Empirical Software Engineering
https://doi.org/10.1007/s10664-020-09836-5

Formal methods in dependable systems engineering: a survey of professionals from Europe and North America

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

Published online: 09 September 2020

Abstract

Context Formal methods (FMs) 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. Our findings are strongly coherent with earlier observations by Austin and Graeme (1993).

Keywords Formal methods · Empirical research · On-line survey · Usage · Usefulness · Practical challenges · Research transfer · Software engineering education & training

Acronyms

CMMI	Capability Maturity Model Integration
DI	respondents with decreased usage intent
EOU	ease of use

Communicated by: Hélène Waeselynck

✉ Mario Gleirscher
Mario.Gleirscher@york.ac.uk

Diego Marmsoler
Diego.Marmsoler@tum.de

[1] Department of Computer Science, University of York, York, UK

[2] Institut für Informatik, Technical University of Munich, Garching, Germany

FM	formal method
GQM	goal-question-metric
HQ	head quarter
ICT	information and communication technology
II	respondents with increased usage intent
IS	information system
LE	less experienced respondents
M	respondents with some motivations to use FMs
MbE	model-based engineering
ME	more experienced respondents
NP	non-practitioners
P	practitioners
PEOU	perceived ease of use
PU	perceived usefulness
RQ	research question
SE	software engineering
SMT	satisfiability modulo theory
TAM	technology acceptance model
TLD	top-level domain
NM	respondents without any motivations to use FMs
UFM	Use of FMs in mission critical SE
U	usefulness

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.[1] Cost savings from reducing such impact have been *the* motivation of (FMs) 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,[2] which 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, (i) 55% *agree*s, i.e., seem to attribute importance to this role of FMs, (ii) 14% *disagree*s, i.e., oppose that view, (iii) 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 1970s and 1980s, several SE and FM researchers had started to examine this usefulness and to identify error possibilities despite the rigour in FMs (Gerhart and Yelowitz 1976), with the aim of responding to critical observations of practitioners (Jackson 1987).

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

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

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 overkill in some cases but are highly recommended 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. 148–149, 165–166) 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 represent firm requirements concisely and support rigorous inspections[3] and testing. He observes that *changing requirements* has 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. 108–109). Bad communication between theorists and practitioners sustains the issue that FMs are taught but rarely applied (Glass (2002) and Holloway and Butler (1996), pp. 68–70). 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 Tom (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 hidden inconsistencies that can be detected when properly formalised? Are such errors compelling arguments for the wider 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) and Sobel and Clarkson (2002)), often primarily conducted in research labs (e.g. Galloway et al. (1998) and Chudnov et al. (2018) 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 *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

[3]For example, walking through development artefacts in a structured and moderated discussion group and with bug pattern checklists (Fagan 1976).

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

By *formal methods*, we refer to *explicit* mathematical models and *sound* logical reasoning about *critical properties* (Rushby 1994)—such as reliability, safety, security, more generally, dependability and performance—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 in the development and analysis of critical systems and to substantially integrating FMs with the used programming methodologies (e.g. structured development, model-based engineering (MbE), assertion-based programming, test-driven development), notations (e.g. UML, SysML), 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 as a good basis for quantitative follow-up studies, we follow the user-focused TAM. Maturity models according to the Capability Maturity Model Integration (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." According to their approach, 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 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 35 studies on FM research evaluation and transfer. For each study, we estimate[4] 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 Graeme (1993) sought 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 interviews with representatives of five companies to discover the main issues involved in FM use, in particular, issues of understandability and the difficulty of creating and utilising 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. 102–103) assess effects on and shortcomings of the adoption of 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

[4]This estimate is based on opinions and attitudes expressed by the original authors and, where unavailable, on our own interpretation when reading the studies.

Table 1 Overview of related work on FM use and adoption, grouped by primary focus and motivation

Study	A	Motivation	Support	E	C	R
Surveys						
Austin and Graeme (1993)	=	LoEv	Interviews		•	•
Snook and Harrison (2001)	=	LoEv	Interviews	•		
Oliveira (2004)	=	Edu./Train.	Course websites	•		
Woodcock et al. (2009)[a]	=	LoEv	Interviews		•	
Davis et al. (2013)	+	TechTx	Interviews		•	•
Liebel et al. (2016)	+	LoEv	Online questionnaire	•		
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				•
Gleirscher et al. (2019)	+	TechTx	SWOT analysis	•	•	
Expert Opinions and Experience Reports						
Jackson (1987)	=	TechTx			•	
Bjorner (1987)	=	TechTx			•	
Barroca and McDermid (1992)	=	SotA	Multiple cases		•	
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						
Gerhart and Yelowitz (1976)	=	LoEv	Multiple cases, O/E	•	•	•
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 et al. (1993) and Craigen (1995); (A)ttitude, (E)valuation/analysis, (C)hallenges, (R)ecommendations, +/=/– ... positive/neutral/negative, LoEv ... lack of empirical evidence, Hyp. Testing ... hypotheses testing, Edu./Train. ... education/training, O/E ... opinion/experience report, SotA ... state of the art, SWOT ... strengths, weaknesses, opportunities, and threats, TechTx ... technology transfer

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 and the state of practice, 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. In Section 6.3, we provide a further discussion of how our findings relate to the findings of these studies, particularly to the works of Austin and Graeme (1993) and Woodcock et al. (2009).

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 (RQs)*:

RQ1 In which typical domains, for which purposes, in which roles, and to what extent have *FMs been used*?

RQ2 Which *relationships* can we observe between *past experience in using FMs* and the *intent to use FMs*?

RQ3 How difficult do study participants perceive widely known FM *challenges* to be?

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

Table 2 lists the (C)oncepts that constitute the construct *Use of FMs in mission-critical SE* (UFM), the corresponding *scales*, the points of measurement, and references to (Q)uestions from the questionnaire.

Measuring Past and Intended Use For RQ1 (UFM), we examine potential application *domains* for FMs (C1), *roles* when using FMs (C2), *motivations* and *purposes* of using FMs (C6, C4), and the extent of UFM at the general (C5) and specific (C3) 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 (C2), FM class (C3), and purpose (C4). We measure UFM_i by relative frequency (Table 4) with respect to a participants' current situation, FM class, and purpose of use. Using a *relative* instead of an *absolute* frequency scale slightly reduces the burden on respondents to make detailed and, hence, uncertain predictions of UFM_i.

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

Table 2 Concepts and scales for the construct "Use of FMs in mission-critical SE" (UFM)

Concept	Id.	Description [Scale]	Point [Question]
Measured twice...			
Domain	**C1**	Application domains of FMs [MC among domains]	Past [Q1], Intent [Q8]
Role	**C2**	Role in using FMs [MC among roles]	Past [Q4], Intent [Q9]
Use	**C3**	Use of FMs [experience level/relative frequency per FM class]	Past [Q5/Q6], Intent [10/11]
Purpose	**C4**	Purpose of using FMs [absolute/relative frequency per purpose]	Past [Q7], Intent [Q12]
Measured once ...			
Experience	**C5**	Level of FM experience [duration ranges in years]	Single [Q2]
Motivation	**C6**	Motivation to use FMs [degree per motivational factor]	Single [Q3]
Obstacles	**C7**	Difficulty of obstacles to using FMs [degree per challenge]	Single [Q13]

MC... multiple-choice

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 C4 (i.e., subjective assessment of effects of FMs by asking SE practitioners and researchers).

We assume UFM is, nowadays, to a large extent implying 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 U. 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, C2, C3, C4) 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 potential obstacles to the 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 (1) an educational background in engineering and the sciences related to critical computer-based or software-intensive systems, preferably having gained their first degree, *or* (2) a practical engineering background in a reasonably critical system or product domain involving software practice. We use *(study or survey) participant* and *respondent* as synonyms. We talk of *FM users* to refer to the part of the population that has already used FMs in one or another way. See Appendix A.1 and Table 8 for a more fine-grained analysis of the population.

4.4 Survey Instrument: On-line Questionnaire

Table 3 summarises the questionnaire we use to measure UFM (Table 2). The scales used for encoding the answers are described in Table 4.

Although we do not collect personal data, respondents could 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. However, we thought it to be unnecessary in our case to instrument the questionnaire or our tooling to allow us to determine the time spent for submitting complete data points.

Face and Content Validity We derived answer options from the literature, our own experience with FMs, SE research training, discussions with other SE researchers and colleagues from industry, pilot responses, and coding of open answers. Particularly, the classification of FM methods (C3; Q5, Q6, Q10, Q11) and the list of obstacles or challenges (C7; Q13) were derived from our own training, literature knowledge prior to this study, and experience as well as from occasional personal discussions with SE experts from academia and industry. 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.

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	2
Q2	How many years of *FM experience* (including the study of FMs, C5) have you gained?	Duration range in years	5.2	3
Q3	Which have been your *motivations* (C6) to use FMs?	Degree per motivational factor	5.2	4
Q4	In which roles (C2) have you used FMs?	MC among roles	5.3	5
Q5	Describe your *level of experience* (C3) for ⟨*class of formal description techniques*⟩.	Level of experience per class	5.3	6
Q6	Describe your *level of experience* (C3) for ⟨*class of formal reasoning techniques*⟩.	Level of experience per class	5.3	7
Q7	I have mainly *used FMs for* (C4) ...	Absolute frequency per purpose	5.3	8
Q8	In which *domains* (C1) in industry or academia do you intend to use FMs?	MC among domains	5.4	9
Q9	In which *roles* (C2) would (or do) you intend to use FMs?	MC among roles	5.4	10
Q10	I (would) *intend to use* (C3) ⟨*class of formal description techniques*⟩⟨*this*⟩ often.	Relative frequency per class	5.4	11
Q11	I (would) *intend to use* (C3) ⟨*class of formal reasoning techniques*⟩⟨*this*⟩ often.	Relative frequency per class	5.4	12
Q12	I (would) intend to *use FMs for* (C4) ⟨*purpose*⟩.	Relative frequency per purpose	5.4	13
Q13	For any use of FMs in my future activities, I consider ⟨*obstacle*⟩ (C7) as ⟨*that*⟩ difficult.	Degree of difficulty per obstacle	5.5	16

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

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 response. We indicate *dnk*-answers whenever we *exclude* 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 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.

We examine C5, C1, C2, and C3 from Table 2 to check how well our *sample covers the given concept categories*. The better the coverage of these categories the wider is the range of analyses possible from our data set. Less covered categories might indicate inappropriate concepts as well as the case that our sample just does not touch this fraction of the target population. Under the assumption that the sample is drawn from the target population in a uniformly random fashion, we would be able to draw conclusions about the constitution of the target population. However, as noted, this assumption was in our case not controllable.

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 (C2), FM class (C3), and purpose (C4) in Section 5.4. Then, in Section 5.5, 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, $\leq$ 3 years), (2) practitioners (P, practised at least once) from non-practitioners (NP, not used or only in course or lab), (3) motivated (M, moderately or strongly motivated by at least one specified factor) from unmotivated respondents (U, no motivating factor specified), (4) respondents' past and future views, and (5) respondents with increased usage intent (II) from ones with decreased usage intent (DI).

We apply association analysis between these categorical and ordinal variables, using *pairs of matrices* (e.g. Fig. 17). We answer RQ4 by arguing from results for RQ1, 2, and 3.

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 (Neuendorf 2016) to confirm that the understanding of the challenge categories is consistent among the authors of the present study. 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 (Appendix A.11) 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 "Stack Exchange Stats" community[5]). Content analysis and coding takes place in a spreadsheet application. A draft of Appendix A has been archived in Gleirscher and Marmsoler (2018).

5 Execution, Results, and Analysis

In this section, we summarise the responses to the questions in Table 3 and answer the RQs 1, 2, and 3 as explained in Section 4.1. To answer RQ1, we describe the sample in Section 5.2 and discuss some facets of FM use in Section 5.3. For RQ2, we summarise data about past use and usage intent in Section 5.4. For RQ3, we analyse further data in Section 5.5.

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

Table 5 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

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 1 summarises the distribution of responses. The channels in Table 5 particularly cover the European and North American areas.

5.2 Description of the Sample (Answering RQ 1)

A size estimation of the channels in Table 5 yields around 65K *channel memberships* (for some channels we make a best guess but, e.g. for LinkedIn the counts are given). Assuming participants are, on average, member of at least three of the channels, we could have *reached* up to 20K *real persons*. Given a recent estimate of worldwide 23 million SE practitioners (Evans Data 2018) and assuming that at least 1% are mission-critical SE practitioners, our *population* might comprise at least 230K persons, possibly around 38K in the US and 61K in Europe.[6] We received $N = 216$ responses resulting in an estimated *response rate* between 1 and 2% and a *population coverage* of at most 0.1% globally and 0.2% in the US and in Europe. 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.

In the following, we summarise the responses to the questions about the application domain (Q1), the level of experience (Q2), and the motivations (Q3) of a FM user.

Guide to the Figures For Likert-type ordered *scales*, we use centred diverging stacked bar charts (see, e.g. Fig. 4) as recommended by Robbins and Heiberger (2011). The *horizontal bars* in each line show the answer fractions according to the legend at the bottom and are annotated with the percentages of the left-most, middle, and right-most answer options. These bars are aligned by the midpoint of the middle group (for 3- and 5-level scales) or by the boundary between the two central groups (for 4-level scales). *Bar labels* often abbreviate the corresponding answer options in the questionnaire. The questionnaire copy in Appendix A.11 contains short definitions, explanations, and examples to clarify the answer

[6] An estimation in Gleirscher et al. (2019) suggests that about 5% of the overall ICT/IS developer population a re embedded systems practitioners in critical and non-critical domains. Moreover, Evans Data (2018) and Wikipedia contributors (2018) describe data from 2016 and 2017, suggesting that 3.87 million (19%) SE practitioners live in the US and about 13.3 million (39%) in Europe, the Middle East, and Africa. According to an analysis of data from Stack Overflow by ATOMICO (2019), there is a "software engineering talent pool" of about 6.1 million in Europe.

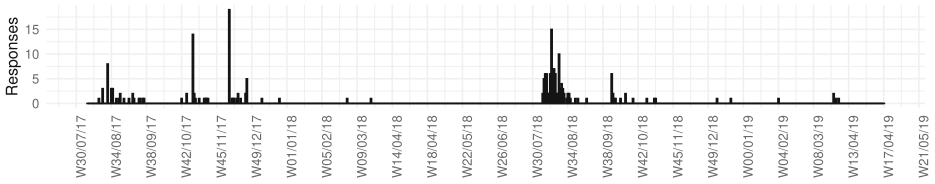


Fig. 1 Distribution of responses over time

options. For sake of brevity, we do not repeat this information here. "M" denotes the median, "CI" the 95% confidence interval for the median calculated according to Campbell and Gardner (1988), "X" the number of excluded data points per answer option, and "NA" the number of invalid data points.

Q1: Application Domain For each domain, Fig. 2 shows the number of participants having experience in that domain.[7] 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. Medical healthcare is an example where participants could have checked more than one answer category because medical devices would belong to "device industry" and emergency management IT would belong to "critical infrastructures". See Appendix A.11 for more information about the answer categories.

Q2: FM Experience Figure 3 depicts participants' years of experience in using FMs, showing that the sample covers all experience levels. However, the fraction of respondents with no experience (i.e., category "0") is comparatively low. 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). A further analysis of the study participants' experience profile is available from Table 8 in Appendix A.1 on page 36.

Q3: Motivation Figure 4 suggests that *regulatory authorities* play a subordinate role in triggering 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. The 88 open responses for this question could either be subsumed in at least one of the given categories (65 in "Own (private) interest", 11 in other categories) or be declared as a comment (3) or not a further motivation (9). Hence, coding did not require an additional answer category to Q3.

5.3 Facets of Formal Methods Use (Answering RQ 1)

In the following, we summarise the responses to the questions about the role of a user (Q4), use in specification (Q5), use in analysis (Q6), and the underlying purpose (Q7) of such use.

Q4: Role Figure 5 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)).

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

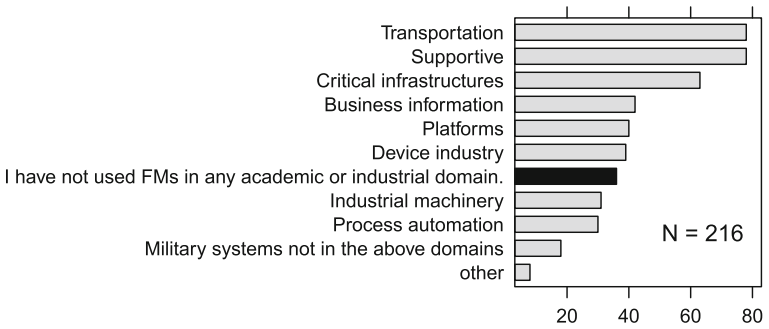


Fig. 2 (Q1) In which application domains in industry or academia have you mainly used FMs? (MC)

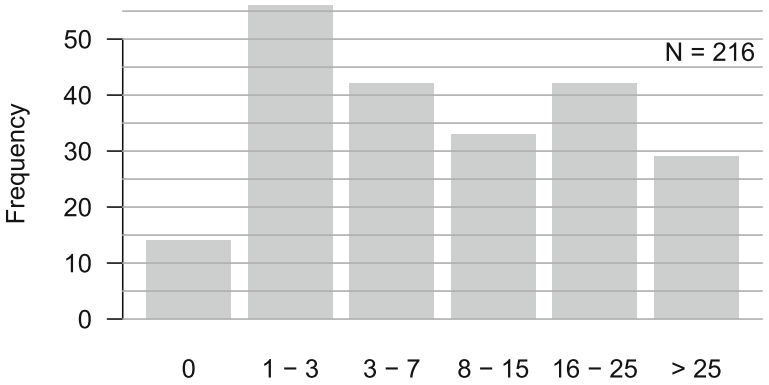


Fig. 3 (Q3) How many years of FM experience (including the study of FMs) have you gained?

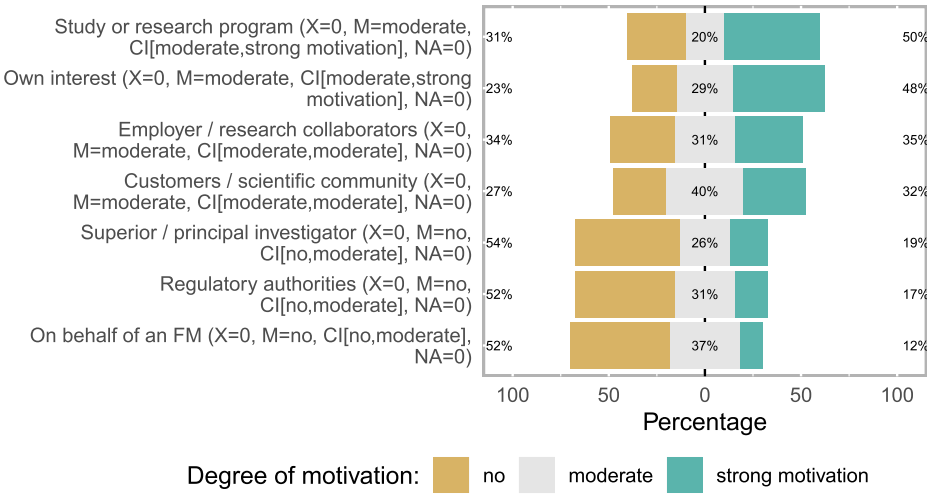


Fig. 4 (Q3) Which have been your motivations to use FMs?

