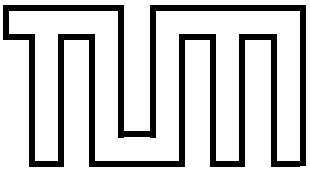


FAKULTÄT FÜR INFORMATIK

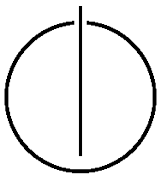
DER TECHNISCHEN UNIVERSITÄT MÜNCHEN

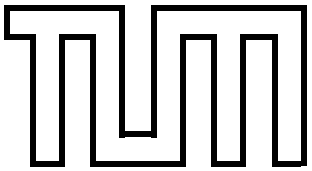
Bachelor Thesis

# Researching the Right Thing

Assessing the Practical Relevance of Requirements Engineering Research

Corinna Coupette





FAKULTÄT FÜR INFORMATIK

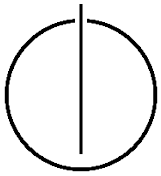
DER TECHNISCHEN UNIVERSITÄT MÜNCHEN

Bachelor Thesis

# Researching the Right Thing

Assessing the Practical Relevance of Requirements Engineering Research


| | |
|---|---|
| Bearbeiter: | Corinna Coupette |
| Betreuer: | PD Dr. habil. Daniel Mendéz Fernández |
| Abgabedatum: | 15.06.2018 |

Ich versichere, dass ich diese Bachelorarbeit selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

München, den 15.06.2018 Corinna Coupette

## Abstract

Requirements Engineering (RE) researchers seek to understand and ultimately improve RE practice. Little is known, however, about what drives the practical relevance of RE research. This thesis aims to shed light on that question, thereby providing a starting point for discussions on how the practical relevance of RE research might be increased. We start by mapping seven years of RE research as presented at the field's most prominent conferences, adding tags to research items to characterize their methods and contents. Against this background, we evaluate the results of an online survey conducted in early 2018. In this survey, practitioners were asked to judge the relevance of the mapped conference publications, explain the reasoning behind their ratings, and name research topics they would like to see investigated. Finally, we discuss the limitations of our results and their implications for the future of RE research.

# Contents

# List of Figures

# List of Tables

# 1 Introduction: Relating RE Research and Practice

Requirements Engineering (RE) researchers seek to understand and ultimately improve RE practice. Little is known, however, about what drives the practical relevance of RE research. This is the starting point of the so-called *RE-Pract survey* [5]. A replication of two recent baseline studies [14, 4], this online survey asks practitioners to judge the practical relevance of RE research presented at the field's most prominent conferences (435 publications from 2010 to 2016, inclusive) based on a hand-crafted, one-sentence summary of its methods and contents.[1] Respondents are further invited to explain the reasoning behind their ratings and name research topics they would like to see investigated [5].

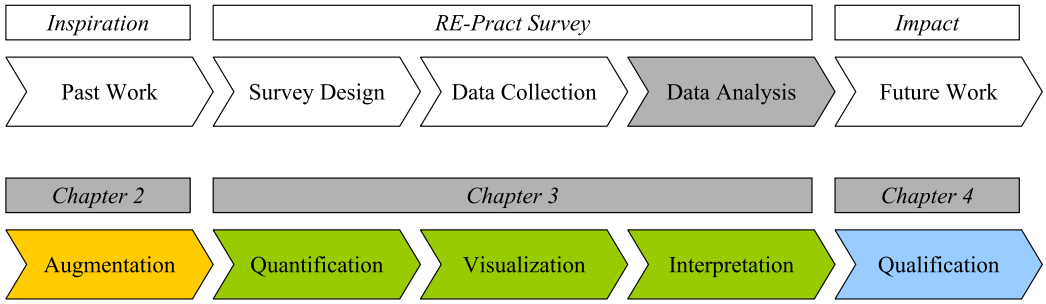



**Figure 1.1:** The RE-Pract Research Pipeline

Figure 1.1 (→ p. 1) shows the position of this thesis in the RE-Pract research pipeline. We present the results of the RE-Pract survey in chapter 3 (*Quantification*, *Visualization*, *Interpretation*), based on completed surveys by 154 respondents, each of whom rated a different, randomly drawn subset of the summarized RE publications. To strengthen the basis of our analysis and compensate for the small sample size, we group the RE publications presented to practitioners by their characteristics before we make quantitative statements about the *perceived* practical relevance of RE research. Therefore, we start by creating a map of this research in chapter 2 (*Augmentation*). Since respondents can only react to the one-sentence summaries they are presented with, we use a semi-automatic tagging scheme based on reg-

