2012). *Industrial Use of Formal Methods: Formal Verification*. Wiley-ISTE. 298 pp. ISBN: 9781848213630.

Bowen, J. P. and M. G. Hinchey (1995a). "Seven more myths of formal methods". In: *IEEE Software* 12.4, pp. 34–41. ISSN: 0740-7459. DOI: 10.1109/52.391826.

— (1995b). "Ten commandments of formal methods". In: *Computer* 28.4, pp. 56–63. ISSN: 0018-9162. DOI: 10.1109/2.375178.

Bowen, Jonathan P. and Michael G. Hinchey (2005). "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*. FMICS '05. Lisbon, Portugal: ACM, pp. 8–16. ISBN: 1-59593-148-1. DOI: 10.1145/1081180.1081183.

Charette, Robert N. (2018). *Fiat Chrysler Is Being Sued Over a Software Flaw*. IEEE. URL: https://spectrum.ieee.org/riskfactor/computing/software/court-allows-lawsuit-to-proceed-against-fiat-chrysler-over-software-flaw.

Chudnov, Andrey et al. (2018). "Continuous Formal Verification of Amazon s2n". In: *Computer Aided Verification*. Springer International Publishing, pp. 430–446. DOI: 10.1007/978-3-319-96142-2_26.

Craigen, D., S. Gerhart, and T. Ralston (1995). "Formal methods reality check: industrial usage". In: *IEEE Transactions on Software Engineering* 21.2, pp. 90–98. ISSN: 0098-5589. DOI: 10.1109/32.345825.

Craigen, Dan (1995). "Formal methods technology transfer: Impediments and innovation (abstract)". In: *CONCUR '95: Concurrency Theory: 6th International Conference Philadelphia, PA, USA, August 21–24, 1995 Proceedings*. Ed. by Insup Lee and Scott A. Smolka. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 328–332. ISBN: 978-3-540-44738-2. DOI: 10.1007/3-540-60218-6_24.

Craigen, Dan, Susan Gerhart, and Ted Ralston (1993). "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*. Ed. by J. P. Bowen and J. E. Nicholls. London: Springer London, pp. 1–5. ISBN: 978-1-4471-3556-2. DOI: 10.1007/978-1-4471-3556-2_1.

Davis, Fred D. (1989). "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology". In: *MIS Quarterly* 13.3, pp. 319–40.

Decision Analyst (2018). *Technology Advisory Board*. Decision Analyst, Inc. URL: https://www.decisionanalyst.com/online/acop/.

Evans Data (2018). *Global Developer Population and Demographic Study*. Tech. rep. Volume 1. Evans Data Corporation. URL: https://evansdata.com/reports/viewRelease.php?reportID=9.

Ferrari, Alessio et al. (2019). *Survey on Formal Methods and Tools in Railways Technical Report on the activities performed within ASTRail, Deliverable D4.1*. DOI: 10.5281/zenodo.2535023.

Fraser, Martin D., Kuldeep Kumar, and Vijay K. Vaishnavi (1994). "Strategies for incorporating formal specifications in software development". In: *Communications of the ACM* 37.10, pp. 74–86. DOI: 10.1145/194313.194399.

Galloway, Andy J., Trevor J. Cockram, and John A. McDermid (1998). "Experiences with the Application of Discrete Formal Methods to the Development of Engine Control Software". In: *IFAC Proceedings Volumes* 31.32, pp. 49–56. DOI: 10.1016/S1474-6670(17)36335-8.

Glass, Robert L. (2002). *Facts and Fallacies of Software Engineering*. Pearson Education (US). ISBN: 978-0321117427.

Gleirscher, Mario and Diego Marmsoler (2018). *Electronic Supplementary Material for "Formal Methods: Oversold? Underused? A Survey"*. Zenodo. DOI: 10.5281/zenodo.1487596.

Gleirscher, Mario and Anne Nyokabi (2018). *System Safety Practice: An Interrogation of Practitioners about Their Activities, Challenges, and Views with a Focus on the European Region*. Tech. rep. York, UK: Department of Computer Science, University of York, UK. arXiv: 1812.08452 [cs.SE].

[7] According to https://en.wikipedia.org/wiki/United_Nations_geoscheme.