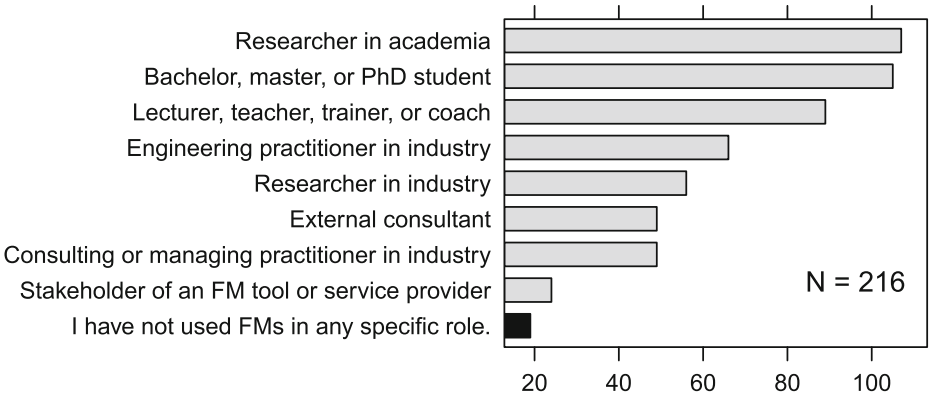


Fig. 5 (Q4) In which roles have you used FMs? (MC)

Q5: Use in Specification The degree of usage of FMs for specification is depicted in Fig. 6. 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 Fig. 7. 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 for Q5, the use of FMs in computational engineering, such as algebraic reasoning about differential equations, is again exceptionally low.

Q7: Purpose Figure 8 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.

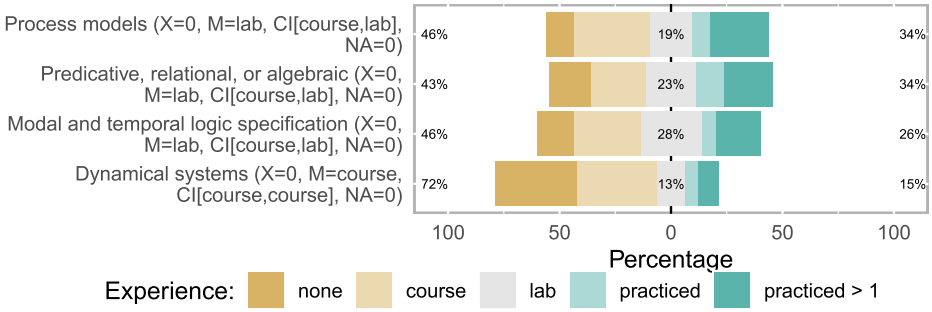


Fig. 6 (Q5) Describe your level of experience with each of the following classes of formal description techniques

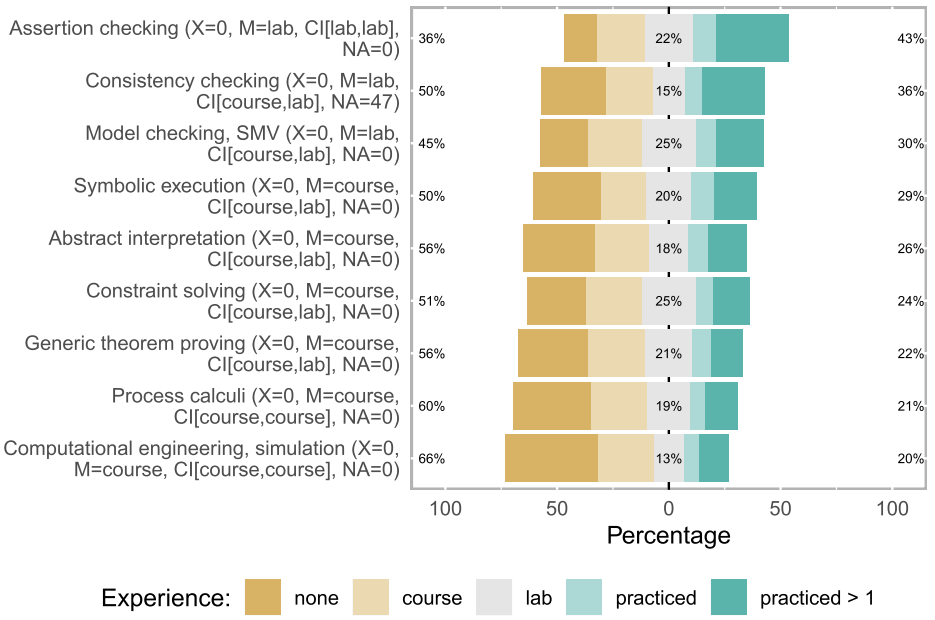


Fig. 7 (Q6) Describe your level of experience with each of the following classes of formal reasoning techniques

5.4 Past Use Versus Usage Intent (Answering RQ 2)

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 9 compares the respondents' past domains of FM application with their intended domains (see Q8). This figure reveals two insights into the participants' intentions to use FMs: (i) Fewer participants do not want to apply FMs in the future (19) than participants that have not used FMs (36, see yellow bars). Ten participants fall into both categories, they have not used FMs and do not intend to use FMs. (ii) The intended

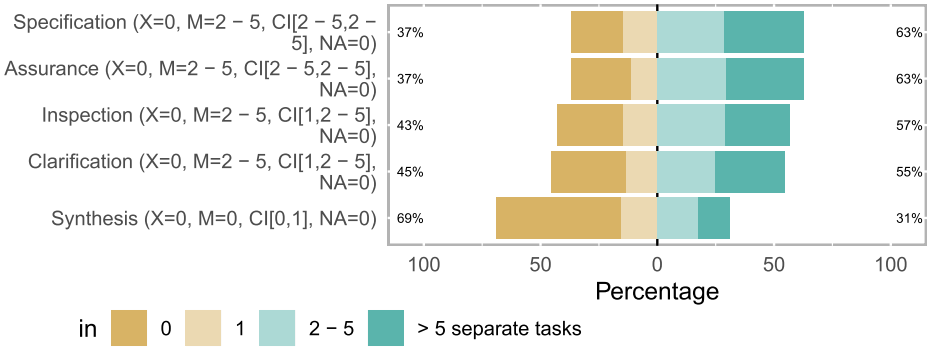


Fig. 8 (Q7) I have mainly used FMs for ...

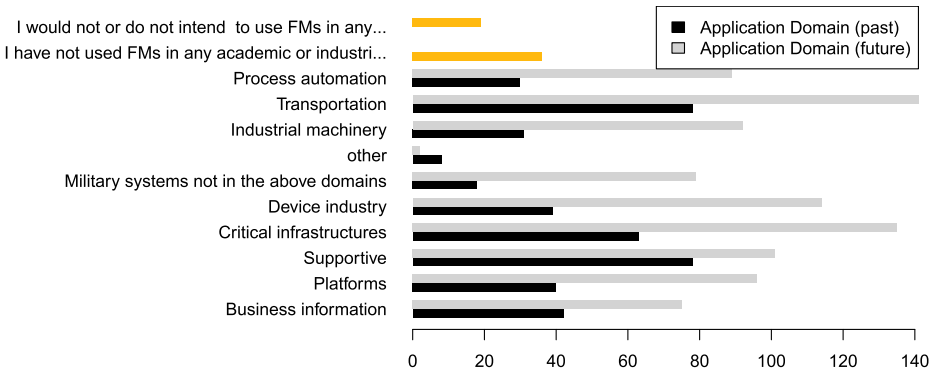


Fig. 9 Number of respondents using FMs by domain (past vs. intent)

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 10 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.

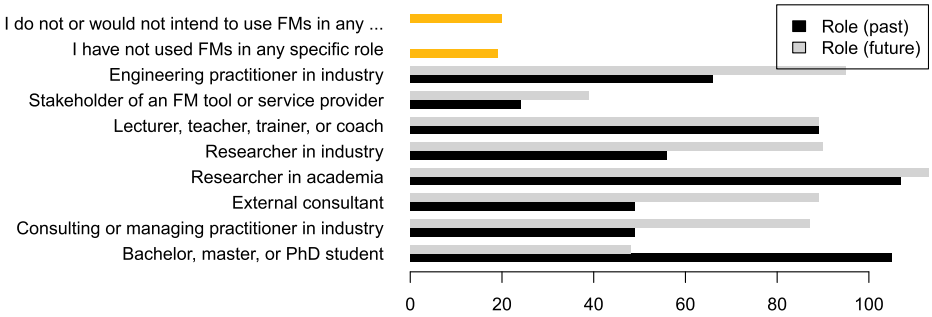


Fig. 10 Number of respondents applying FMs by role (past vs. intent)

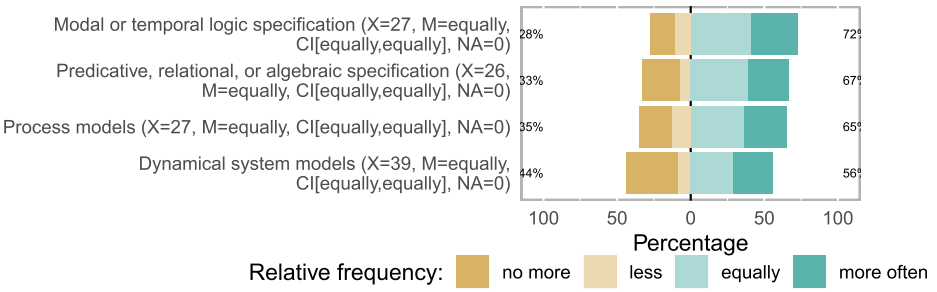


Fig. 11 (Q10) I (would) intend to use ...

Q10: Intended Use for Specification Figure 11 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.

Q11: Intended Use for Analysis The respondents' intended use of FMs for the analysis of specifications (i.e., formal reasoning techniques) is depicted in Fig. 12. Except for process calculi, we observe a general tendency of the participants to *increase* their future FM use.

Q12: Intended Purpose Figure 13 indicates why respondents intend to apply FMs. Again, there is a tendency of the participants to *increase* FM use across all listed purposes.

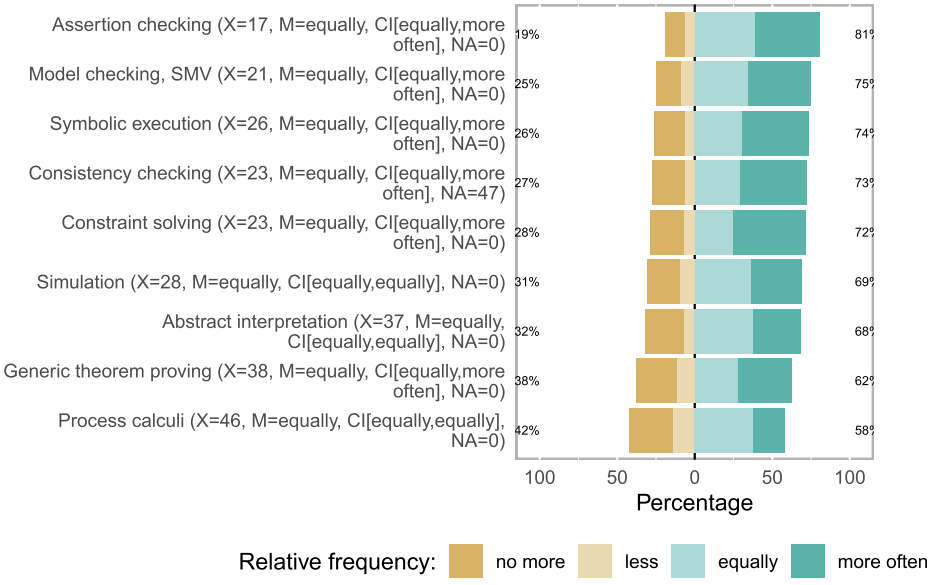


Fig. 12 (Q11) I (would) intend to use ...

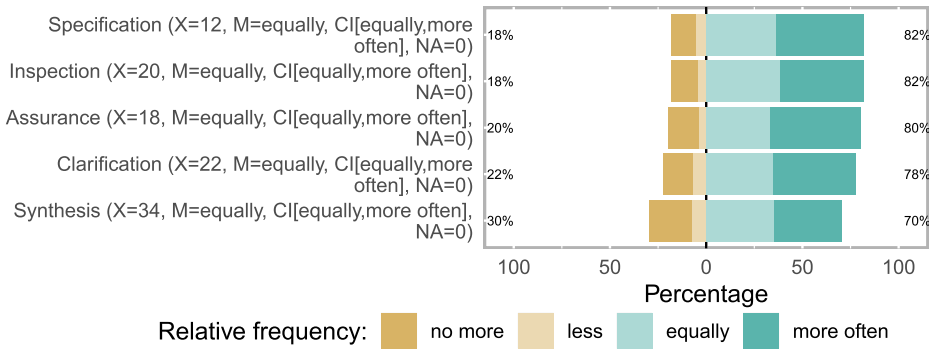


Fig. 13 (Q12) I (would) intend to use FMs for ...

Q7 and Q12: Comparison of Code- and Model-based FMs In the following, we regard *practitioners* with experience level "applied several times in engineering practice" or "applied once in engineering practice" and frequency "applied in 2 to 5 separate tasks" or "applied in more than 5 separate tasks" (see Table 4). We compare *users of code-based FMs* (CBs; including "abstract interpretation", "assertion checking", "symbolic execution", "consistency checking"; with N=128) with *users of model-based FMs* (MBs; including "process calculi", "model checking", "theorem proving", and "simulation"; with N=114). While some of the FM classes can be seen as both, code- and model-based, we made a choice based on our experience but left out "constraint solving" because it is a fundamental technique intensively applied in both.

The comparison of past and future use for code-based (top half of Fig. 21 in Appendix A.4) and model-based FMs (bottom half of Fig. 21), for example, in *inspection* (e.g. error detection, bug finding) shows the following:

- CBs show slightly more frequently an increased intent (the "more often" group) than MBs; for both sub-groups, respondents with 2 to 5 and with more than 5 past uses.
- MBs show slightly more frequently a decreased intent (the "no more" group) than CBs.

Looking at *assurance* (e.g. proof, error removal) shows the following:

- MBs show slightly more frequently an increased intent than CBs when looking at respondents who have used FMs more than 5 times. However, MBs indicate slightly less frequently an increased intent than CBs when looking at respondents with 2 to 5 uses.
- CBs indicate more *dnk*s after 2 to 5 uses and slightly more frequently a decreased intent after 5 uses in comparison with MBs.

Q1, Q5, and Q6: Practised FM Classes by Application Domain We asked respondents about their use of each FM class *independent* of the application domain and about their general use of FMs in each such domain. Hence, we can only approximate past usage per FM class and application domain assuming that the overall usage per respondent is uniformly distributed among the specified FM classes and domains. For that, we interpret (and count) each respondent who specifies a domain in combination with "applied once in engineering practice" or "applied several times in engineering practice" for an FM class as a practitioner who *has used* (UFM_p) or, respectively, *wants to use* (UFM_i) FMs of that class in that domain. More generally, we count a respondent who specifies n domains, say d_1 to d_n, in

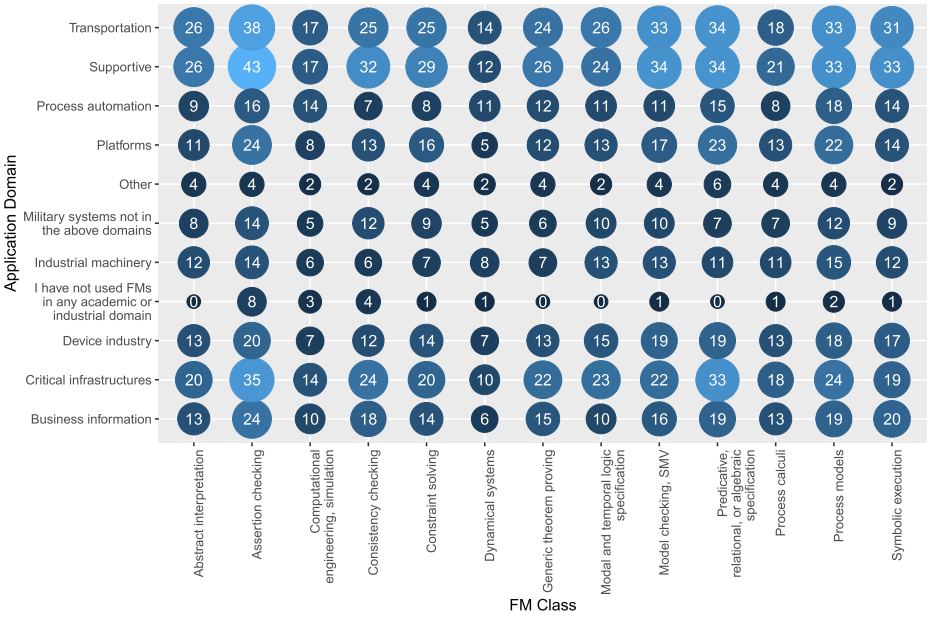


Fig. 14 Approximation (likelihood) of practised use (UFM_p) by FM class and application domain

combination with "applied once in engineering practice" or "applied several times in engineering practice" for m FM classes, say c_1 to c_m, as a practitioner who *has used* (UFM_p) or, respectively, *wants to use* (UFM_i) FMs of the classes c_1 to c_m in the domains d_1 to d_n. Figs. 14 and 15 show these approximations for UFM_p and UFM_i.