1 For an overview of related work, see [5] as well as section 2.2.

ular expressions to arrive at a RE research taxonomy, operating solely on the one-sentence paper summaries. While defensible in the present context, this procedure raises methodological doubts, which are discussed in chapter 4, along with further survey limitations (*Qualification*). Chapter 5 concludes with an outlook on the potential implications of our findings for the future of RE research.

In economic terminology, this thesis thus sets out to characterize the relationship between RE research supply and RE research demand as perceived through the eyes of RE practitioners. From this perspective, chapter 2 takes stock of the supply side, while chapter 3 sketches the demand side, and chapter 4 investigates the gap between the two, highlighting the challenges associated with its detection. Thereby, we hope to stimulate the debate on how the return on RE research investment might be increased.

# 2 Taxonomy: Characterizing RE Research Supply

In the RE-Pract survey, respondents are asked to rate a randomly selected subset of 435 RE research papers, based on one-sentence summaries crafted by the survey designers. To evaluate the importance of different RE research topics and methodologies addressed in the literature as perceived by practitioners, the RE research papers included in the survey need to be preprocessed so that they can be grouped by content criteria, allowing for an aggregation of ratings beyond the level of the individual paper. In this chapter, we therefore derive a map of RE research as presented in the 435 papers that are included in the survey. We start by describing the input we need to work with (→ 2.1). Considering both the limitations of this input and previous work on taxonomies for Software Engineering (SE) and RE research, we develop a procedure to assign tags to RE research papers (→ 2.2) which, when applied to our input papers, yields a map of RE research supply (→ 2.3).

## 2.1 Input: Seven Years of RE Research

The input to our mapping effort is described in [5]. The authors of the RE-Pract survey select 435 papers presented between 2010 and 2016 (inclusive) at the International Symposium on Empirical Software Engineering and Measurement (ESEM), the European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE), the International Symposium on the Foundations of Software Engineering (FSE), the International Conference on Software Engineering (ICSE), the International Requirements Engineering Conference (RE), and the International Working Conference on Requirements Engineering: Foundations for Software Quality (REFSQ),[1] deliberately limiting the publication pool to full papers related to RE. For each paper, the conference venue and year are documented, along with the authors' affiliation (academic, industry, or mixed) and

1 Franch et al. speak of 418 papers published between 2010 and 2015, *inter alia*, because the papers presented at the 2016 conferences are included only after the publication of [5].

whether the paper was submitted to a conference's industry track.[2] Table 2.1 (→ p. 4) summarizes the distribution of the selected papers over conferences and years, demonstrating the dominance of the RE Conference in the sample of papers to be evaluated by the RE-Pract survey participants. As Table 2.2 (→ p. 4) shows, most papers in the sample were presented outside of the industry track by researchers with academic affiliations.

| Venue<br>Year | ESEC/FSE | ESEM | FSE | ICSE | RE | REFSQ | Σ |
|---|---|---|---|---|---|---|---|
| 2010 | – | 3 | – | 5 | 38 | 15 | 61 |
| 2011 | 3 | 1 | – | 4 | 33 | 10 | 51 |
| 2012 | – | 4 | 1 | 3 | 35 | 14 | 57 |
| 2013 | 4 | 2 | – | 9 | 34 | 23 | 72 |
| 2014 | – | 7 | – | 10 | 44 | 18 | 79 |
| 2015 | 1 | 1 | – | 1 | 28 | 18 | 49 |
| 2016 | – | 4 | 7 | 7 | 33 | 15 | 66 |
| Σ | 8 | 22 | 8 | 39 | 245 | 113 | 435 |

**Table 2.1:** RE Research by Year and Conference Venue

| Industry track<br>Academic vs. industry | No | Yes | Σ |
|---|---|---|---|
| Academic | 296 | 20 | 316 |
| Industry | 7 | 28 | 35 |
| Mixed | 43 | 41 | 84 |
| Σ | 346 | 89 | 435 |