Gleirscher, Mario, Simon Foster, and Jim Woodcock (2019). *New Opportunities for Integrated Formal Methods*. Unpublished working paper. York, UK: Department of Computer Science, University of York. arXiv: 1812.10103 [cs.SE]. Accepted subject to minor revisions for ACM Computing Surveys.

Gnesi, Stefania and Tiziana Margaria (2013). *Formal Methods for Industrial Critical Systems: A Survey of Applications*. Wiley-IEEE Press. ISBN: 9781118459898.

Google (2018). *Google Forms Service*. Google, Inc. URL: http://forms.google.com.

Graydon, P.J. (2015). "Formal Assurance Arguments: A Solution in Search of a Problem?" In: *Dependable Systems and Networks (DSN), 2015 45th Annual IEEE/IFIP International Conference on*, pp. 517–528. DOI: 10.1109/DSN.2015.28.

Hall, Anthony (1990). "Seven Myths of Formal Methods". In: *IEEE Software* 7.5, pp. 11–19. DOI: 10.1109/52.57887.

Heisel, Maritta (1996). "A Pragmatic Approach to Formal Specification". In: *Object-Oriented Behavioral Specifications*. Springer. ISBN: 978-0-7923-9778-6. DOI: 10.1007/978-0-585-27524-6_4.

Heitmeyer, Constance L. (1998). "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, pp. 18–26.

Hinchey, Michael G. and Jonathan P. Bowen (1996). "To formalize or not to formalize?" In: *IEEE Computer* 29.4, pp. 18–19.

Holloway, C. M. (1997). "Why engineers should consider formal methods". In: *16th DASC. AIAA/IEEE Digital Avionics Systems Conference. Reflections to the Future. Proceedings*. Vol. 1, pp. 16–22. DOI: 10.1109/DASC.1997.635021.

Holloway, C. M. and R. W. Butler (1996). "Impediments to industrial use of formal methods". In: *Computer* 29.4, pp. 25–26. DOI: 10.1109/MC.1996.488298.

Jackson, Michael (1987). "Power and Limitations of Formal Methods for Software Fabrication". In: *Journal of Information Technology* 2.2, pp. 72–76. DOI: 10.1177/026839628700200204.

Jeffery, Ross et al. (2015). "An empirical research agenda for understanding formal methods productivity". In: *Information and Software Technology* 60, pp. 102–112. DOI: 10.1016/j.infsof.2014.11.005.

Kaner, Cem and David Pels (1998). *Bad Software*. Wiley. ISBN: 978-0471318262.

— (2018). *Bad Software: Website*. URL: http://badsoftware.com.

Kitchenham, B., S. Linkman, and D. Law (1997). "DESMET: a methodology for evaluating software engineering methods and tools". In: *Computing & Control Engineering Journal* 8.3, pp. 120–126. DOI: 10.1049/cce:19970304.

Kitchenham, Barbara A. and Shari L. Pfleeger (2008). "Guide to Advanced Empirical Software Engineering". In: Springer. Chap. Personal Opinion Surveys, pp. 63–92.

Klein, Gerwin et al. (2018). "Formally verified software in the real world". In: *Communications of the ACM* 61.10, pp. 68–77. DOI: 10.1145/3230627.

Knight, John C. et al. (1997). "Why Are Formal Methods Not Used More Widely?" In: *Fourth NASA Formal Methods Workshop*, pp. 1–12.

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

Lai, Richard and Wilfred Leung (1995). "Industrial and Academic Protocol Testing: The Gap and the Means of Convergence". In: *Computer Networks and ISDN Systems* 27.4, pp. 537–547. DOI: 10.1016/0169-7552(93)E0110-Z.

Leiner, D. J. (2014). *SoSci Survey*. Tech. rep. URL: https://www.soscisurvey.de.

Liebel, Grischa et al. (2016). "Model-based engineering in the embedded systems domain: an industrial survey on the state-of-practice". In: *Software & Systems Modeling* 17.1, pp. 91–113. DOI: 10.1007/s10270-016-0523-3.

Mathieson, Kieran (1991). "Predicting User Intentions: Comparing the Technology Acceptance Model with the Theory of Planned Behavior". In: *Information Systems Research* 2.3, pp. 173–191. DOI: 10.1287/isre.2.3.173.

Miller, Steven P., Michael W. Whalen, and Darren D. Cofer (2010). "Software model checking takes off". In: *Communications of the ACM* 53.2, pp. 58–64. DOI: 10.1145/1646353.1646372.