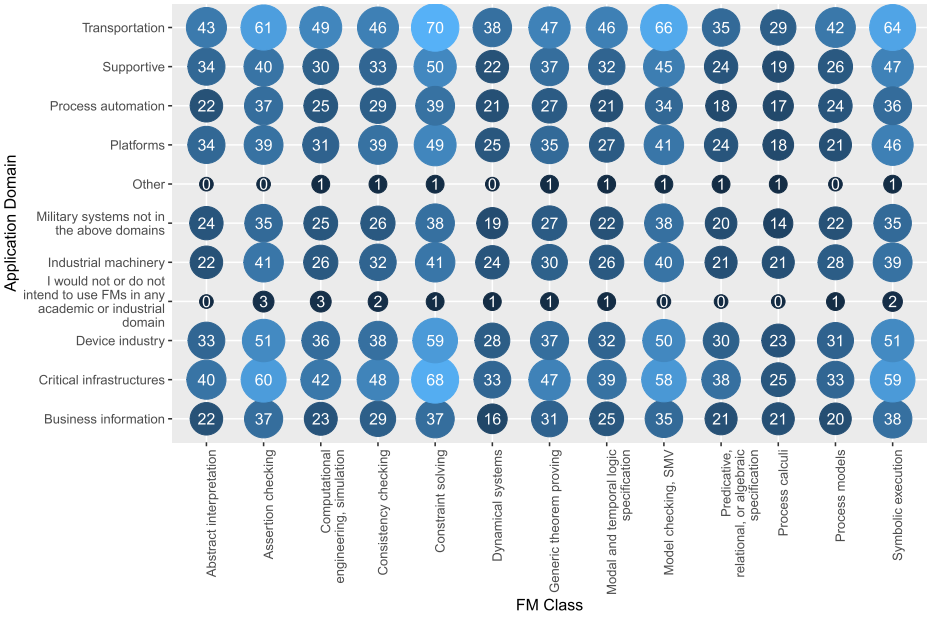


Fig. 15 Approximation (likelihood) of increased usage intent (UFM_i) by FM class and application domain

5.5 Perception of Challenges (Answering RQ 3)

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 16 shows the respondents' ratings of all challenges. Most of them 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*, the 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. Appendix A.6 provides some background material for the following association analyses.

Less Experienced (LE) Versus more Experienced (ME) Respondents (Q2) The comparison of the difficulty ratings of LEs with the ratings of MEs shows that (i) LEs less often perceive the given challenges as tough, t (ii) MEs significantly more often rate *scalability* as tough, (iii) 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*. Scalability, the forerunner in Fig. 16, exhibits the most tough-ratings from NPs in *synthesis* and from Ps in *assurance* and *clarification* (see the top half of Fig. 22 in Appendix A.6).

Decreased Intent (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: (i) *Scalability* and *skills and education*, both forerunners in Fig. 16, show the most tough-ratings from IIs for *assurance* (67%) and *inspection* (66%) and from DIs for *synthesis* (53%). (ii) The trend in Fig. 16 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) The top half of Fig. 17 shows for NPs, the trend in Fig. 16 is **largely independent of the FM class**, except for *consistency checking* and *logic* leading with *tough* proportions of 49%.

The bottom half of Fig. 17 shows for Ps, difficulty ratings across FM classes vary more: The foremost challenges in Fig. 16 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 histograms in the lower right corners in Fig. 17 indicate that (i) NPs' difficulty ratings vary less than Ps' ratings, (ii) NPs' ratings are more independent from the FM classes, and (iii) NPs' difficulty ratings are lower on average than Ps' ratings. Appendix A.6 contains several such association matrices with more detailed data in the matrix cells.

Table 6 Feedback on given and additional challenges (see Appendix A.8 for a full list of references)

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 Fig. 16; 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 NMs; 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, documentation)	Q	16 studies, e.g. Bjorner (1987) O'Hearn (2018)	Top-rated by DIs; but comparatively small tough-proportions from practitioners

Table 6 (continued)

Challenge name & description	Src.	Supported by (oldest, newest)	Findings for RQ3 (Section 5.5)
Maintainability: Stable proofs, easily modifiable specifications, and adaptable verification results	Q	Barroca and McDermid (1992), Knight et al. (1997), and Parnas (2010)	Comparatively small tough-proportions from practitioners
Resources: Sufficient resources, good cost-benefit ratio (despite adoption, training, licenses)	4R	11 studies, Hall (1990) and 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	6R	12 studies, Bjorner (1987) and O'Hearn (2018)	
Practicality & Reputation: Benefit awareness and sufficient empirical evidence for benefits	7R	5 studies, Lai and Leung (1995) and Parnas (2010)	

Src.. . . source, Q . . . in questionnaire, *n*R . . . additionally raised by *n* Respondents

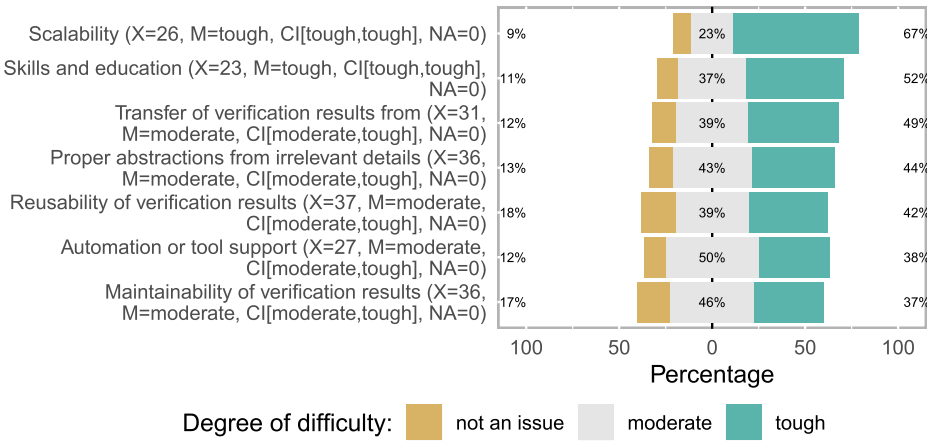


Fig. 16 (Q13) For any use of FMs in my future activities, I consider ⟨*obstacle*⟩ as [not an|a moderate|a tough] issue

Decreased Intent (DI) Versus Increased Intent (II) by FM Class (Q10, Q11) The trend in Fig. 16 is supported by many tough ratings (48%) for *transfer of verification results* from DIs in *consistency checking*. However, DIs in *process calculi* provide 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 Fig. 16, 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 the given challenges as moderate to tough, **regardless of the motivating factor.** The trend in Fig. 16 seems explainable by many tough ratings from respondents motivated by *regulatory authorities* (69%), not motivated by *tool providers* (56%), and not motivated by *superiors/principal investigators* (56%, see Fig. 24 in Appendix A.6). Us' tough-ratings are **notably lower than Ms'** tough-ratings.

Past and Future Views by Role (Q4, Q9) Although participants show role-based discrepancies between their past and intended use of FMs (Fig. 10), the **perception of difficulty** of the rated challenges seems to be **largely similar**, following the trend in Fig. 16. The high ranking of *scalability* (and *reusability of verification results*) is supported by 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 Fig. 16 is underpinned by highest tough-proportions for respondents from the *transportation*, *military systems*, *industrial machinery*, and *supportive* domains.

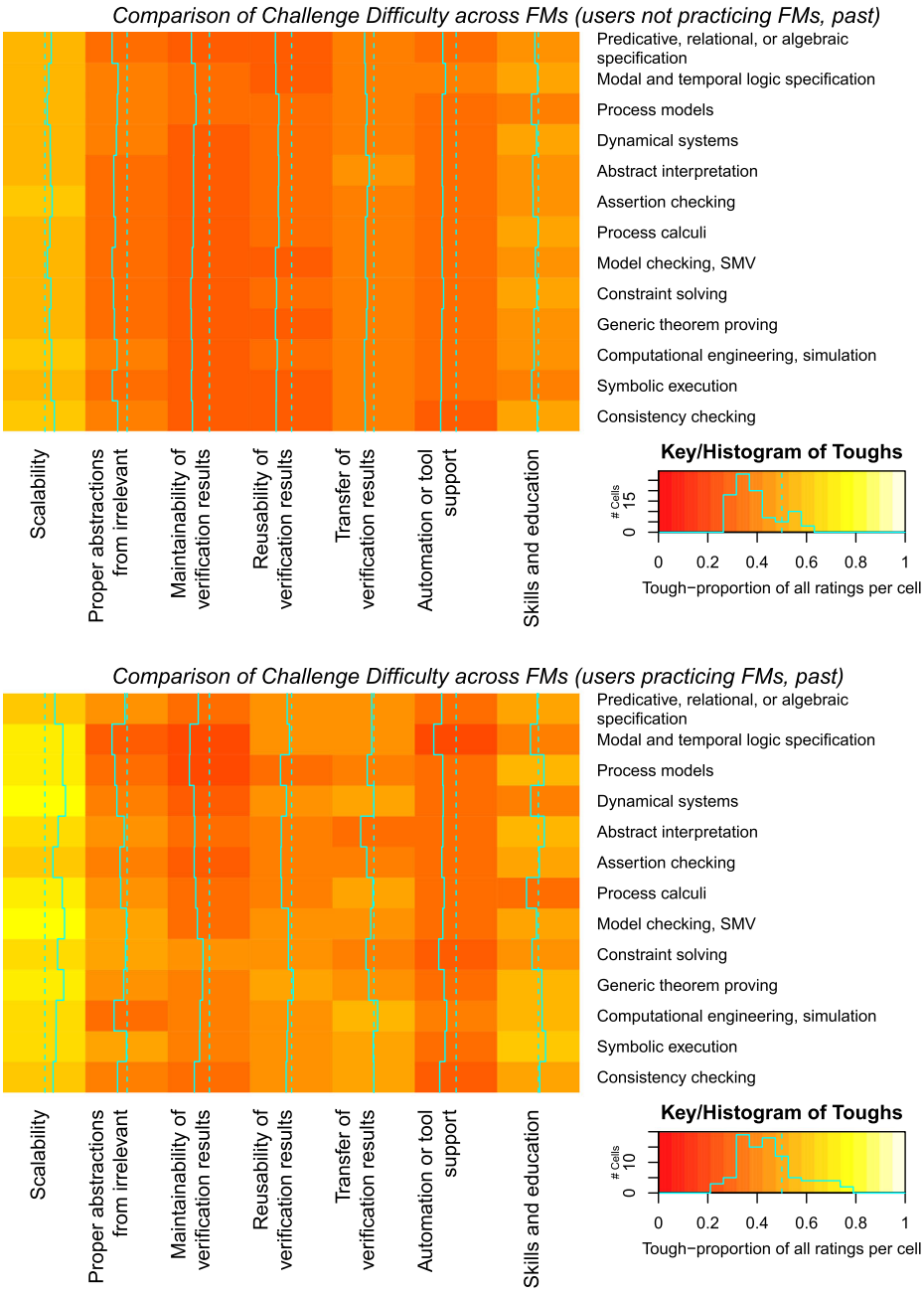


Fig. 17 Difficulty of challenges (cols): NPs (top) compared to Ps (bottom) by class of used FM (rows). *Legend:* In each cell of an association matrix, both the solid vertical line and the colour (gradient from red to white) represent the tough proportions (from 0 to 100%), with the dotted vertical line marking the 50% margin. The histogram (to the lower right corner of each matrix) counts the combinations (cells) in each 5%-band of tough ratings. E.g. ∼70% of "process calculi" users perceive "scalability" as a tough challenge

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. All findings are then collected in Table 7 on page 26.

Findings for RQ 1

F1 *Regulatory authorities* with their norms, codes, or policies represent only a minor motivating factor to use FMs. *Intrinsic motivation* (maybe market-triggered) seems to be stronger. This finding is consistent with what we know from the literature survey in Gleirscher et al. (2019): FMs are not formally required by corresponding standards today, not even for the highest safety integrity levels. If regulatory authorities change their recommendations to requirements, then this might spike as a motivating factor.

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 norms or codes of regulatory authorities.
F2 The fraction of respondents with no experience at all is comparatively low.
F3 Respondents use FMs the least in computational engineering and for dynamical systems.
RQ2: Which *relationships* can we observe between *past experience in using FMs* and *intent to use FMs*?
F4 Increased intent to use FMs observable across all application domains.
F5 Amount of experience is positively associated with the strength of usage intent.
F6 The responses do not show any significant differences between code- and model-based FMs.
F7 Respondents show high likelihoods of an increased intent to use FMs such as "model checking" or "assertion checking" in areas such as "transportation" or "critical infrastructures".
RQ3: How difficult do study participants perceive widely known FM *challenges*?
F8 Scalability and skills & education lead the challenge difficulty ranking.
F9 Maintainability of proof results is found to be the least worrying challenge.
F10 Reusability of proof results is rated as tough by several practitioner groups.
F11 FM users with decreased usage intent rate *tool deficiencies* as their top obstacle.
F12 Respondents identified resources, process compatibility, and reputation as further obstacles.
F13 All considered challenges are generally perceived as moderate or tough.
F14 Among the FM classes, process models are most positively associated with tough scalability.
RQ4: What can we say about the *perceived ease of use* and the *perceived usefulness* of FMs?
F15 Respondents perceive the usefulness of FMs as mainly positive and intend to increase their use.
F16 Respondents perceive the ease of use of FMs as mainly negative.
Relationship to Existing Evidence (from the literature):
F17 Proof maintainability and reusability are least covered by the literature.
F18 We repeat Austin and Graeme (1993), excluding benefit analysis but with a broader sample and more detailed questions.

F2 The low fraction of respondents with no experience in Fig. 3 may have been caused (1) by our choice of expert channels in Table 5 where the likelihood of encountering FM users is probably higher than in more generic SE channels (e.g. Stack Overflow) and (2) by the fact that SE students will usually have an FM course or some lectures about FMs such that they would choose "1–3 years" in Q2 and "studied in course" in Q3.

F3 We observe the least use of FMs in computational engineering and for reasoning about dynamical systems, for example, reasoning about the correctness of algorithms, and their implementation in embedded software, controlling such systems. One explanation for this is that our sample mainly comprises software and systems engineers who will work less intensively with such FMs than, for example, mechanical or control engineers. Another explanation is that such FMs are still less widely known, less well developed, or less well supported by tools than FMs focusing on the reasoning about pure software.

Findings for RQ 2

F4 It seems that in *all given domains* (Fig. 9, 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 by teachers and researchers is saturated. This saturation indicates a stable intent to teach FMs, to perform research in FMs, or to otherwise use FMs in teaching or research. However, there is an increased intent to apply FMs in *industrial contexts* in the future. One explanation could be that engineers have already wanted to use FMs but have not had the opportunity or were not told or permitted to do so. Another explanation for an increased intent of FM non-users could be due to some bias when answering questions about whether someone would do (e.g. try out) something.

F5 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 as described in Section 4.6). We observe that the *more experience* one has with using a specific FM class, the *more likely they will apply it in the future* (see the two charts in the Appendices A.3 and A.5). No experience with a specific FM class correlates with a *low intent* to use that class. Participants not having used FMs and, hence, unfamiliar with them might not have had the need in the first place. Only little experience with a certain FM class *significantly increases the intent to apply it again in the future*. Similar observations can be made for the use of FMs in general for a specific purpose.

F6 The differences in past and intended use between code- and model-based FMs (Section 5.4), for example, when looking at inspection and assurance, are marginal. Moreover, we cannot find a significant difference or a trend between these two categories of FMs when considering different purposes, experience levels, and usage frequencies.

F7 The approximation in the Figs. 14 and 15 allows the, albeit vague, interpretation of the numbers as the likelihood that respondents *have used* (Fig. 14) or *want to use* (Fig. 15) a particular FM in a particular domain. Assuming this model, Fig. 15 indicates the highest likelihoods of an increased UFM_i for methods such as "assertion checking", "constraint solving", "model checking", and "symbolic execution" in domains such as "transportation", "critical infrastructures", and the "device industry".

Findings for RQ 3

F8 *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. Fig. 18 may explain the latter by showing a high fraction of students among the 46 non-practitioners.

F9 *Maintainability of proof results* or other verification artefacts was found to be the least difficult challenge. However, in the lower half of Fig. 17, the challenge column "maintainability" shows relatively low frequencies for "modal and temporal logic" and "model checking" (possibly because of the high level of automation) whereas "theorem proving" (possibly because of a low level of automation) and "constraint solving" (possibly because of being too versatile or generic for the present purpose) show the highest frequencies of tough ratings. See Fig. 26 in the Appendix A.6 for more details.

F10 *Reusability of proof results* was rated as tough by several practitioner groups.

F11 FM users with decreased usage intent rate *tool deficiencies* as their top obstacle to FM adoption.

F12 Furthermore, our respondents raised three additional challenges (i.e., resources, process compatibility, and practicability & reputation) 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.

F13 Challenges are perceived *as moderate or tough*, largely similarly between the pairs of groups we distinguish in Section 4.6.

F14 With 72% of tough ratings for *scalability*, process calculi (e.g. ACP, CCS, CSP) perform in the midfield despite their high reputation as compositional methods. *Scalability* of process models (e.g. Petri nets, Mealy machines, labelled transition systems, Markov models) is also ranked in the middle field of tough challenges. The ranking of these models, however, is unsurprising in the light of the difficult scalability of model checking, a frequently used verification technique for process models and the leader in this ranking (cf. Fig. 17). One explanation for the high number of tough-ratings from NPs in *synthesis* could be that NPs might either not associate FMs with synthesis in general, or because automated synthesis of sophisticated artefacts is known to be an unsolved problem in many cases, independent of the use of FMs.

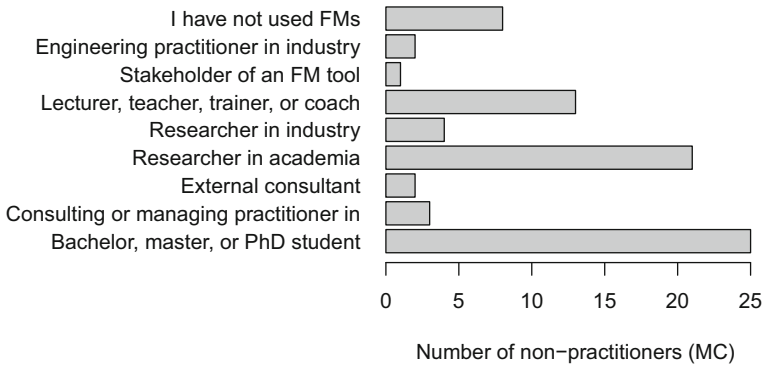


Fig. 18 The past role profile of the 46 non-practitioners (out of 216 respondents) helps to explain finding F8

6.2 Relationship to TAM for Methods (Answering RQ 4)

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, F4 suggests an increased intent to use FMs in the future. Moreover, F4 suggests a positive association of the degree of experience with UFM_i, that is, more experience increases the intent. **F15** 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* to be more positive than negative.

Inspired by Riemenschneider et al. (2002), in the last paragraph of Section 4.2, we justify the use of challenge scales to collect data for PEOU and PU. We justify the validity of the FM-specific challenge scale using the studies in Table 1. The column "supported by" in Table 6 indicates studies discussing the corresponding challenges. From these discussions, we infer that tackling these challenges 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 tools and automation are 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 F13, we observe that respondents rate most challenges as moderate to tough, largely independent of other variables (F8).