**Table 2.2:** RE Research by Affiliation and Conference Track

For each of the 435 papers, the survey authors formulate a one-sentence summary based on its abstract (where available) or its body [5]. Inspired by the approach pursued in [14], the summaries are crafted by one pair of researchers and validated by another pair of researchers. All sentences are sought to have similar structure, including information on the paper's methodology, its topic, and—where applicable—the people involved in the research. The summary authors also aim to control their vocabulary, partially reusing terminology from the existing literature on RE taxonomies, e.g., the research type facets proposed in [19]. This procedure results in sentences such as the following:

— "An experiment with practitioners for evaluating the adequacy and feasibility of an

2 Other metadata documented include a paper's title, its length in pages, and the names of its authors, all of which we do not use in this thesis.

existing software product line design method in order to prepare software product lines for likely future adaptation needs"

— "A method for eliciting business goals and linking them to quality requirements of the system in order to allow software architects to understand the business goals of the system"

— "A multi-case study on the use of goal-oriented requirements engineering techniques to improve traceability among enterprise architectures and business goals"

When practitioners are asked to rate the practical relevance of RE research papers in the RE-Pract survey, they have access to such one-sentence paper summaries only. Since the primary purpose of our taxonomy is to enable the aggregation of RE-Pract survey responses, this imposes a major constraint on our taxonomy design: It must deliver meaningful results when applied to the one-sentence representation of the papers created by the survey authors, rather than when applied to the papers themselves. If we want to keep the analyst-induced distortions of our results to a minimum, we may only use as an input to our mapping procedure the information presented to the RE-Pract survey participants because that is all our respondents can consider in their ratings. With this limitation in mind, we now describe our mapping process.

## 2.2 Procedure: Developing a RE Research Taxonomy

To construct our taxonomy, we need to map the 435 publications included in the RE-Pract survey. We can draw upon two principal sources: the literature and the data. A literature-driven approach starts by scrutinizing the taxonomies suggested in prior work (→ 2.2.1), whereas a data-driven approach starts by perusing the structure of the one-sentence summaries we seek to map (→ 2.2.2). Both approaches bear their own risks: While the literature-driven approach is prone to *underfitting* the data we seek to map, the data-driven approach is prone to *overfitting* it. Therefore, we ultimately adopt a hybrid approach, striking the balance between the literature and the data (→ 2.2.3).

### 2.2.1 Literature-Driven Approach: Leveraging Existing Taxonomies

There are numerous taxonomies (or similar conceptual constructs) that might be helpful to organize our 435 RE research papers. Some suggestions are contained in documents pro-

vided by international organizations, associations, or standard-setting bodies, others are offered in RE or SE research publications. In the following, we sketch the insights to be gleaned from these materials and highlight their limitations.

#### International Organizations, Associations, and Standard-Setting Bodies

When it comes to taxonomies for research paper classification more generally, the Institute of Electrical and Electronics Engineers (IEEE) and the Association for Computing Machinery (ACM) keyword taxonomies come to mind—but both are far too broad for the use case at hand. The IEEE taxonomy features only "Requirements engineering", with a subheading of "Technical requirements", and "Requirements management" as categories, both under the rubric "Systems engineering and theory". The ACM taxonomy does not even use the term "Requirements engineering", mentioning only "Requirements analysis" (under "Software and its engineering" → "Software creation and management" → "Requirements analysis") and "Security requirements" (under "Security and privacy" → "Formal methods and theory of security") without further subheadings.