Miyoshi, T. and M. Azuma (1993). "An empirical study of evaluating software development environment quality". In: *IEEE Transactions on Software Engineering* 19.5, pp. 425–435. DOI: 10.1109/32.232010.

Mohagheghi, Parastoo et al. (2012). "An empirical study of the state of the practice and acceptance of model-driven engineering in four industrial cases". In: *Empirical Software Engineering* 18.1, pp. 89–116. DOI: 10.1007/s10664-012-9196-x.

Murphy, G. C., R. J. Walker, and E. L. A. Banlassad (1999). "Evaluating emerging software development technologies: lessons learned from assessing aspect-oriented programming". In: *IEEE Transactions on Software Engineering* 25.4, pp. 438–455. DOI: 10.1109/32.799936.

Neuendorf, Kimberly A. (2016). *The Content Analysis Guidebook*. 2nd. Sage. ISBN: 9781412979474.

Neumann, Peter G. (2018). "Risks to the Public". In: *ACM SIGSOFT Software Engineering Notes* 43.2, pp. 8–11. DOI: 10.1145/3203094.3203102.

O'Hearn, Peter W. (2018). "Continuous Reasoning". In: *Proceedings of the 33rd Annual ACM/IEEE Symposium on Logic in Computer Science - LICS'18*. ACM Press. DOI: 10.1145/3209108.3209109.

Oliveira, Jose N. (2004). "A Survey of Formal Methods Courses in European Higher Education". In: *Teaching Formal Methods*. Springer Berlin Heidelberg, pp. 235–248. DOI: 10.1007/978-3-540-30472-2_16.

Oliveira, José Nuno et al. (2018). *Formal Methods Body of Knowledge (FMBoK)*. URL: http://formalmethods.wikia.com/wiki/FMBoK.

Parnas, David Lorge (2010). "Really Rethinking 'Formal Methods'". In: *IEEE Computer* 43.1, pp. 28–34. DOI: 10.1109/mc.2010.22.

Petersen, Kai et al. (2008). "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*. URL: http://ewic.bcs.org/content/ConWebDoc/19543.

Pfleeger, S. L. and L. Hatton (1997). "Investigating the influence of formal methods". In: *Computer* 30.2, pp. 33–43. DOI: 10.1109/2.566148.

Poston, R. M. and M. P. Sexton (1992). "Evaluating and selecting testing tools". In: *IEEE Software* 9.3, pp. 33–42. DOI: 10.1109/52.136165.

Riemenschneider, C. K., B. C. Hardgrave, and F. D. Davis (2002). "Explaining software developer acceptance of methodologies: a comparison of five theoretical models". In: *IEEE Transactions on Software Engineering* 28.12, pp. 1135–1145. DOI: 10.1109/tse.2002.1158287.

Rushby, John (1994). "Critical system properties: Survey and taxonomy". In: *Reliability Engineering & System Safety* 43.2, pp. 189–219. DOI: 10.1016/0951-8320(94)90065-5.

SEI (2010). *CMMI for Development*. Tech. rep. CMU/SEI-2010-TR-033. CMU.

Shull, Forrest, Janice Singer, and Dag I. K. Sjøberg, eds. (2008). *Guide to Advanced Empirical Software Engineering*. London: Springer.
Snook, C and R Harrison (2001). "Practitioners' views on the use of formal methods: an industrial survey by structured interview". In: *Information and Software Technology* 43.4, pp. 275 –283. ISSN: 0950-5849. DOI: 10.1016/S0950-5849(00)00166-X.
Sobel, A.E.K. and M.R. Clarkson (2002). "Formal methods application: an empirical tale of software development". In: *IEEE Transactions on Software Engineering* 28.3, pp. 308–320. DOI: 10.1109/32.991322.
The R Project (2018). *R*. The R Project. URL: https://www.r-project.org.
Wing, J. M. (1990). "A specifier's introduction to formal methods". In: *Computer* 23.9, pp. 8–22. DOI: 10.1109/2.58215.
Wohlin, Claes et al. (2012). *Experimentation in Software Engineering*. Springer. ISBN: 9783642290435.
Woodcock, Jim et al. (2009). "Formal Methods: Practice and Experience". In: *ACM Comput. Surv.* 41.4, 19:1–19:36. ISSN: 0360-0300. DOI: 10.1145/1592434.1592436.