F16 Overall, it thus seems that our respondents perceive the *ease of use of FMs* to be more negative than positive. According to Table 6, many of the surveyed studies discuss *skills & education* (12 studies) and *tools & automation* (16) as important challenges. Moreover, Fig. 16 suggests that conceptual difficulties (possibly, from a lack of education and training, from difficulties in FM teaching, from a lack of FM students) seem to be at least as responsible for the negative ease of use as the lower ranked tool deficiencies. Indeed, in a recent discussion of "push-button verifiers", O'Hearn (2018) highlights that both conceptual expertise and tool deficiencies are still significant bottlenecks. However, an investigation of respondents' experiences with FM tools in comparison to their experiences with FM concepts goes beyond the possibilities of the data collected for this study.

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. **F17** However, the fact that maintainability and reusability were least covered by our literature is, on the one hand, in line with F9 but, on the other hand, not with F9 and typical cultures of reuse in practice.

Beyond the general findings about FM benefits in Austin and Graeme (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 both questions on benefits and on suggestions to overcome obstacles. Regarding the latter,

Austin and Graeme (1993) mention the improvement of education and standardisation, the preparation of case studies, and the definition of FM effectiveness metrics.

F18 The report of Austin and Graeme (1993) from the National Physical Laboratory archive was unfortunately no more available to us. We finally managed to get access to a paper copy provided by a friendly colleague. This, however, only happened after conducting this survey. Anyway, we found that our conclusions are nearly identical to Austin and Graeme's. The data from Fig. 16 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 Graeme (1993) is scalability, rated by our respondents to be the toughest obstacle.

F5 is in line with other observations in Woodcock et al. (2009) and Bicarregui 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). Finally, our study generalises the main findings about barriers in Davis et al. (2013) to several geographies and application domains, however, using an on-line questionnaire instead of interviews and not asking for barrier mitigations.

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 partial (○) or full (✓) mitigation.

6.4.1 Construct Validity

Why would the construct (Section 4.2) appropriately represent the phenomenon?

maj ↯: Inappropriate questions and conceptual misalignment / 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 Graeme (1993) finally confirms the appropriateness of our questions. However, we might have needed additional questions to check for conceptual alignment, for example, to more precisely determine whether the respondents' understanding of *FMs* and of the *use or application of FMs* closely matches ours. However, from 18 respondents giving feedback on our questionnaire, only one commented on the definition and one on the classification of FMs. That suggests that many respondents did not have or were not aware of misunderstandings worth mentioning. ○

min ↯: Questionnaire limited for measurement of PEOU (e.g. per FM class) and PU / We avoid deriving conclusions specific to a FM or a corresponding 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 QR13. 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. ✓

6.4.2 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" (Fig. 4) and of Q6 and Q11 with the option "consistency checking" (Fig. 7). 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. ✓

min ↯: Inter- vs. intra-UFM inference / Our study design is not suitable for "inter-UFM predictions", for example, to predict that (dis)satisfied model checking practitioners have an increased (a decreased) intent to use theorem proving. However, the argumentation in the Sections 4.2 and 6.2 aims at "intra-UFM predictions", that is, inferring an increased or decreased intent to use model checking from the quantity and quality of past experience in using model checking. Such predictions may inherit possible limitations of TAM studies. ○

6.4.3 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 (i) improve targeting by repetitively advertising on multiple appropriate channels, (ii) spot unreliable contact information, (iii) provide incentive (study results via email), (iv) keep the questionnaire short and comprehensible, (v) avoid forced answers, and (vi) 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. However, from an estimated population of around 100K (i.e., the rounded sum of 38K and 61K), the minimum sample size for 95% confidence intervals with continuous scale error margins of less than 7% is 196, consistent with the ballpark figure in Gleirscher et al. (2019, p. 117:29). Our sample (N=216) exceeds this number. The 95% confidence intervals for the Likert items show that the margin of error for the median sometimes deviates by one category (e.g. Fig. 4).

In this first study, we aim at understanding common perceptions, such as *"FMs are not practically useful"* or *"FMs are difficult to apply"*. Because these statements address FMs as a whole, we believe such local errors do not affect our general conclusions. However, the response rate (1 to 2%) and population coverage (0.1%, cf. Section 5.1) were too low to avoid such errors and refute specific null-hypotheses, such as *"FM m is effective for role r and purpose p in domain d"* (by the FM community) or *"FM m is difficult to apply for role r and purpose p in domain d"* (by SE practitioners), with satisfactory statistical power. ✓

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. Selection bias, a problem in snowball

sampling (Biernacki and Waldorf 1981), is limited by good visibility and accessibility of the target population in these channels (Section 5.2) as well as little use and control of referral chains among respondents. Our sample includes 50% practitioners according to Section 4.6, ≈ 21% NP (incl. laypersons), and ≈ 31% pure academics. A bias towards FM experts (Fig. 3) 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. critical IT systems in the finance and e-voting sectors. ○

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

min ↯: Lack of FM knowledge / 11 to 18% of our respondents did not know specific challenges (Fig. 16). For RQ1 (Figs. 2 and 16), *dnk*-data points have no influence because the findings of RQ1 directly describe and interpret the status quo of UFM_p. For test purposes, we included *dnk*-data points in the analyses of RQ2 and RQ3 (Figs. 11 and 16), 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 geographies such as China, India, or Brazil. ○

6.4.4 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 c (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. The addition of that item only after the 47th respondent might explain this difference. ○

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 intend 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 that FM*.

Barroca and McDermid (1992) present evidence to show that FMs can be used in industry effectively and more widely. Their observation from 1992 is that FM use had been limited, benefits were clear but limitations were subtle. In response to Barroca and McDermid's finding "FMs are both oversold and under-used", our insights from the analysis of RQ 2 and 3 lead us to conclude that today FMs are probably more underused than oversold. However, our data also suggests that these methods still need substantial 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 the questionnaire difficult for FM beginners.

Implications Towards a Research Agenda In the spirit of Jeffery et al. (2015) and complementing the suggestions from the SWOT analysis in Gleirscher et al. (2019), we want to make another step in setting out an agenda for future FM research.

To address *scalability*, we need more research on how compositional methods (e.g. automated assume-guarantee reasoning, Cofer et al. (2012); automated assertion checking, Leino and Rustan (2017)) can be better leveraged in practical settings. To address *skills and education*, we need an enhanced and up-to-date *FM body of knowledge* (FMBoK; Oliveira et al. (2018)). From his survey of "FMs courses in European higher education", Oliveira (2004) observes that (i) "model-oriented specification", "formalising distribution, concurrency and mobility", and "logical foundations of formal methods" showed to be the topic areas most frequently taught by FM lecturers, and (ii) Z, B, SML, CSP, and Haskell showed to be the most popular formal notations and languages taught in these courses. A comparison of the current state with Oliveira's observations can help to evaluate and revise current FM curricula (e.g. for undergraduate SE as suggested in Davis et al. (2013)) and to derive recommendations for improved FM courses fostering good modelling, composition, and refinement skills in SE practice. 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 *process compatibility*, we need more research in *continuous reasoning* (e.g. O'Hearn (2018) and Chudnov et al. (2018)), a revival of activities, possibly even regulations, in tool integration and model data interchange, and guidance on how to update engineering development processes. To address *reputation*, we need to provide more incentives for practitioners to use 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. Cost-savings analyses of FM applications (e.g. Jeffery et al. (2015)) supported by strong empirical designs (i.e., controlled field experiments) can help to collect the necessary evidence for decision making, successful knowledge transfer, and for implementing this vision.

This survey underpins and enhances the analysis of strengths and weaknesses of FMs in Gleirscher et al. (2019) and can be a guide (1) for consulting and managing practitioners when considering the introduction of FMs into a engineering organisation, (2) for research managers when shaping a grant programme for FM experimentation and transfer, and (3) for associate editors when organising a journal special section on applied FM research.

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 (i) to find out which particular FM (and tool) is used in which domain for which particular purpose and role (e.g. was SMT solving used for model checking in certification or for task scheduling at run-time?), (ii) to measure where particular techniques work well (e.g. which types of formal contracts work well in control software requirements management in a DO-178C context?), (iii) to measure key indicators for successful use of FMs, (iv) to identify management techniques needed to accommodate the changes in working practices, and, finally, (v) to provide guidance to future projects wishing to adopt FMs.

In a next survey, we like to ask about typical FM benefits, about suggestions for barrier mitigation (Davis et al. 2013), pose more specific questions on scalability and useful abstraction, the geographical[8] and educational background, and for conceptual alignment. Further analysis of obstacles, benefits, and usage intent could also benefit from a more fine-grained distinction between FMs directly applied to program code and FMs focusing on more abstract models. We would 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.

The research design, and even our current data set, allows the derivation of the usage intent (UFM_i) for each FM class, application domain, and obstacle. These UFM_i values could be used to analyse whether a particular FM might be (1) underused (i.e., domains with an increased usage intent indicate a potential for more applicability) or (2) oversold (i.e., domains with a lower usage intent and were obstacles are perceived as being particularly tough and, hence, FMs as being less effective).

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. He connected us with John Fitzgerald, who made his paper copy of Austin and Graeme

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

(1993) available to us such that we were able to complete our investigation. We are grateful to John Fitzgerald and also to 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.

Funding Partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the Grant no. 381212925.

Compliance with Ethical Standards

Conflict of interests The authors declare that they have no conflict of interest.

Appendix: A Supplementary Material for "Formal Methods in Dependable Software Engineering: A Survey"

In the following, we provide additional material to the survey, including

1. a more detailed analysis of responses to certain questions (Appendix A.1),
2. further visualizations of the collected data (Appendices. A.2 and A.6),
3. more details on our analysis of related work (Table 9 in Appendix A.7),
4. more details on the mapping from studies to challenges (Appendix A.8),
5. a copy of the advertisement flyer (Appendix A.9),
6. a screenshot of the Twitter poll (Appendix A.10), and
7. a copy of the whole questionnaire (Appendix A.11).

A.1 Data for Analysis of RQ,1 and Estimation of External Validity

Based on the responses for question Q1, the Table 8 provides an overview of categories of respondents referred to in our analysis (particularly, in Section 5.2 and Fig. 10) along with the corresponding counts based on the sample from 31.3.2019 with $N = 220$.

For question Q1, by practitioner, we mean "practitioner in dependable or mission-critical software engineering." To include respondents from all areas of the population or at any study stage, we generalize "practitioner" by the term "user". Below, for the questions Q5 and Q6, we then refer to "formal method user" and "FM-non-user".

7.1 A.2 Geographical Analysis of the Sample

Figure 19 shows geographical aspects of the sample for this study.

A.3 Usage Intent (UFM_i) by Purpose (for Analysis of RQ2)

The comparison in Fig. 20 (and in the figures of the Appendices A.4 and A.5) contains two columns.

Table 8 Overview of categories of respondents for Q1, Q3, Q4, Q5, and Q6

Category of respondents to . . .	Description	Count	Fraction
. . . Question Q4:			
Respondents with academic educational background (AEB)	Researchers in academia; bachelor, master, or PhD students; lecturers, teachers, trainers, coaches	156	72%
Academics with pure transfer experience	AEB cut with researchers in industry, consulting and managing practitioners, and external consultants; **without** engineering practitioners in industry and without tool provider stakeholders	35	16%
Academics with practical experience	AEB cut with engineering practitioners in industry	41	21%
Academics with experience in transfer and practice	AEB cut with researchers in industry, consulting and managing practitioners, and external consultants; **cut with** engineering practitioners in industry and without tool provider stakeholders	31	14%
Practitioners incl. transfer practitioners and industrial consultants, all with academic background	AEB cut with researchers in industry, consulting and managing practitioners, external consultants, and engineering practitioners in industry	86	40%
Pure academics	AEB without respondents specifying additional roles	66	31%
Respondents not specifying an educational background (NEB)	The complement of AEB	60	27%
Respondents not specifying an educational background and being researchers in industry	NEB intersected with researchers in industry	13	6%
Consultants	NEB intersected with consulting or managing practitioners and external consultants	23	11%
Pure practitioners	NEB cut with engineering practitioners in industry	23	11%
Tool provider stakeholders not specifying an educational background	NEB cut with stakeholder of an FM tool or service provider	5	2.3%
Non-academic FM non-users	NEB cut with “I have not used FMs in any specific role.”	19	9%
Practitioners incl. industrial consultants	Consulting and managing practitioners, external consultants, and engineering practitioners in industry	108	50%
FM users (all)	Respondents having used FMs in one or another way and context in the past	212	98%
FM users (beyond students)	Excl. “only-in-course” respondents	202	93.5%

Table 8 (continued)

Category of respondents to ...	Description	Count	Fraction
... Question Q1:			
FM non-users	Respondents who chose "I have not used FMs in any academic or industrial domain."	36	17%
... Question Q3:			
Respondents with no motivation	Respondents who selected "no" for all motivating factors	9	4%
... Questions Q5 and Q6:			
Non-practitioners including FM non-users	Respondents who chose "no experience or no knowledge", "studied in (university) course" or "applied in lab, experiments, case studies" for all FM techniques. This group includes laypersons.	46	21%
Sample (N)	All valid responses	216	100%

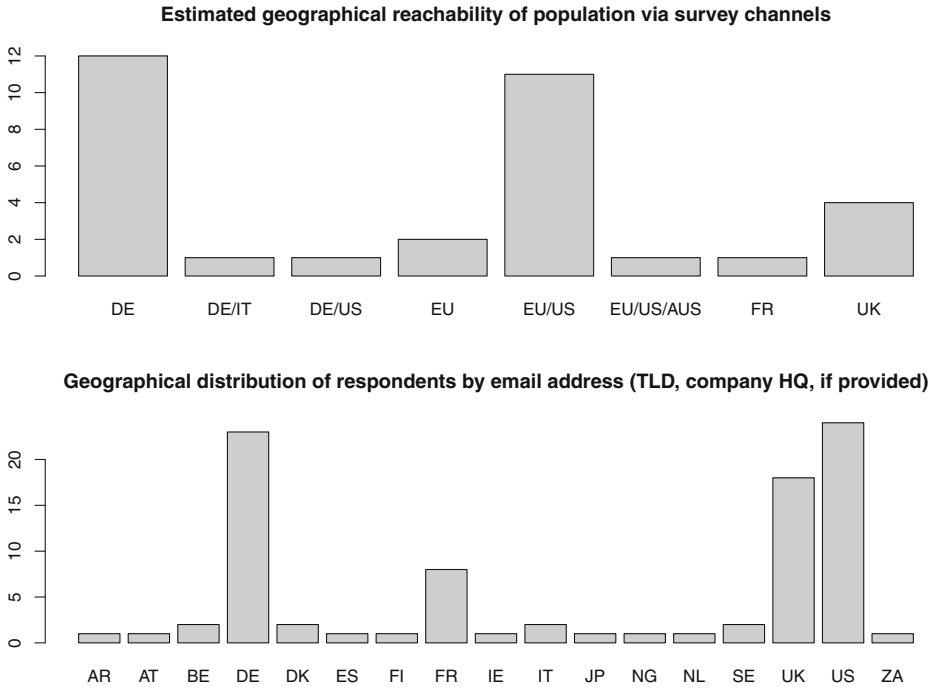


Fig. 19 Geographical analysis of the sample. *Legend:* top-level domain (TLD), head quarter (HQ)

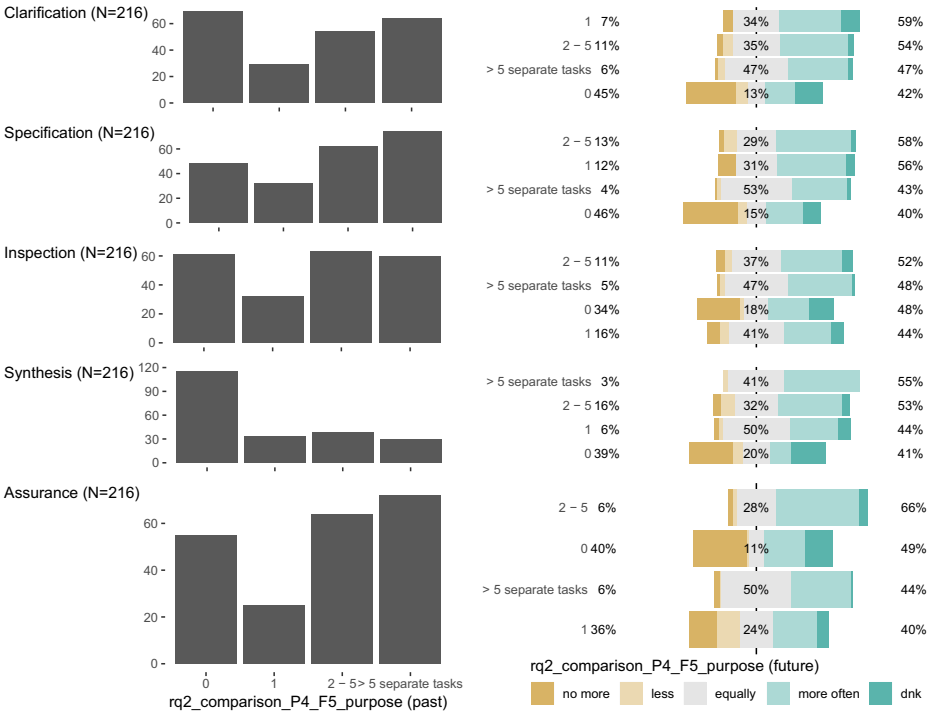


Fig. 20 Comparison of past and future usage intent by purpose

The left column describes for each purpose (e.g. specification) how often (e.g. in 2 to 5 separate tasks) respondents have used FMs in the past (UFM_p).