Moving from paper classification schemes to international standards affecting the RE domain, the *ISO/IEC/IEEE 29148* standard on *Systems and software engineering — Life cycle processes — Requirements engineering* [11], though aimed at practitioners, offers more guidance for the taxonomy we seek to build. While its table of contents offers little insight into the structure of the RE domain, the standard highlights some key concepts and distinctions that seem suitable as building blocks of an RE research taxonomy. For example, RE is said to be "concerned with discovering, eliciting, developing, analyzing, determining verification methods, validating, communicating, documenting, and managing requirements" (5.2.1), which suggests requirements discovery, elicitation, development, analysis, verification, validation, communication, documentation, and management as activities that form part of the RE process. Further, the adjectives *necessary*, *implementation free*, *unambiguous*, *consistent*, *complete*, *singular*, *feasible*, *traceable*, and *verifiable* are listed as desirable characteristics of individual requirements (5.2.5), while *complete*, *consistent*, *affordable*, and *bounded* are mentioned as desiderata for sets of requirements (5.2.6). "Requirements type" is listed as a requirements attribute (5.2.8.1), with "functional", "performance", "interface", "design constraints", "process requirements", and "non-functional" given as examples (5.2.8.2). Non-functional requirements are further broken down into "quality requirements" and "human factors re-

quirements", with "transportability, survivability, flexibility, portability, reusability, reliability, maintainability, and security" highlighted as examples of quality requirements (5.2.8.2). However, while the standard provides details concerning individual tasks and documents in a typical RE process, it contains no systematization of the ways to achieve or produce them; it is focused on the "what," rather than the "how," of RE.[3] As much of RE research is concerned with the latter, this limits the standard's applicability to our mapping effort.

Finally, the materials published by the International Requirements Engineering Board (*IREB e.V.*), which are also tailored to practitioners, constitute another potential resource for our taxonomy construction efforts. The CPRE Foundation Level Syllabus [6] offers little beyond the ISO/IEC/IEEE standard, and the handbooks as well as the syllabi for the advanced certifications are too specialized for our current task. The glossary [7], however, constitutes a convenient checklist against which we can assess the completeness of our mapping scheme.

### RE and SE Research

The materials available from international organizations or standard-setting bodies focus on structuring RE as a *practice domain*. As we seek to design a taxonomy to structure RE as a *research domain*, we thus turn to the literature for additional guidance. One of the earlier RE-specific research classification schemes is by Zave, over twenty years old [21], and summarized in Figure 2.1 (→ p. 8).

Without going into the details, at least two things appear remarkable about this proposal. First, the first dimension of the taxonomy explicitly revolves around problems, rather than around tasks, "because [problems] are more stable than tasks" [21, p. 316]. This focus is emphasized by recent research on problems in RE, including the *Naming the Pain in Requirements Engineering* (NaPiRE) initiative [15, 16], and it is in contrast to the materials provided by international organizations and standard-setting bodies, which—given their peculiar function as standardization or training tools—focus on tasks and documents. Second, the second dimension of the taxonomy addresses the type of the research to be classified, an element naturally missing in all sources focused on RE practice.

3 As the "what" of RE can be expected to be much more stable than the "how", this is not a *bug* of the standard but rather a *feature*.