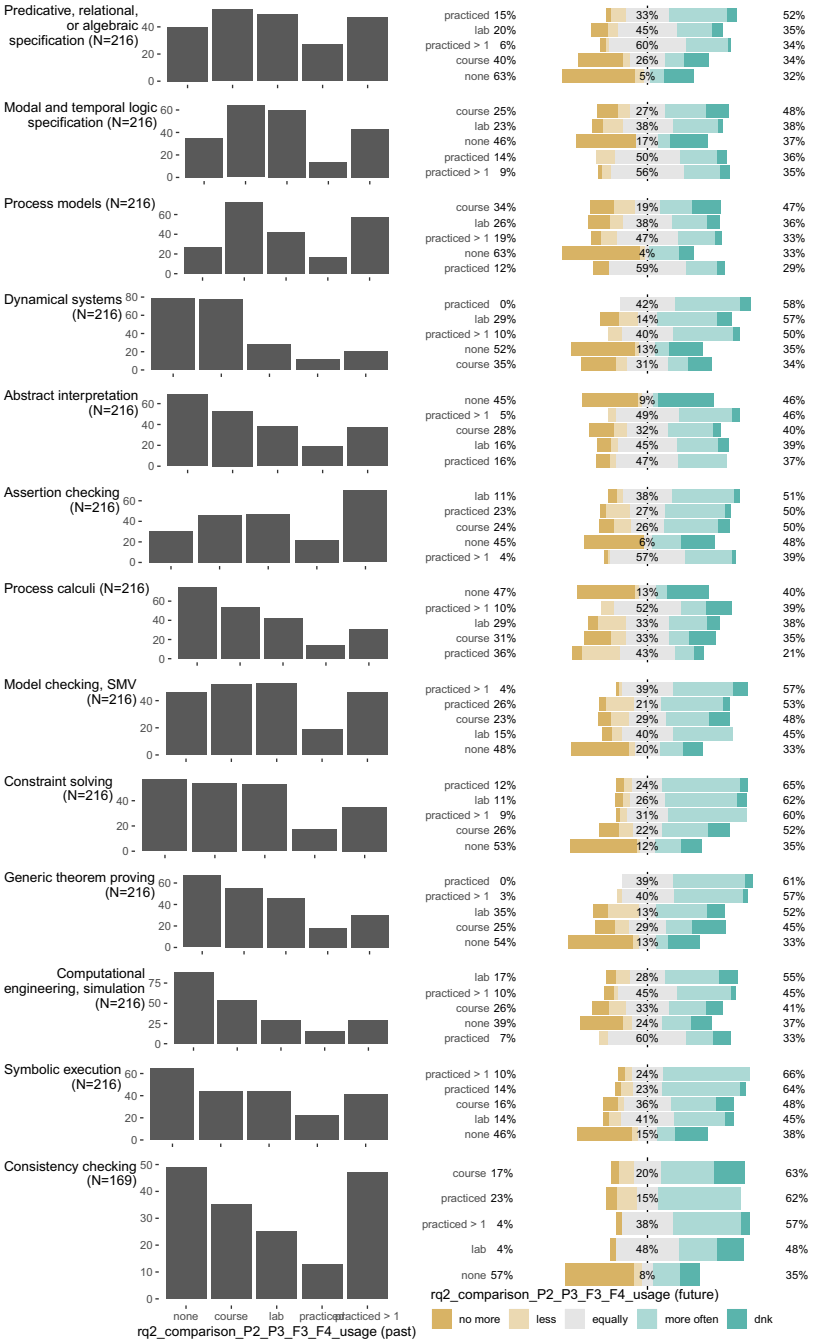
The right column describes for each purpose the usage intent (UFM_i) depending on how often respondents have used FMs in the past (UFM_p). The horizontal bars representing the UFM_p frequency categories are listed in descending order by the overall size of both UFM_i groups "more often" and "dnk". We chose to keep *dnk*-answers visible despite the readability inconvenience caused by the *dnk*s influencing the ordering. However, in the majority of cases the largest group of respondents intending to increase FM use in the future is visible first or near the top.

A.4 Code-based vs. Model-based FMs for Assurance vs. Inspection

The data for this comparison is summarised in Fig. 21.

Fig. 21 Comparison of past and future usage for code-based (top half) and model-based FMs (bottom half) by purpose

A.5 Usage Intent (UFM_i) by FM Class (for Analysis of RQ2)



A.6 Data for the Analysis of RQ3

Figure 22 and the following figures in this section show pairs of matrices, so-called "heat-maps", useful for association analysis between categorical and ordinal variables. The cells

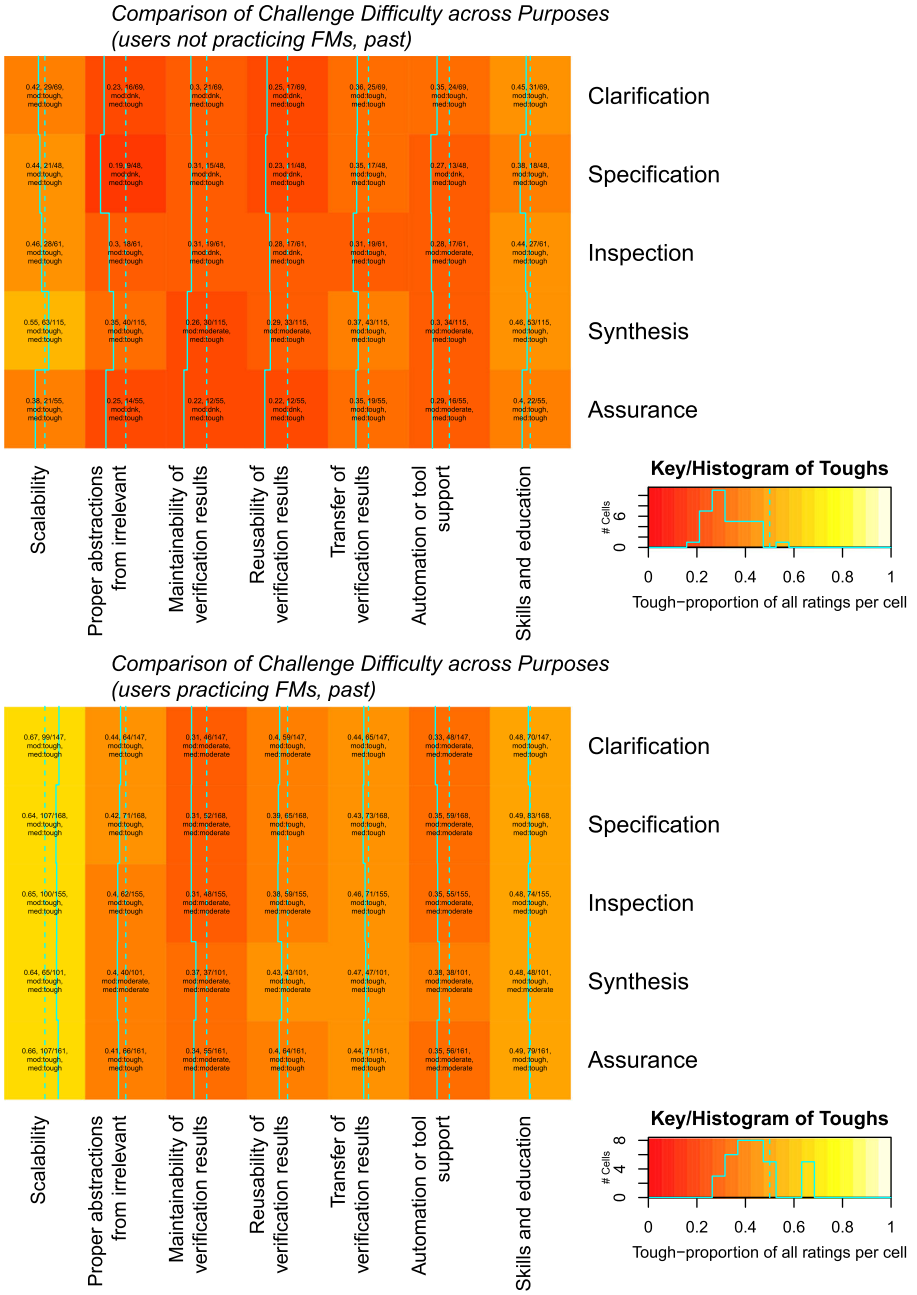


Fig. 22 Comparison of challenge difficulty across purposes (UFM_p)

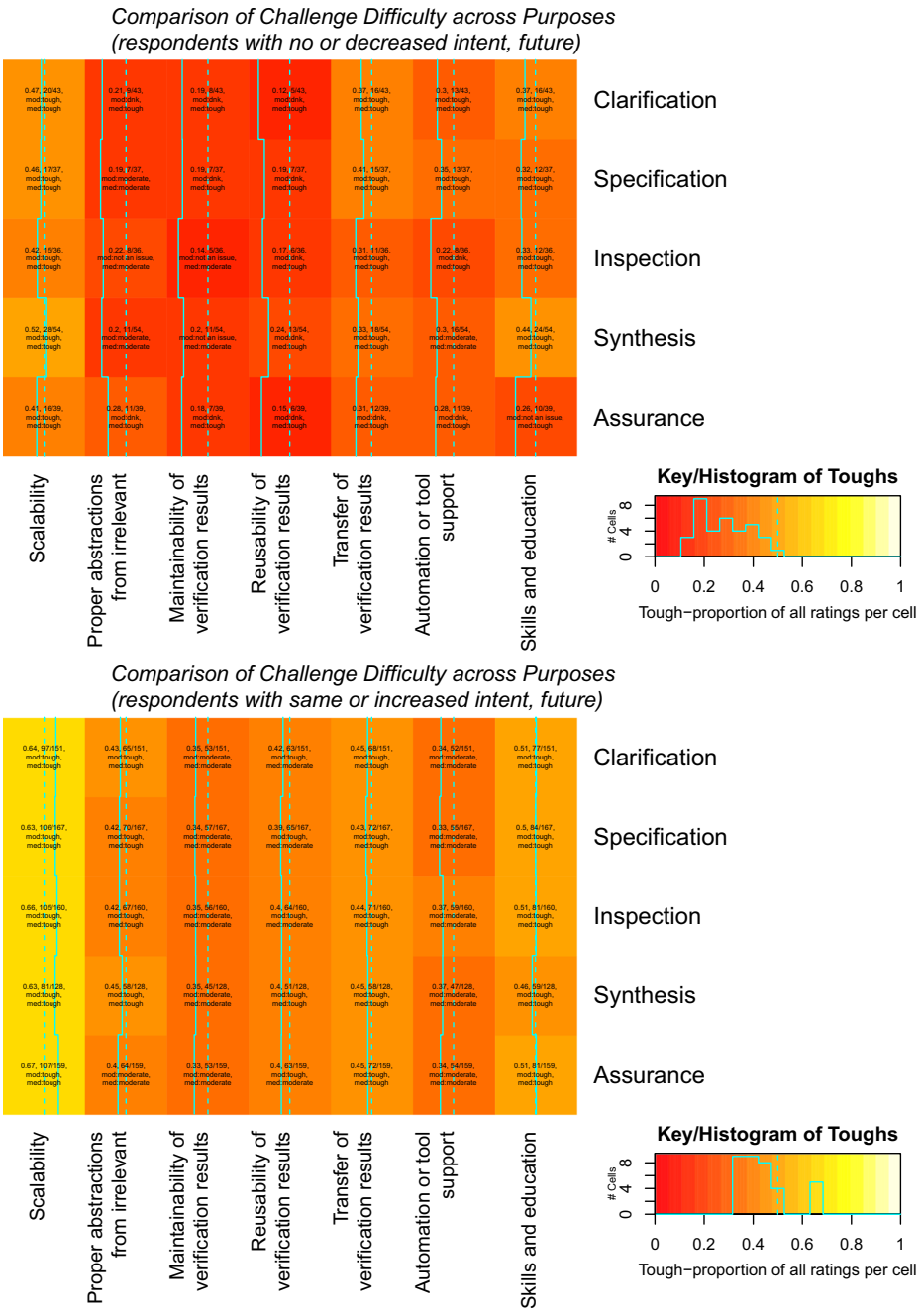


Fig. 23 Comparison of challenge difficulty across purposes (UFM_i)

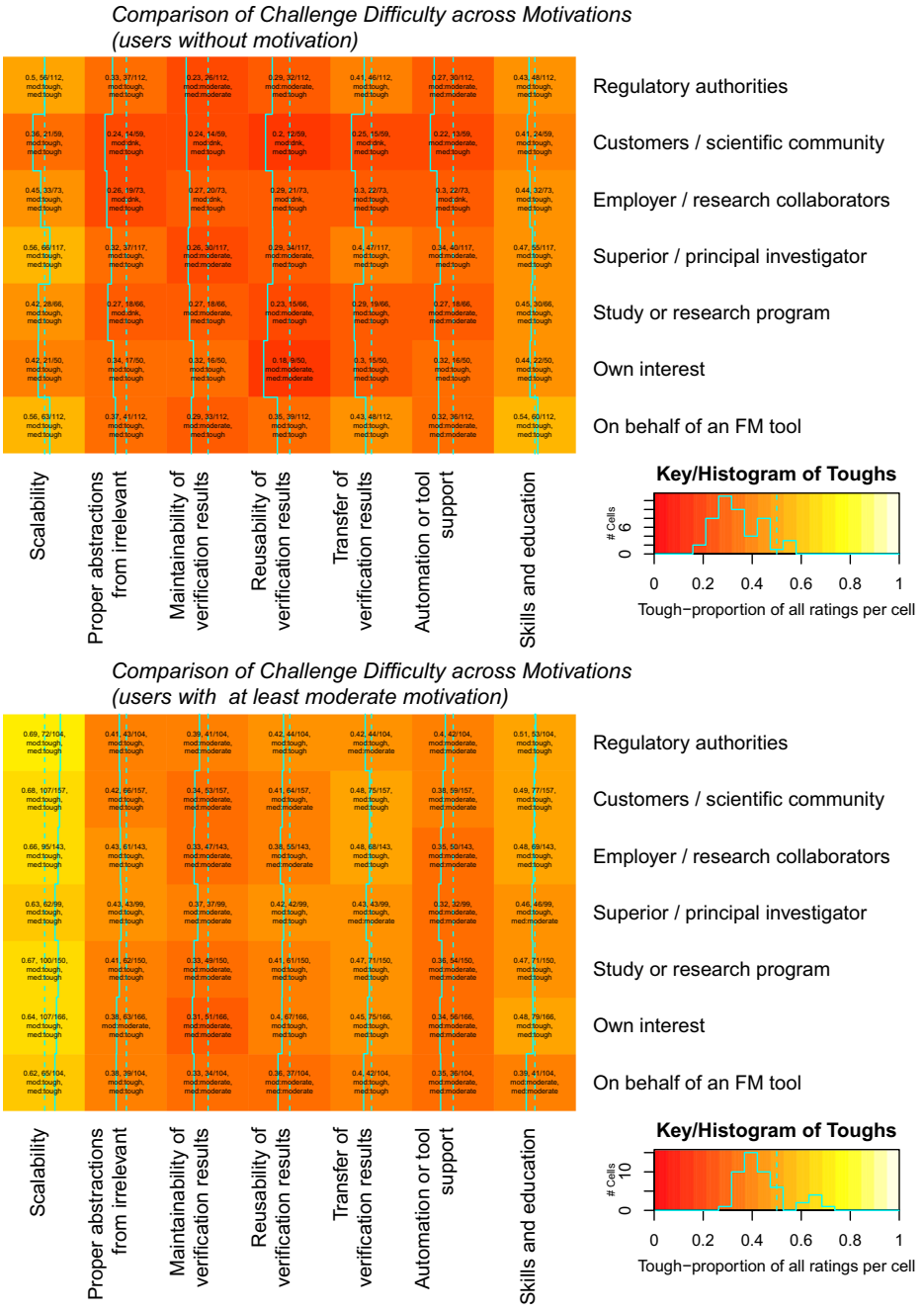


Fig. 24 Comparison of challenge difficulty across motivations

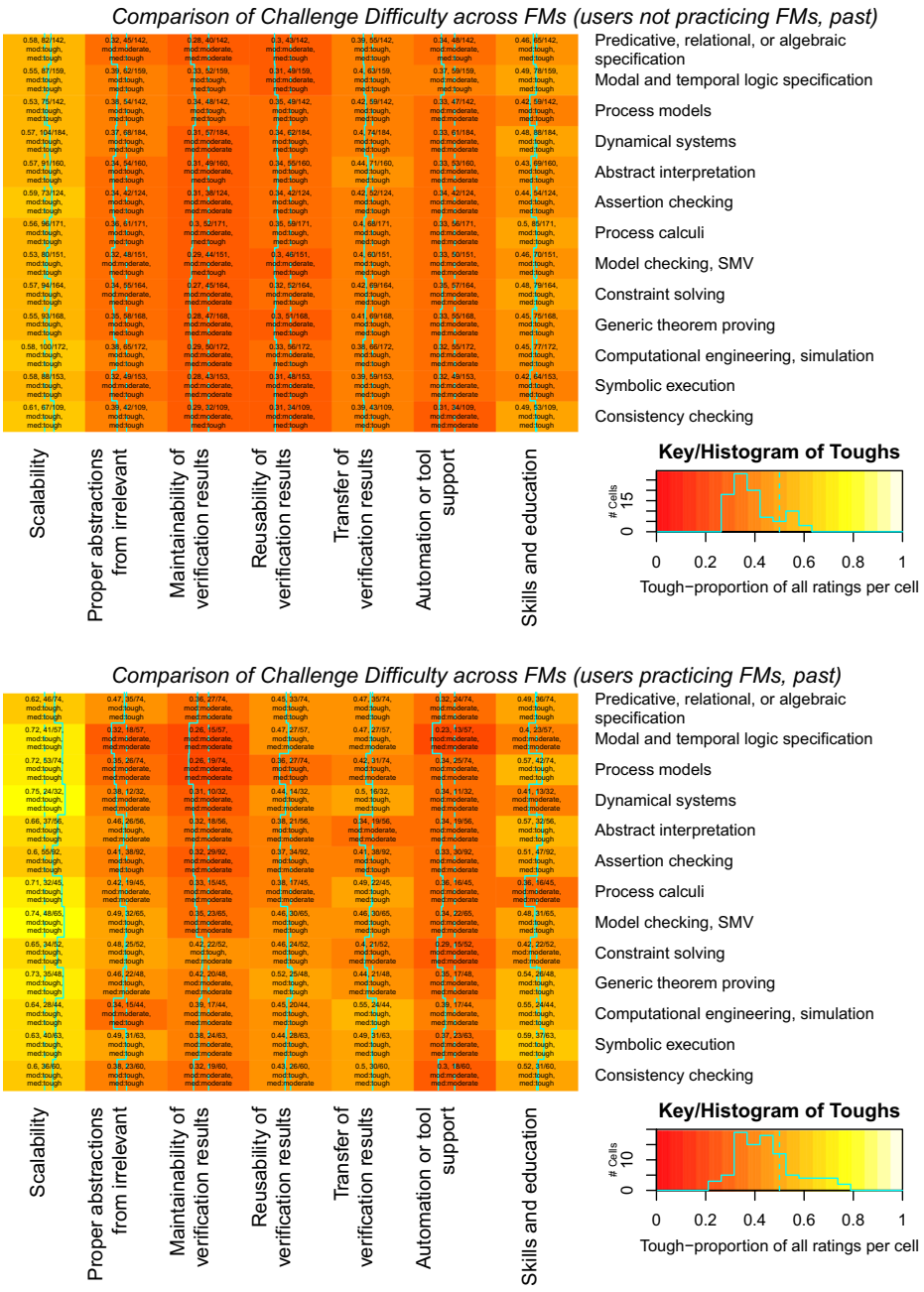


Fig. 25 Comparison of challenge difficulty across FM classes (UFM_p)

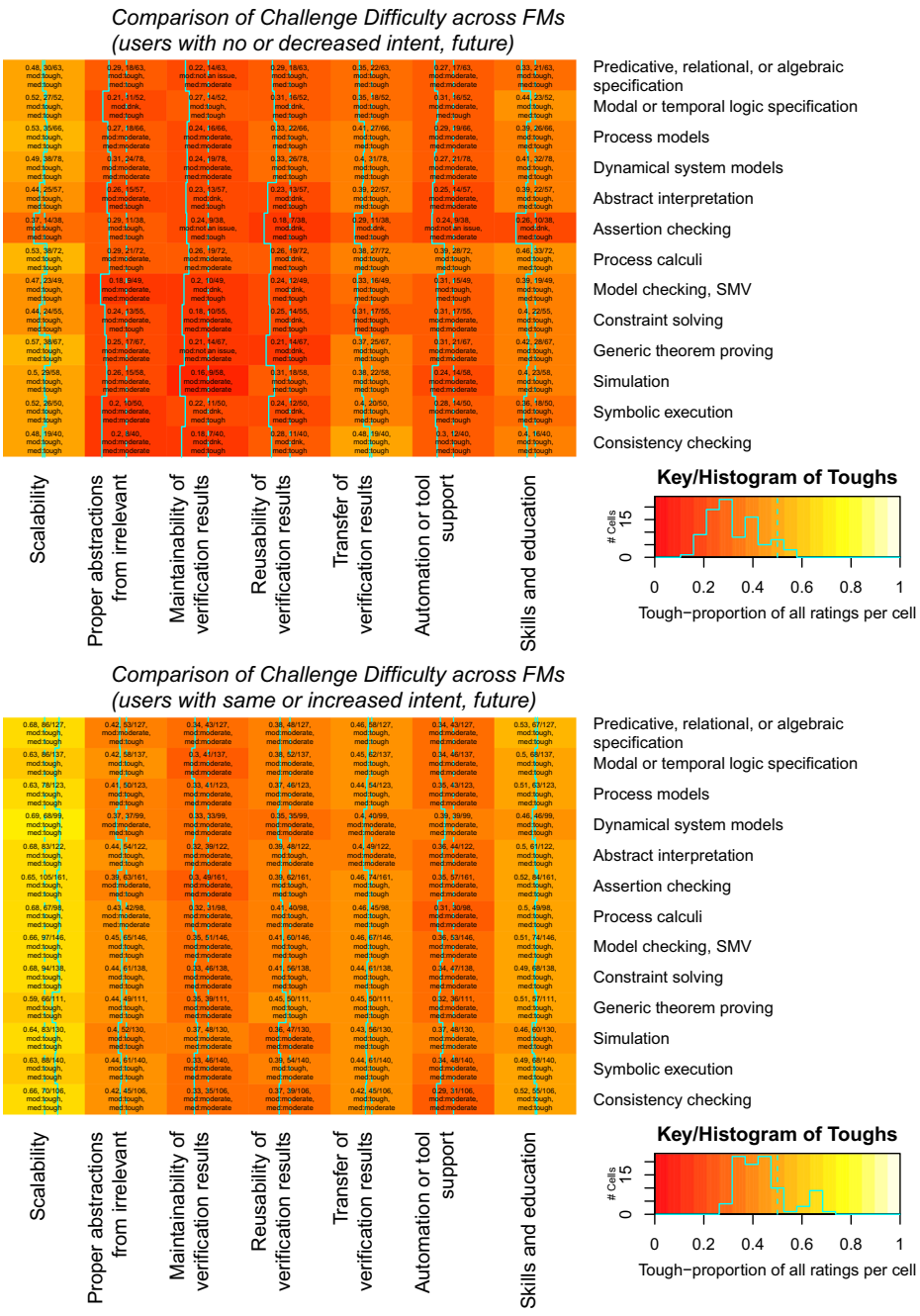


Fig. 26 Comparison of challenge difficulty across FM classes (UFM_i)

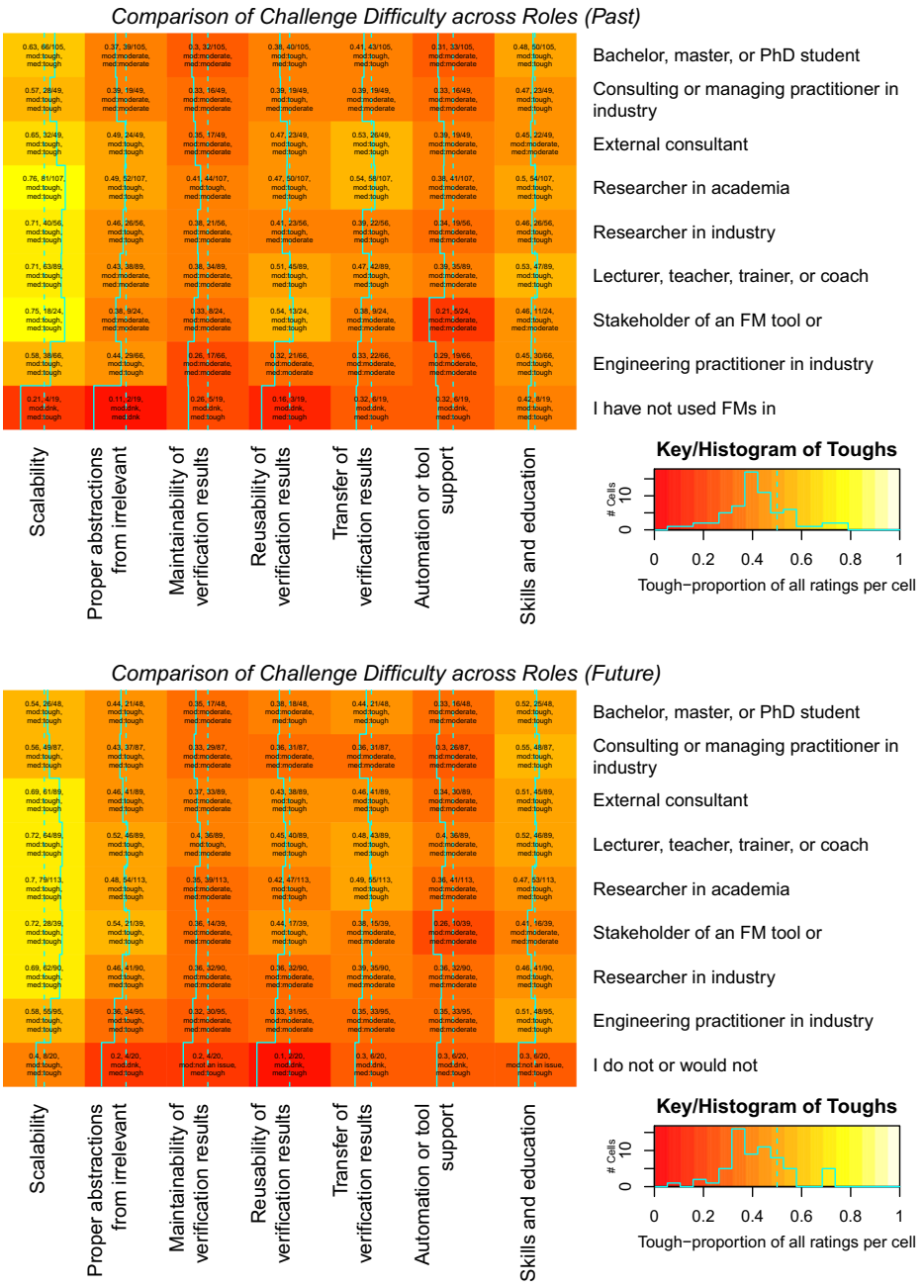


Fig. 27 Comparison of challenge difficulty across roles

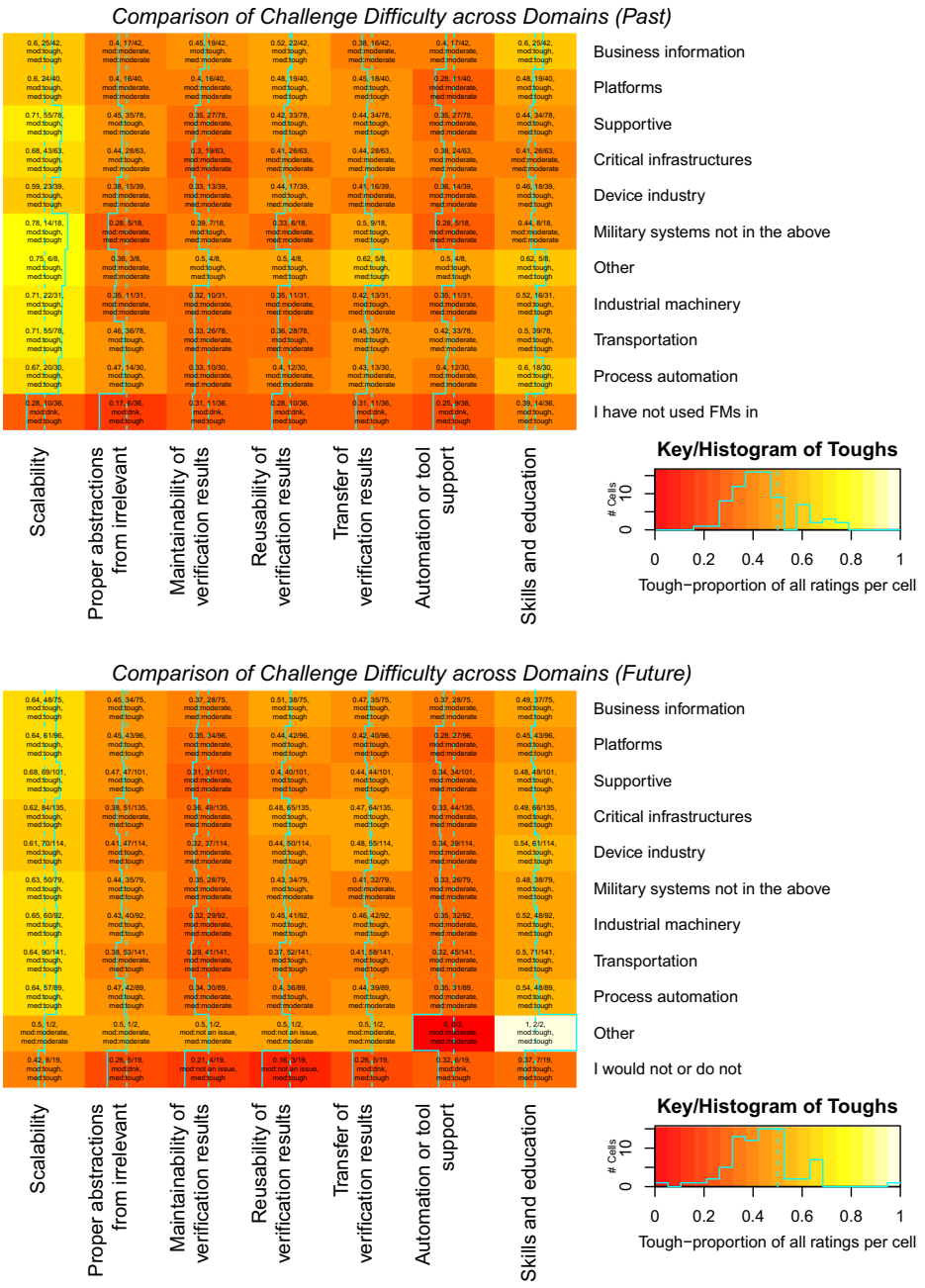


Fig. 28 Comparison of challenge difficulty across domains

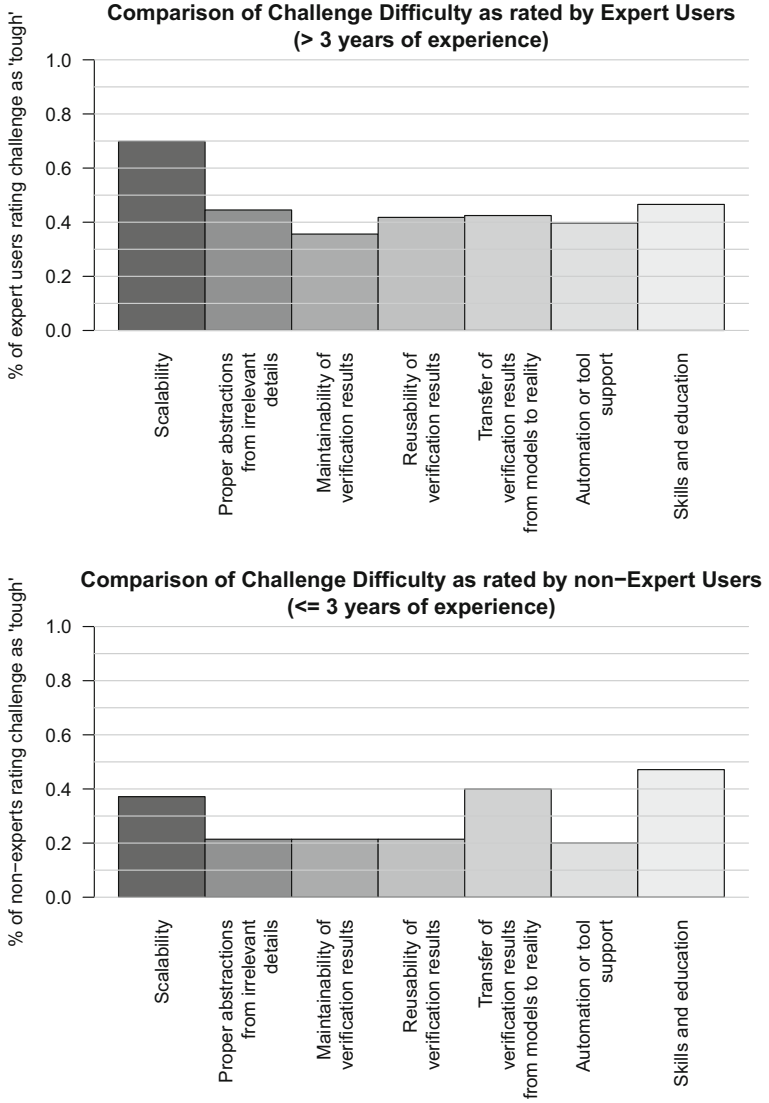


Fig. 29 Comparison of expert and non-expert users by their perception of challenge difficulty

in the matrices represent combinations of the scales, each cell containing data about the *mode* and *median* of "degree of difficulty" ratings, their *proportion* of *tough* ratings, and the *actual numbers* of data points. Both the colour gradient (red to white) and the solid vertical lines in the cells represent the tough proportions (left = 0 to right = 100%), with the dotted vertical line signifying the 50% margin.

A.7 Details on the Systematic Map

Table 9 contains the data we collected from the literature for the systematic map.

Table 9 Details for the classification of related work

Reference	Motivation	Approach	Result	Relation	List of obstacles
Gerhart and Yelowitz (1976)	Pinpoint fallibility in the use of FMs	Evaluation of methodologies, error classification and analysis, using several examples of fundamental algorithms	Identification of three classes of errors: specification errors, systematic construction errors, proved program errors. Discussion of potential causes of errors of each class.	Observations contain obstacles, recommendations their alleviation	Formality gap, appropriate abstraction and structuring, lack of skills and education
Jackson (1987)	Generic evaluation	Expert opinion	FMs have inherent limitations	Aim of method transfer, design of applicable FMs	Inherent informality of formalisation, difficult to communicate to customers, lack of expressiveness/freedom of expression, lack of methodology
Bjorner (1987)	Proposes a software Development method based on FM	Personal opinion, experience	Identify 3 main challenges: education, hiring, and tool support.	Challenges could be considered obstacles	Lack of skills and education, lack of tools, changeability/compatibility with existing process (method culture)
Hall (1990)	Present and test FM myths.	They evaluate FMs on one larger case study (50.000 lines of Objective C Code) where they use Z (550 Z schemas to define 280 operations) to develop a CASE tool.	Formal methods are powerful tools which must be better understood by developers at large.	Rejection of common Hypothesis about FMs. Evaluation of a single FM by means of a case study.	Myth 1: improper abstraction; transfer of v.results; (myth 2-4: skills and directed education;) myth 5: time-budget restrictions; myth 6: improper abstraction, tool support (usability); myth 7: scalability

Table 9 (continued)

Reference	Motivation	Approach	Result	Relation	List of obstacles
Wing (1990)	FM adoption	Literature review, summary, and analysis	Overview/taxonomy of FMs; analysis of limitations	Justifies the FM classification used in our questionnaire	Proper abstraction (formality gap, neglected environmental assumptions)
Bloomfield et al. (1991)	Introduction to formal methods	Reference to technology transfer, existing case studies, and tool support	At the present time, formal methods are good for the description of sequential properties of systems, and for communication protocols, although they do not yet address temporal properties and concurrency particularly well.	Investigation of state of the art, no comparison with future.	Handling of incomplete specs, lack of verification tools, costly training, changing management style
Austin and Parkin (1993)	Lack of acceptance in industry	Literature survey and questionnaire	Identify obstacles and suggest to improve education and standardisation and to perform case studies and define metrics.	Our questionnaire has less focus on representation and methodology and excludes questions on benefits, suggestions. Their sample mainly covers Z/VDM users in the UK. Our analysis of past use is more elaborate.	Math, tools, lack of cost/benefit evidence, change resistance

Table 9 (continued)

Reference	Motivation	Approach	Result	Relation	List of obstacles
Craigen et al. (1993)	Determine state of the art about the use of FM in practice	Analysis of 12 case studies	FMs are beginning to be used seriously and successfully by industry	Study of FMs in industrial practice	Stated as recommendations: scalability, lack of tool support, lack of skills/education, transfer of verif. obl./results from/to code; resource constraints
Fraser et al. (1994)	Lack of FM adoption, improvement of RE	Discuss benefits and problems of FM adoption, literature study	Present a two-dimensional framework for assessing strategies for incorporating formal specifications in software development	Add suggestions on FM introduction	Lack of method and tool support, lack of skills/training, not suitable for requirements prototyping, lack of cost/benefit evidence
Bowen and Hinchey (1995a)	Re-examine Hall's myths, introduce 7 new Myths.	Argumentation and mentioning of case studies (although no reference to the studies are given).	More real links between industry and academia are required, and the successful use of formal methods must be better publicized.	Rejection of common Hypothesis about FMs. Evaluation with reference to case studies.	Time-budget-restrictions, lack of tool support, lack of integration in current process, scalability (multi-tech)
Bowen and Hinchey ((Bowen and Hinchey 1995b))	Identify maxims that may help in the application of formal methods in an industrial setting.	Based on observations (by ourselves and others) on a number of recently completed and in-progress projects	10 Hypotheses on how to improve the success of FM usage	Investigation of FM usage; Argumentation and Examples.	Stated as commandments: lack of tool support (documentation guidelines), proper abstraction, budget restrictions (bad cost-benefit ratio), lack of experts (skills and education); compatibility with current process (lack of quality culture); lack of reuseability

Table 9 (continued)