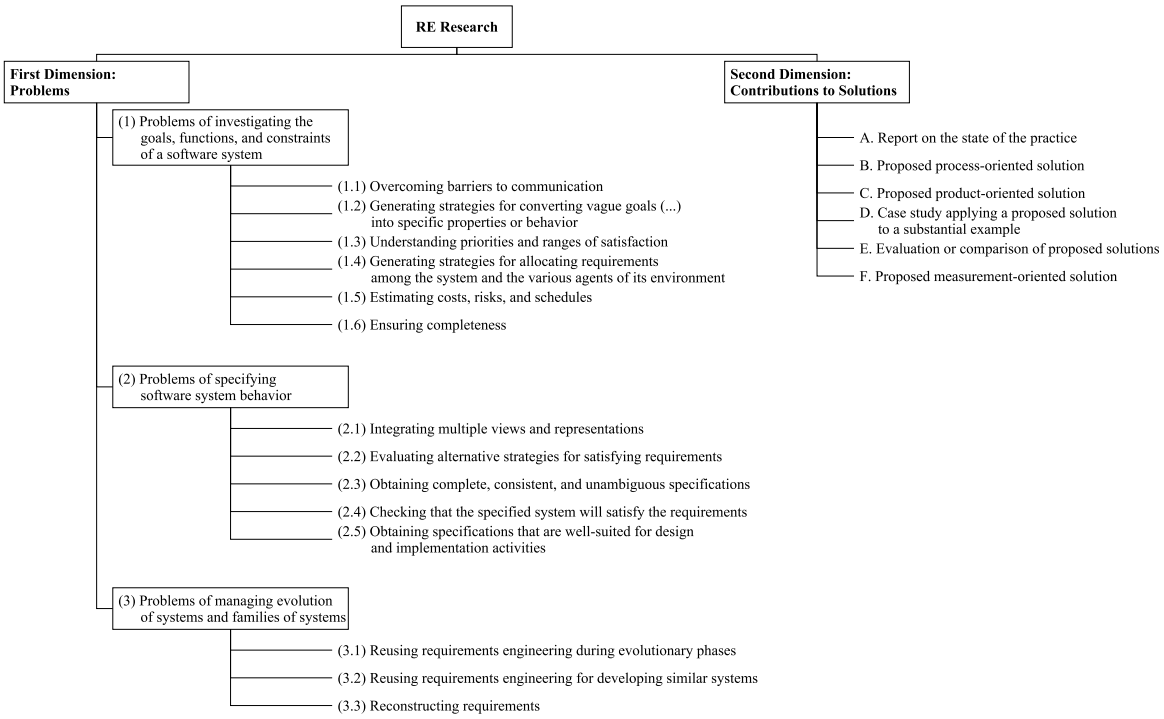



**Figure 2.1:** RE Research Classification as Proposed by Zave in [21]

The second dimension of Zave's taxonomy is further developed by Wieringa et al. in [19]. Wieringa et al. distinguish six paper classes, (1) evaluation research, (2) proposal of solution, (3) validation research, (4) philosophical papers, (5) opinion papers, and (6) personal experience papers, in order to then propose different evaluation criteria for each class of papers to be used in the context of conference paper acceptance [19]. Their scheme has been taken up in subsequent papers (e.g., [5, 18]), despite the potential confusion caused by the distinguishing parts of the labels for two central categories, evaluation research and validation research. These words, *evaluation* and *validation*, are not intuitively connected to their categories' meaning as ascribed the authors (evaluation research: "This is the investigation of a problem in RE practice or an implementation of an RE technique in practice."; validation research: "This paper investigates the properties of a solution proposal that has not yet been implemented in RE practice." [19, p. 105]), and they are used with different meanings in RE practice (e.g., validation of requirements, evaluation of projects).[4] Yet, the authors of the RE-Pract survey attempt to encode Wieringa et al.'s classification scheme (also known as

4 Thus, in the RE community, *evaluation* and *validation* have come to be used as descriptors of both *empirical research strategies* and *steps in the RE process*.

"research type facets") in the one-sentence summaries presented to the survey participants [5], and we address the challenges associated with this choice in section 2.2.3.

Last but not least, most inspiration for the construction of our RE research taxonomy can be drawn from a number of "Body of Knowledge" (BoK) projects. The oldest of these projects is the Software Engineering Body of Knowledge (SWEBOK), which was first published in 2004 and is now in its third edition [13].[5] The SWEBOK is divided into 15 chapters, each devoted to a different Knowledge Area (KA). Its first chapter is dedicated to software requirements, and it features the diagram reproduced in Figure 2.2 (→ p. 9).