Reference	Motivation	Approach	Result	Relation	List of obstacles
Lai and Leung (1995)	Investigate reasons why academic methods are not adapted by industry	Personal experience	Provide 5 reasons: practicality, too academic, Education, Resistance to change, difficulty to re-invest	The reasons can be considered as obstacles	Practicality (being too academic): scalability, skills and education, resistance to change (compat with existing process); budget restrictions
Heisel (1996)	In practice FMs are not widely used	Personal Opinion/Observation	Proposes a pragmatic approach to FM	Method tries to overcome obstacles	Lack of skills/education, lack of experts, improper abstraction (low useability, low correctness); budget restrictions; compatibility with existing process
Hinchey and Bowen (1996)	Identify reasons for industry's reluctance to take formal methods to heart.	Experience in editing a collection of essays on the industrial application of FM.	Misconception of Myths, Standards, Tools, and Education are obstacles to use FM in industry.	They identify obstacles. However, they to not provide evidence against or in favor of them.	Wrong skills and education, lack of clarification (bad reputation/misconception); lack of standards/regulation; lack of tools;
Holloway and Butler (1996)	FM adoption	Position statement, experience report, expert opinion	List of impediments to FM adoption	Do not measure usage intent but highlight lack of transfer efforts	Inadequate tools and examples, inadequate transfer
Lai (1996)	Academic methods (FM) not used in communication industry	Personal opinions; critize research transfer and suggest improvements that might also be helpful for FM transfer to practice	8 Reasons why academia needs to do industry research, 7 catalysts, 12 industry relevant factors	Reasons can be considered obstacles	Lack of empirical evidence, lack of skills/education, scalability, proper abstractions; compatibility with existing process, lack of tool support

Table 9 (continued)

Reference	Motivation	Approach	Result	Relation	List of obstacles
Pfleeger and Hatton (1997)	Investigate effectiveness of FMs	Case study and effect analysis: Comparison of change requests for FM-based and non-FM-based code fragments as a result of postdelivery problems caused by these fragments.	FM have a positive effect on code quality	Empirical evidence for FM effectiveness shown for FMs in design phase but not in more general, however only one system	Cost-benefit (fault removal effectiveness)
Knight et al. (1997)	Formal Methods are not accepted by industry	Elementary Field Experiment	Evaluation of Z, PVS, and state-charts	Evaluation of FMs in practice	Integration in existing process (tools, methods, environments, people); proper and useful abstractions (incl. usability/comprehensibility); tool support (group development, collaborative eng.); evolution and spec/proof maintainability; budget constraints
Heitmeyer (1998)	FM tools are useful but not used in industry since people are not skilled enough	Reference to a few case studies	Provide a set of guidelines how to make FM more usable	Usability as obstacle	Tool support, proper abstraction, process changeability/compatibility

Table 9 (continued)

Reference	Motivation	Approach	Result	Relation	List of obstacles
Snook and Harrison (2001)	Lack of empirical investigation on the use of FM	5 Structured Interviews	Improved quality of software with little or no additional lifecycle costs	Empirical investigation of the benefits of FM in industry	Lack of skills/education, improper abstractions (understandability); transfer of verif.results (models to reality)
Bowen and Hinchey (2005)	Re-examination of their 1995 commandments 10 Years later	Personal experience, reference to literature	empirically validate commandments with little conclusion	They investigate the use of FM in industry	Tool support still an issue despite some case studies
Bicarregui et al. (2009)	No recent study on the industrial use of FMs.	Structured Questionnaire on 62 industrial projects	4 challenges where identified	Empirical Study identifying obstacles	Lack of tool support, lack of empirical evidence; lack of experience (skills/education), budget restrictions
Woodcock et al. (2009)	State of the art of industrial use of FM (extension of Bicarregui et al. (2009))	Structured Questionnaire on 62 industrial projects (claimed to be most comprehensive review ever made of formal methods application in industry)	Identify several challenges	Empirical Study identifying obstacles	Budget/resource constraints (high entry costs, cost-benefit), lack of tool support (automation)
Miller et al. (2010)	FMs are not widely used in Industry	3 Case studies about the use of Model Checking in Industry	Model Checking can be effectively used to find errors early in the development process for many classes of models	They investigate the applicability of one FM in industry	Lessons from case studies: scalability and proper abstraction (useability)

Table 9 (continued)

Reference	Motivation	Approach	Result	Relation	List of obstacles
Parnas (2010)	FMs are not widely used in industry	Argumentation/Personal Experience	Provides reasons why FMs are not used and suggestions for improvement	Provides Obstacles	Proper abstraction, lack of empirical evidence, lack of tool support; maintainability/transfer (refinement,step-by-step)
Mohagheghi et al. (2012)	Adoption of MbE in SE	Tool evaluations based on TAM (Riemenschneider et al. 2002,) interviews, survey	Evaluation of PEOU, PU, current, and future use	Also use IVs current and future use; little focus on FM; but MDE is sometimes based on FM, MDE adoption can improve FM adoption;	Lack of training, lack of maturity, broken tool chains, high cost of adoption (tool integration)
Davis et al. (2013)	Identify barriers to FM adoption and suggestions for barrier mitigation	Interviews with 31 practitioners from the US aerospace domain	Top barriers: education, tools, work environment; top mitigations: education, tool integration, evidence of FM benefits; occasional non-barriers: evidence on savings, FM complexity, training/skills	Similar research questions, open-ended interview questions; restricted to one domain and geography	Education, tools, environment, engineering, certification, misconceptions, scalability, evidence of benefits, cost
Liebel et al. (2016)	Adoption of MbE (incl. FM) in embedded SE	Online survey on needs, positive/negative effects, and shortcomings of MDE adoption	SotA and challenge assessment: FMs not used widely; their data suggests a need of FM adoption; 30% of the responses from industry declare the need for FMs as a reason to adopt MDE; median of responses suggests that MbE adoption has a positive effect on FM adoption;	Few FM users as participants processes	Lack of tool support, bad reputation, rigid development

Table 9 (continued)

Reference	Motivation	Approach	Result	Relation	List of obstacles
Ferrari et al. (2019)	Lack of FM adoption in railway domain	Review of FM literature, FM projects, and FM tools according to DESMET (Kitchenham et al., 1997); survey among practitioners	UML dominates as the MbE language, many FMs and FM-based tools are used, B dominates as the FM; tool ranking/selection matrix	Analyse maturity of FMs from literature review; evaluate relevance/quality/maturity of FM and FM tool features from subjective assessment of survey respondents;	Difficulty to learn, lack of tool qualifications, lack of expressiveness
Klein et al. (2018)	FM adoption	Large case study, measurement of proof effort	FMs can scale to real systems, mixed assurance levels are possible	Evidence for scalability contradicting the belief of our responses	Incompleteness of theorems for abstracting from all hardware features

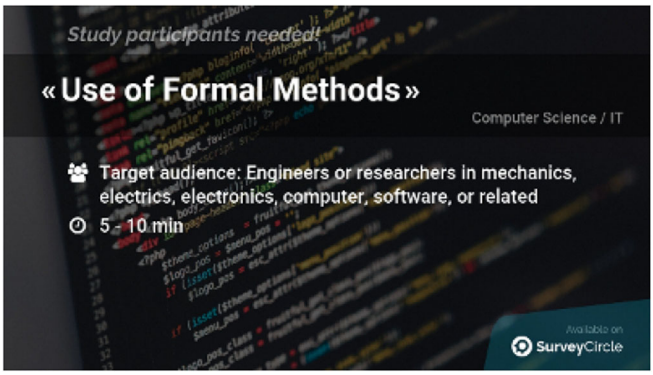
A.8 Mapping of Studies to Challenges for RQ3

In addition to Table 6 in Section 5.5, Table 10 provides the complete lists of surveyed studies mapped to the corresponding challenges.

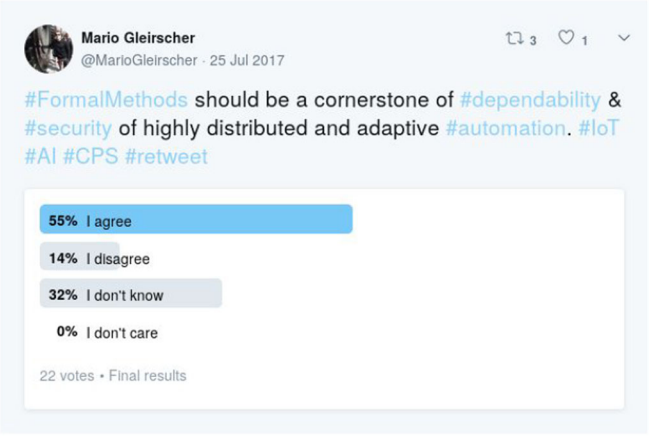
Table 10 Mapping of studies to challenge names (with the number of studies in parentheses)

Challenge name	Supported by
Scalability (7)	Hall (1990), Miller et al. (2010), Bowen and Hinchey (1995a), Lai and Leung (1995), Lai (1996), Craigen et al. (1993), and Craigen et al. (1995)
Skills & Education (12)	Bjorner (1987), Bicarregui et al. (2009), Hall (1990), Barroca and McDermid (1992), Hinchey and Bowen (1996), Bowen and Hinchey (1995b), Lai and Leung (1995), Lai (1996), Heisel (1996), Craigen et al. (1993), Craigen et al. (1995), and Snook and Harrison (2001)
Transfer of Proofs (8)	Jackson (1987), Parnas (2010), Hall (1990), Craigen et al. (1993), Craigen et al. (1995), Snook and Harrison (2001), Bloomfield et al. (1991), and Barroca and McDermid (1992)
Reusability (2)	Barroca and McDermid (1992) and Bowen and Hinchey (1995b)
Abstraction (11)	Jackson (1987), Parnas (2010), Miller et al. (2010), Hall (1990), Barroca and McDermid (1992), Bowen and Hinchey (1995b), Lai (1996), Heitmeyer (1998), Heisel (1996), Knight et al. (1997), and Snook and Harrison (2001)
Tools & Automation (16)	Bjorner (1987), O'Hearn (2018), Hall (1990), Bloomfield et al. (1991), Bowen and Hinchey (1995a), Hinchey and Bowen (1996), Bowen and Hinchey (1995b), Bowen and Hinchey (2005), Bicarregui et al. (2009), Woodcock et al. (2009), Parnas (2010), Lai (1996), Heitmeyer (1998), Craigen et al. (1993), Craigen et al. (1995), and Knight et al. (1997)
Maintainability (3)	Barroca and McDermid (1992), Knight et al. (1997), and Parnas (2010)
Resources (11)	Hall (1990), Woodcock et al. (2009), Craigen et al. (1993), Craigen et al. (1995), Bloomfield et al. (1991), Bowen and Hinchey (1995a), Bowen and Hinchey (1995b), Lai and Leung (1995), Heisel (1996), Knight et al. (1997), and Bicarregui et al. (2009)
Process Compatibility (12)	Bjorner (1987), O'Hearn (2018), Bloomfield et al. (1991), Bowen and Hinchey (1995a), Bowen and Hinchey (1995b), Lai and Leung (1995), Hinchey and Bowen (1996), Lai (1996), Heitmeyer (1998), Heisel (1996), Knight et al. (1997), and Craigen et al. (1995)
Practicality & Reputation (5)	Lai and Leung (1995), Parnas (2010), Lai (1996), Glass (2002), and Bicarregui et al. (2009)

A.9 Copy of the Advertisement Flyer



A.10 Screenshot of the Twitter Poll



A.11 Copy of the Questionnaire

The PDF export of our on-line questionnaire **on the next page** corresponds to the questionnaire we used for the sample taken until 31.3.2019 with $N = 220$. Since 26.5.2019, an extended version of the questionnaire had been available online at

https://goo.gl/forms/FnKNQtTmI3A6BekM2.

We crafted this questionnaire using Google Forms (Google 2018). We use numbered identifiers for each question category, demographic questions are prefixed with a "D", questions about past FM use (UFM_p) with a "P", about future or intended FM use (UFM_i) with an "F", questions about obstacles with an "O". Open questions are suffixed by an "o".

Use of Formal Methods

Dear participant, thank you for your interest in this 8-10min survey on the use of formal methods (FMs).

This survey does NOT require previous knowledge in FMs or in their actual application in a practical context. However, this survey targets persons with an educational background in engineering and sciences OR with a practical engineering background in a reasonably critical systems or product domain.

By "FMs", we refer to explicit mathematical models and sound formal logical reasoning about critical properties---such as reliability, safety, availability, data privacy or, more generally, dependability and security---of electrical, electronic, and programmable electronic or software systems in critical application domains. FMs include, for example, formal specification, theorem proving, model checking, formal contracts, SMT solving, process algebras.

By "use of FMs", we refer to the application of FMs to engineered systems in the context of education, research, and, particularly, the field of industrial practice and by using formal languages together with manual or automated tool-based techniques.

This survey is anonymous. However, you can provide your email address if you are interested in receiving our final results afterwards.

The underlying study is conducted by Mario Gleirscher at University of York and Diego Marmsoler at Technical University of Munich.

* Erforderlich

Demographic Questions

1. **D1. In which application domain(s) in industry or academia (if any) have you mainly used FMs?** *

Wählen Sie alle zutreffenden Antworten aus.

- [ ] I have not used FMs in any academic or industrial domain.
- [ ] Critical infrastructures (e.g. telecom, energy, road/air/naval/rail traffic, smart buildings or cities)
- [ ] Process automation (e.g. chemical process plants, power plants, warehouse logistics, production lines)
- [ ] Industrial machinery (e.g. stationary robotics, production machines)
- [ ] Transportation (e.g. automotive, utility vehicles, naval, aeronautics, train systems, freight logistics, cable cars, mobile robotics, drones/UAVs)
- [ ] Device industry (e.g. medical, health-care, semi-conductors, consumer electronics)
- [ ] Military systems not in the above domains (e.g. for command, control, surveillance)
- [ ] Business information (e.g. database applications, banking, finance, ERP, PLM, web services, cloud apps)
- [ ] Platforms (e.g. operating systems, middle-ware, firmware, drivers, database systems, libraries)
- [ ] Supportive (e.g. CASE tools, checking or verification tools, CAD/CAM systems)
- [ ] Sonstiges: _____

2. **D2. How many years of FM experience (including the study of FMs) have you gained?** *

Markieren Sie nur ein Oval.

- ◯ 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

3. **D3. Which have been your motivations (if any) to use FMs?** *
   Markieren Sie nur ein Oval pro Zeile.

	no motivation	moderate motivation	strong motivation (or requirement)
Regulatory authorities	◯	◯	◯
Customers / scientific community	◯	◯	◯
Employer / research collaborators	◯	◯	◯
Superior(s) / principal investigator(s)	◯	◯	◯
Study or research program	◯	◯	◯
Own (private) interest	◯	◯	◯
On behalf of an FM tool or service provider	◯	◯	◯

4. **D3o. Which have been your further motivations to use FMs (if any)?**

Past and Current Use of Formal Methods

The following questions aim at your EXPERIENCE with the use of FMs in your PAST and CURRENT activities and projects.

NOTE: If you are not able to say anything about past or current use, please, choose the corresponding "not yet used...", "no experience...", or "not at all" options and proceed to the next page!

5. **P1. In which role(s) have you used FMs?** *
   Wählen Sie alle zutreffenden Antworten aus.

- ☐ I have not used FMs in any specific role.
- ☐ Engineering practitioner in industry (e.g. programmer)
- ☐ Consulting or managing practitioner in industry (e.g. architect, requirements or systems engineer)
- ☐ External consultant (e.g. external requirements or systems engineer)
- ☐ Researcher in industry
- ☐ Researcher in academia
- ☐ Lecturer, teacher, trainer, or coach
- ☐ Bachelor, master, or PhD student
- ☐ Stakeholder of an FM tool or service provider
- ☐ Sonstiges:

Experience in Formal Methods Use

6. **P2. Describe your level of experience with each of the following classes of formal description techniques?** *

Markieren Sie nur ein Oval pro Zeile.

	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
Predicative, relational, or algebraic specification	◯	◯	◯	◯	◯
Modal and temporal logic specification	◯	◯	◯	◯	◯
Process models (e.g. Petri nets, Mealy machines, LTS, Markov processes)	◯	◯	◯	◯	◯
Dynamical systems (i.e. differential equations)	◯	◯	◯	◯	◯

7. **P3. Describe your level of experience with each of the following classes of formal reasoning techniques?**
   Markieren Sie nur ein Oval pro Zeile.

	no experience or no knowledge	studied in (university) lectures	applied in lab, experiments, case studies	applied once in engineering practice	applied several times in engineering practice
Abstract interpretation					
Assertion checking (e.g. for pre/post specification, contracts)					
Process calculi (e.g. CSP, CCS, pi, mu, hybrid)					
Model checking, SMV (of e.g. temporal or probabilistic properties)					
Constraint (SAT, SMT) solving (e.g. for static code analysis), optimisation techniques					
Generic (first-order, HOL) theorem proving (using e.g. term rewriting, functional programming)					
Computational engineering, simulation (using e.g. differential calculus, numerical methods)					
Symbolic execution (e.g. scenario testing, model animation)					
Consistency checking (e.g. syntax or bug pattern checking)					

9. **P4. I have mainly used FMs for ...** *
Markieren Sie nur ein Oval pro Zeile.

	... not at all.	... once.	... in 2 to 5 separate tasks.	... in more than 5 separate tasks.
... clarification (i.e. explicit description for analyzing a problem)	⬭	⬭	⬭	⬭
... specification (e.g. contracts, documentation and communication of requirements and design)	⬭	⬭	⬭	⬭
... inspection (i.e. error detection, e.g. non-conformance checking, model-based testing)	⬭	⬭	⬭	⬭
... synthesis (e.g. transformation, compilation)	⬭	⬭	⬭	⬭
... assurance (e.g. error removal, property verification, refinement or equivalence proofs, argumentation)	⬭	⬭	⬭	⬭

10. **P4o. I have used FMs for other purposes (if any):**

Intended Future Use of Formal Methods

The following questions aim at your INTENT to use FMs in your FUTURE activities and projects.

NOTE: Your intend to use FMs will also be interpreted as the "mere possibility of FM usage in the corresponding ways" according to and based on your responses.

Future Extent of Formal Methods Use

11. **F1. In which application domain(s) in industry or academia (if any) would (or do) you intend or recommend to use FMs?** *
Wählen Sie alle zutreffenden Antworten aus.