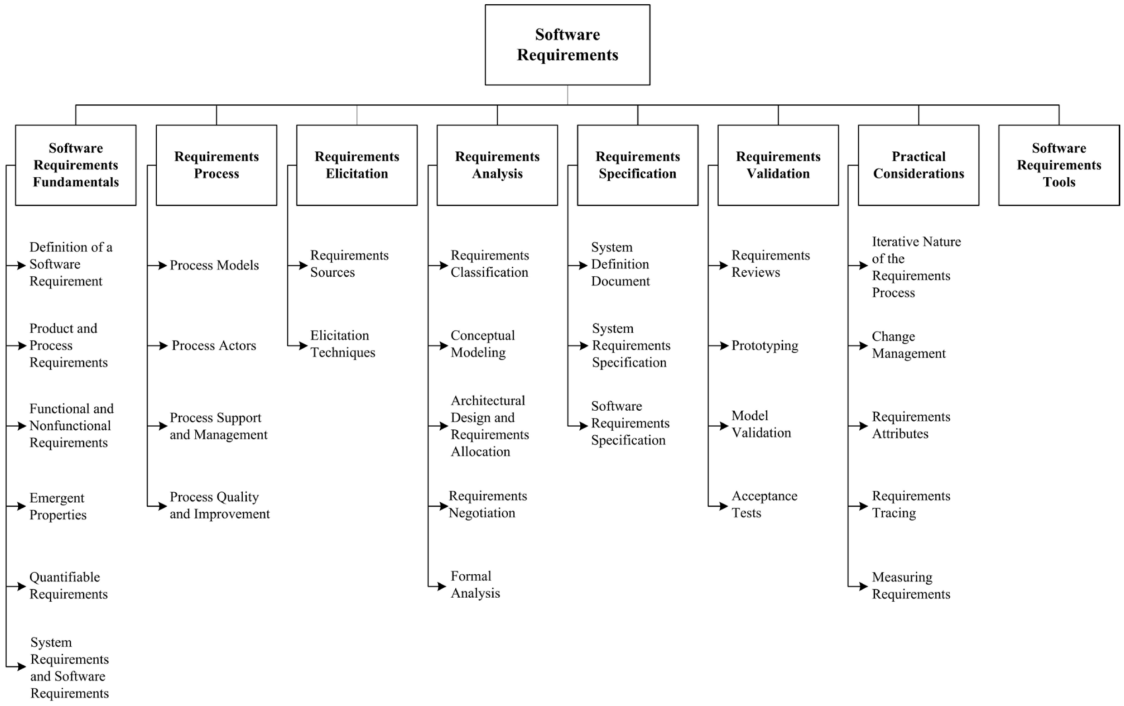



**Figure 2.2:** Breakdown of Topics from the Software Requirements KA (reproduced from [13])

This structure is almost identical to the structure proposed for the Requirements Engineering Body of Knowledge (REBoK)[6] developed in Japan [1, 2, 3], which treats (1) Requirements Engineering Fundamentals, (2) Requirements Engineering Process, (3) Requirements Eliciation, (4) Requirements Analysis, (5) Requirements Specification, (6) Requirements Verification, Validation and Evaluation, (7) Requirements Planning and Management, and (8) Prac-

5 The SWEBOK itself relies heavily on the *ISO/IEC 12207* standard, whose 2008 version [9] *ISO/IEC/IEEE 29148* elaborates on. When the first SWEBOK was published, only the 1995 version of *ISO/IEC 12207* was available [8]. The standard was last updated in 2017 and is now also endorsed by the IEEE [10].

6 The small "o" in the abbreviation (compare SWEBOK) appears to be intentional.

tical Considerations as separate KAs.[7] The emphasis on the steps of the RE process is striking; distinctions addressing the contents of requirements and the challenges associated with RE are relegated to the lower levels of the taxonomy. Also, as expectable for a BoK structure, there is no way to categorize research contributions by research type. All in all, there exists no taxonomy in the literature that we could readily employ to classify our 435 research papers.

### 2.2.2 Data-Driven Approach: Exploiting Sentence Summary Structure

Having considered the taxonomies available in the literature, we now turn to the structure available in our data. As explained in section 2.1, the one-sentence summaries presented to the participants of the RE-Pract survey are hand-crafted by its authors. The authors, all active in the RE research community, leverage their background knowledge in RE and their internal discussions to give all summaries a similar structure. We can recover (parts of) this structure from the summaries themselves for use in the construction of our taxonomy. More precisely, we can derive terms, which we call *tags*, characterizing our papers with a view to both their research methods and their research contents.

#### Research Methods

One distinguishing facet of a paper's research method is the type of research presented. The RE-Pract survey authors seek to follow Wieringa et al. by phrasing their summaries in terms of the research type facets presented by these authors (→ 2.2.1), frequently using "in order to" to specify apparent research goals. However, the attention of an external observer (e.g., the survey respondent whose perspective we are trying to mimic) is first drawn to a much more mundane kind of structure: regularities in the phrasing at the start of the summaries. The 435 one-sentence summaries start with only around 50 different word triples (3-grams), some of which are again almost identical; frequent examples include "A case study," "An experiment with," or "A method for". Upon closer inspection, it becomes evident that the vocabulary used at the start of the summaries is indicative of the type of research, which may be further classified using conceptual distinctions. Additionally, where the research involves interaction with a specific population, this is included in the summaries using a simple

7 The identically named REBoK initiative by Penzenstadler et al. [17] appears to have been discontinued.

"with"-statement, e.g., "An experiment with students" or "An interview-based study with practitioners". Thus, from the wording of the sentences, we can easily extract information regarding the *how* and the *with whom* of the summarized research papers.