- [ ] I would not or do not intend (or recommend) to use FMs in any academic or industrial domain.
- [ ] Critical infrastructures (e.g. telecom, energy, road/air/naval/rail traffic, smart buildings or cities)
- [ ] Process automation (e.g. chemical process plants, power plants, warehouse logistics, production lines)
- [ ] Industrial machinery (e.g. stationary robotics, production machines)
- [ ] Transportation (e.g. automotive, utility vehicles, naval, aeronautics, train systems, freight logistics, cable cars, mobile robotics, drones/UAVs)
- [ ] Device industry (e.g. medical, health-care, semi-conductors, consumer electronics)
- [ ] Military systems not in the above domains (e.g. for command, control, surveillance)
- [ ] Business information (e.g. database applications, banking, finance, ERP, PLM, web services, cloud apps)
- [ ] Platforms (e.g. operating systems, middle-ware, firmware, drivers, database systems, libraries)
- [ ] Supportive (e.g. CASE tools, checking or verification tools, CAD/CAM systems)
- [ ] Sonstiges: _____

12. **F2. In which role(s) would (or do) you intend to use FMs?** *
Wählen Sie alle zutreffenden Antworten aus.

- [ ] I do not or would not intend to use FMs in any specific role.
- [ ] Engineering practitioner in industry (e.g. programmer, test or verification engineer)
- [ ] Consulting or managing practitioner in industry (e.g. architect, requirements or systems engineer)
- [ ] External consultant (e.g. external requirements or systems engineer)
- [ ] Researcher in industry
- [ ] Researcher in academia
- [ ] Lecturer, teacher, trainer, or coach
- [ ] Bachelor, master, or PhD student
- [ ] Stakeholder of an FM tool or service provider
- [ ] Sonstiges: _____

13. **F3. I (would) intend to use ...** *

Markieren Sie nur ein Oval pro Zeile.

	... 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.
... predicative, relational, or algebraic specification					
... modal or temporal logic specification					
... process models (e.g. Petri nets, Mealy machines, LTS, Markov processes)					
... dynamical system models (i.e. differential equations)					

14. **F4. I (would) intend to use ...** *

Markieren Sie nur ein Oval pro Zeile.

	... 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.
... abstract interpretation					
... assertion checking (e.g. for pre/post specification, contracts)					
... process calculi (e.g. CSP, CCS, pi, mu, hybrid)					
... model checking, SMV (of e.g. temporal or probabilistic properties)					
... constraint (SAT, SMT) solving (e.g. for static code analysis), optimisation techniques					
... generic (first-order, HOL) theorem proving (using e.g. term rewriting, functional programming)					
... simulation (i.e. computational engineering using e.g. differential calculus, numerical methods)					
... symbolic execution (e.g. scenario testing, model animation)					
... consistency checking (e.g. syntax or bug pattern checking)					

15. **F4o. I (would) intend to use other FMs, semi-FMs (i.e. without formal semantics and proof system), or highly systematic procedure (if any, please, provide some details):**

Future Purposes of Formal Methods Use

16. **F5. I (would) intend to use FMs for ...** *
Markieren Sie nur ein Oval pro Zeile.

	... 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.
... clarification (i.e. explicit description for analyzing a problem)					
... specification (e.g. contracts, documentation and communication of requirements and design)					
... inspection (i.e. error detection, e.g. non-conformance checking, model-based testing)					
... synthesis (e.g. transformation, compilation)					
... assurance (e.g. error removal, property verification, refinement or equivalence proofs, argumentation)					

17. **F5o. I (would) intend to use FMs for other purposes (if any):**

Potential Obstacles to the Intended Use of Formal Methods

18. **O1. For any potential use of FMs in my future activities and projects, I consider ...** *

Markieren Sie nur ein Oval pro Zeile.

	... not as an issue.	... as a moderate challenge.	... as a tough challenge.	I don't know.
... scalability (e.g. towards large or heterogeneous systems)				
... proper (automated) abstractions from irrelevant details				
... maintainability of verification results (e.g. stable proofs)				
... reusability of verification results (e.g. parametric proofs)				
... transfer of verification results from models to reality				
... automation or tool support (incl. notations, DSLs, IDEs)				
... skills and education (e.g. methods known and ready to use)				

19. **O1o. Which further obstacles (if any) would potentially hinder you to use FMs as intended?**

Thank you for your participation!

For SurveyCircle users (www.surveycircle.com): The Survey Code is: XXXX-XXXX-XXXX-XXXX

20. **Please, feel free to provide us any feedback on the questionnaire or on its topic:**

21. **If you are interested in our results you can provide us your email address below:**

References

Aichernig BK, Tom M (eds) (2003) Formal methods at the crossroads. From panacea to foundational support. Springer, Berlin. ISBN: 3-540-20527-6

ATOMICO (2019) The State of European Tech 2019. Section 6.4. URL: https://web.archive.org/web/20191220234928/http://2019.stateofeuropeantech.com/chapter/people/article/strong-talent-base/

Austin S, Graeme P (1993) Formal methods: A survey. Tech. rep. Teddington, Middlesex, UK: National Physical Laboratory

Barroca LM, McDermid JA (1992) Formal methods: Use and relevance for the development of safety-critical systems. Comp J 35(6):579–99. https://doi.org/10.1093/comjnl/35.6.579

Basili VR (1985) Quantitative evaluation of software bibmethodology. Tech. rep. TR-1519. University of Maryland. URL: https://drum.lib.umd.edu/bitstream/handle/1903/7520/Quantitative+Evaluation.pdf?sequence=1 (visited on 05/30/2019)

Bicarregui JC et al (2009) Industrial practice in formal methods: A review. In: Cavalcanti A, Dams DR (eds) FM 2009: Formal Methods. Springer, Berlin, pp 810–813. ISBN: 978-3-642-05089-3

Biernacki P, Waldorf D (1981) Snowball sampling: Problems and techniques of chain referral sampling. In: Sociological methods&research 10.2, pp 141–163. https://doi.org/10.1177/004912418101000205

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, IEEE Computer Society Press, pp 17–29. ISBN: 0-89791-216-0. https://doi.org/10.5555/41765.41768. URL: http://dl.acm.org/citation.cfm?id=41765.41768

Bloomfield RE, Froome PKD, Monahan BQ (1991) Formal methods in the production and assessment of safety critical software. In: Reliability Engineering & System Safety 32, vol 1-2, pp 51–66

Boulanger J-L (2012) Industrial use of formal methods: Formal verification. Wiley-ISTE. 298 pp. ISBN: 9781848213630

Bowen JP, Hinchey MG (1995a) Seven bibmore bibmyths of formal methods. In: IEEE Software 12.4, pp. 34–41. ISSN: 0740-7459. https://doi.org/10.1109/52.391826

Bowen JP, Hinchey MG (1995b) Ten commandments of formal methods. In: Computer 28.4, pp. 56–63. ISSN: 0018-9162. https://doi.org/10.1109/2.375178

Bowen JP, Hinchey MG (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. ACM, Lisbon, pp 8–16. ISBN: 1-59593-148-1. https://doi.org/10.1145/1081180.1081183

Campbell MJ, Gardner MJ (1988) Calculating confidence intervals for some non-parametric analyses. In: British Medical Journal, vol 296, p 1454

Charette RN (2018) Fiat chrysler is being sued over a software flaw. IEEE. https://web.archive.org/web/20180629231601/https://spectrum.ieee.org/riskfactor/computing/software/courtallows-lawsuit-to-proceed-against-fiat-chrysler-over-software-flaw

Chudnov A et al (2018) Continuous formal verification of amazon s2n. In: Computer aided verification. Springer International Publishing, pp 430–446. https://doi.org/10.1007/978-3-319-96142-2 26

Cofer DD et al (2012) Compositional verification of architectural models. In: NASA formal methods - 4th international symposium, NFM 2012. Proceedings, Norfolk, pp 126–140. https://doi.org/10.1007/978-3-642-28891-3n 13

Craigen D (1995) Formal methods technology transfer: Impediments and innovation (abstract). In: Lee I, Smolka SA (eds) CONCUR'95: Concurrency theory: 6th international conference Philadelphia, PA,

USA, August 21–24, 1995 Proceedings. Springer, Berlin, pp 328–332. ISBN: 978-3-540-44738-2. https://doi.org/10.1007/3-540-60218-6_24
Craigen D, Gerhart S, Ralston T (1993) An international survey of industrial applications of formal methods. In: Bowen JP, Nicholls JE (eds) Z User Workshop, London 1992: Proceedings of the 7th annual Z user meeting, London 14–15 December 1992. Springer, London, pp 1–5. ISBN: 978-1-4471-3556-2. https://doi.org/10.1007/978-1-4471-3556-2_1
Craigen D, Gerhart S, Ralston T (1995) Formal methods reality check: industrial usage. In: IEEE Transactions on Software Engineering 21.2, pp 90–98. ISSN: 0098-5589. https://doi.org/10.1109/32.345825
Davis FD (1989) Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. In: MIS Quarterly 13, vol 3, pp 319–40
Davis JA et al (2013) Study on the barriers to the industrial adoption of formal methods. In: Formal methods for industrial critical systems. Springer, Berlin, pp 63–77. https://doi.org/10.1007/ 978-3-642-41010-9_5
Decision Analyst (2018) Technology advisory board. Decision analyst, Inc. https://web.archive.org/web/20191214142906/https://www.decisionanalyst.com/online/acop/
Evans Data (2018) Global Developer Population and Demographic Study. Tech. rep. Volume 1. Evans Data Corporation. https://web.archive.org/web/20191015060004/https://evansdata.com/reports/viewRelease.php?reportID=9
Fagan ME (1976) Design and code inspections to reduce errors in program development. In: IBM Systems Journal 15.3, pp 182–211. https://doi.org/10.1147/sj.153.0182
Ferrari A et al (2019) Survey on formal methods and tools in railways technical report on the activities performed within ASTRail, Deliverable D4.1. https://doi.org/10.5281/zenodo.2535023
Fraser MD, Kumar K, Vaishnavi VK (1994) Strategies for incorporating formal specifications in software development. In: Communications of the ACM 37.10, pp 74–86. https://doi.org/10.1145/194313.19439
Galloway AJ, Cockram TJ, McDermid JA (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. https://doi.org/10.1016/S1474-6670(17)36335-8
Gerhart S, Yelowitz L (1976) Observations of fallibility in applications of modern programming methodologies. In: IEEE Trans. Software Eng. 2.3, pp 195–207. https://doi.org/10.1109/TSE. 1976.233815
Glass RL (2002) Facts and fallacies of software engineering. Pearson Education (US). ISBN978-0321117427
Gleirscher M, Marmsoler D (2018) Electronic supplementary material for formal methods: Oversold? underused? a survey. Zenodo. https://doi.org/10.5281/zenodo.1487596
Gleirscher M, Nyokabi A (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]
Gleirscher M, Foster S, Woodcock J (2019) New opportunities for integrated formal methods. ACM Comput Surv 52 (6):117:1–117:36. ISSN: 0360-0300. https://doi.org/10.1145/3357231 arXiv:1812.10103 [cs.SE]
Gnesi S, Margaria T (2013) Formal methods for industrial critical systems: A survey of applications. Wiley-IEEE Press. ISBN: 9781118459898
Google (2018) Google forms service. Google, Inc. http://forms.google.com
Graydon PJ (2015) Formal assurance arguments: A solution in search of a problem? In: 2015 45th annual IEEE/IFIP international conference on dependable systems and networks (DSN), pp 517–528. https://doi.org/10.1109/DSN.2015.28
Hall A (1990) Seven myths of formal methods. In: IEEE Software 7.5, pp 11–19. https://doi.org/10.1109/52.57887
Heisel M (1996) A pragmatic approach to formal specification. In: Object-oriented behavioral specifications. Springer. ISBN: 978-0-7923-9778-6. https://doi.org/10.1007/978-0-585-27524-6_4
Heitmeyer CL (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, vol 1486. Lyngby, DenmarkLyngby, Denmark, pp 18–26
Hinchey MG, Bowen JP (1996) To formalize or not to formalize? In: IEEE computer 29, vol 4, pp 18–19
Holloway CM (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. https://doi.org/10.1109/DASC.1997.635021
Holloway CM, Butler RW (1996) Impediments to industrial use of formal methods. In: Computer 29.4, pp 25–26. https://doi.org/10.1109/MC.1996.488298
Jackson M (1987) Power and limitations of formal methods for software fabrication. In: Journal of Information Technology 2.2, pp 72–76. https://doi.org/10.1177/026839628700200204
Jeffery R et al (2015) An empirical research agenda for understanding formal methods productivity. In: Information and software technology 60, pp 102–112. https://doi.org/10.1016/j.infsof.2014.11.005

Kaner C, Pels D (1998) Bad software, Wiley. ISBN: 978-0471318262
Kaner C, Pels D (2018) Bad software: Website. https://web.archive.org/web/20191210042547/http://badsoftware.com/
Kitchenham BA, Pfleeger SL (2008) Guide to advanced empirical software engineering. In: Springer. Chap. Personal Opinion Surveys, pp 63–92
Kitchenham B, Linkman S, Law D (1997) DESMET: A methodology for evaluating software engineering methods and tools. In: Computing & Control Engineering Journal 8.3, pp 120–126. https://doi.org/10.1049/cce:19970304
Klein G et al (2018) Formally verified software in the real world. In: Communications of the ACM 61.10, pp 68–77. https://doi.org/10.1145/3230627
Knight JC 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. https://doi.org/10.1007/978-0-387-35062-2_1
Lai R, Leung W (1995) Industrial and academic protocol testing: The gap and the means of convergence. In: Computer Networks and ISDN Systems 27.4, pp 537–547. https://doi.org/10.1016/0169-7552(93)E0110-Z
Leiner DJ (2014) SoSci Survey. Tech. rep. https://web.archive.org/web/20191202015133/https://www.soscisurvey.de/
Leino K, Rustan M (2017) Accessible Software Verification with Dafny. In: IEEE Software 34.6, pp 94–97. https://doi.org/10.1109/bibms.2017.4121212
Liebel G et al (2016) Model-based engineering in the embedded systems domain: an industrial survey on the state-of-practice. In: Software and systems modeling 17.1, pp 91–113. https://doi.org/10.1007/s10270-016-0523-3
Mathieson K (1991) Predicting user intentions: Comparing the technology acceptance model with the theory of planned behavior. In: Information Systems Research 2.3, pp. 173–191. https://doi.org/10.1287/isre.2.3.173
Miller SP, Whalen MW, Cofer DD (2010) Software bibmodel checking takes off. In: Communications of the ACM 53.2, pp 58–64. https://doi.org/10.1145/1646353.1646372
Miyoshi T, Azuma M (1993) An empirical study of evaluating software development environment quality. In: IEEE Transactions on Software Engineering 19.5, pp 425–435. https://doi.org/10.1109/32.232010
Mohagheghi P 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. https://doi.org/10.1007/s10664-012-9196-x
Murphy GC, Walker J, Banlassad ELA (1999) Evaluating emerging software development technologies: lessons learned from assessing aspect-oriented programming. In: IEEE Transactions on Software Engineering 25.4, pp 438–455. https://doi.org/10.1109/32.799936
Neuendorf KA (2016) The content analysis guidebook. 2nd. Sage. ISBN: 9781412979474
Neumann PG (2018) Risks to the public. In: ACM SIGSOFT software engineering notes 43.2, pp 8–11. https://doi.org/10.1145/3203094.3203102
O'Hearn PW (2018) Continuous reasoning. In: Proceedings of the 33rd annual ACM/IEEE symposium on logic in computer science - LICS'18. ACM Press. https://doi.org/10.1145/3209108.3209109
Oliveira JN (2004) A survey of formal methods courses in European higher education. In: Teaching formal methods. Springer, Berlin, pp 235–248. https://doi.org/10.1007/978-3-540-30472-2_16
Oliveira JN et al (2018) Formal methods body of knowledge (FMBoK). https://web.archive.org/web/20200109111534/https://formalmethods.wikia.org/wiki/FMBoK
Parnas Dl (2010) Really rethinking 'Formal Methods'. In: IEEE Computer 43.1, pp. 28–34. https://doi.org/10.1109/mc.2010.22
Petersen K 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, pp 26–27. https://doi.org/10.14236/ewic/ease2008.8
Pfleeger SL, Hatton L (1997) Investigating the influence of formal methods. In: Computer 30.2, pp 33–43. https://doi.org/10.1109/2.566148
Poston RM, Sexton MP (1992) Evaluating and selecting testing tools. In: IEEE Software 9.3, pp 33–42. https://doi.org/10.1109/52.136165
Riemenschneider CK, Hardgrave BC, Davis FD (2002) Explaining software developer acceptance of methodologies: A comparison of five theoretical models. In: IEEE transactions on software engineering 28.12, pp 1135–1145. https://doi.org/10.1109/tse.2002.1158287
Robbins NB, Heiberger RM (2011) Plotting Likert and other rating scales. In: Joint statistical meeting, pp 1058–66

Rushby J (1994) Critical system properties: Survey and taxonomy. In: Reliability engineering and system safety 43.2, pp 189–219. https://doi.org/10.1016/0951-8320(94)90065-5
SEI (2010) CMMI for Development. Tech. rep. CMU/SEI-2010-TR-033. CMU
Shull F, Singer J, Sjøberg DIK (eds) (2008) Guide to advanced empirical software engineering. Springer, London
Snook C, Harrison R (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. https://doi.org/10.1016/S0950-5849(00)00166-X
Sobel AEK, Clarkson MR (2002) Formal methods application: an empirical tale of software development. In: IEEE transactions on software engineering 28.3, pp 308–320. https://doi.org/10.1109/32.991322
The R Project (2018). R. The R Project. URL: https://web.archive.org/web/20200109063512/https://www.r-project.org/
Wikipedia contributors (2018) Software engineering demographics – Wikipedia, The Free Encyclopedia. URL: https://en.wikipedia.org/w/index.php?title=Software_engineering_demographics&oldid=823840899 (visited on 01/09/2020)
Wing JM (1990) A specifier's introduction to formal methods. In: Computer 23.9, pp 8–22. https://doi.org/10.1109/2.58215
Wohlin C et al (2012) Experimentation in software engineering. Springer. ISBN: 9783642290435
Woodcock J et al (2009) Formal methods: Practice and experience. ACM Comput Surv 41(4):9:1–19:36. ISSN: 0360-0300. https://doi.org/10.1145/1592434.1592436

Mario Gleirscher is a postdoctoral researcher in the Computer Science Department at the University of York, U.K. He received the M.Sc. degree in computer science with a minor in mathematics and the Ph.D. degree in computer science, both from the Technical University of Munich, Germany. He is also a qualified production engineer and has collected several years of practical experience as a consultant, method engineer, and software developer. He has been awarded a DFG research fellowship in 2017. His interests cover applied formal methods, particularly, algebraic methods, formal reasoning about risk in machines, and controller design for risk-aware autonomous machines.

Diego Marmsoler is a postdoctoral researcher at the Software and Systems group of Prof. Manfred Broy at the Technical University of Munich. He obtained a B.Sc. from the Free University of Bozen-Bolzano and an M.Sc. from the Technical University of Munich, Ludwig Maximilian University of Munich, and Augsburg University. He received a Ph.D. in Computer Science from the Technical University of Munich in 2019. His research focuses on the formal specification and verification of distributed, component-based systems. In particular, he works on the integration of various formal methods for the verification of such systems.