### Research Contents

Information regarding the *what* of the papers in question, i.e., their research content, is much harder to discern from the one-sentence summaries than information regarding the research method (*how* and *with whom*). First, there are many more themes of RE research than there are research methods, which means that the content part of our taxonomy will naturally be more granular than its method part. Second, the one-sentence summaries describe the focus of the research in question from the perspective of the respective paper. The perspectives of the summarized papers, however, reflect different perspectives on RE in general and are therefore very diverse. Compare, for example, the following three summaries:

— "An experiment with students for comparing two requirements elicitation approaches when instantiating a Software Product Line (SPL) in order to understand which approach is more suitable for eliciting requirements when using SPLs."[8]

— "A method for automatically recovering software traceability links between various software artifacts based on topic modelling (requirements, design, code, bug reports, test cases)"[9]

— "An analysis on the integration of non-functional requirements into model-driven development processes in order to include this type of requirements into such processes"

Here, the dominant content cues to the uninitiated reader are "elicitation," "traceability," and "non-functional requirements"—i.e., the first cue refers to a step in the RE process, the second cue refers to a desirable quality characteristic of a requirement or its specification, and the third cue refers to a specific subset of requirements. Since survey respondents might base their research ratings primarily on an intuitive reaction to their "first impression" of a one-sentence summary, we can assume that the content part of their rating is

8 The punctuation inconsistencies stem from the original summaries, which are reproduced here without modifications.

9 The summaries are (partly) written in British English, hence the occasional divergence in spelling between the main text and the quotations.

driven by the dominant content cues provided in the paper summaries.[10] As these content cues evoke entirely different reference frames (e.g., the reference frame for "elicitation" is the RE process, whereas the reference frame for "non-functional requirements" is a typology of requirements), we need to include these reference frames as facets in our taxonomy if we seek to disentangle the motives behind the paper ratings after the taxonomy-enabled aggregation of ratings. Since the reference frames themselves are defined by the concepts and categories RE practitioners are familiar with, which are reflected in the literature, this mandates a hybrid approach to the construction of our RE research taxonomy.

### 2.2.3 Hybrid Approach: Data-Driven Synthesis of Existing Taxonomies

Thus far, we have used the term "taxonomy" informally, employing it for all structured sets of RE-related terms. As Usman et al. highlight in their systematic mapping study of taxonomies in SE, a taxonomy is a scheme of classification "allowing for the description of terms and their relationships in the context of a knowledge area" [18]. The authors bemoan a lack of detail in the documentation of the procedures used to develop and deploy existing taxonomies, and they propose a revised taxonomy development method comprising 13 taxonomy development activities. We slightly modify and reorganize these activities to apply them in our case, arriving at the 14-point taxonomy fact sheet presented in Table 2.3 (→ p. 13). In the following, we provide details on the rationale behind each of the entries.

The first five points of the fact sheet, P1–P5, relate to *taxonomy planning*. P1 states the SWEBOK Knowledge Area (here: Software Requirements), P2 defines the focus of the taxonomy within the knowledge area (RE research papers), and P3 clarifies the taxonomy goal (aggregation for the purposes of survey evaluation). P4 and P5 lay out the overall structure of the taxonomy and the procedure used to apply it. Given the diversity in the research papers we seek to map, we settle for a facet-based tagging scheme, assigning tags to papers along several dimensions in order to characterize them rather than bucketing them into clear-cut categories. The tagging is done using regular expressions (regex) and then refined manually (removing tags where the regex are too general and adding tags where they are too narrow).

There are six considerations concerning *taxonomy construction*. CT1–CT3 specify the sources

10 On the challenge resulting from the mixture of content and method cues in the one-sentence summaries, see the discussion in chapter 4.

of the terms to be used in the taxonomy, the method used to extract the terms from the source, and the mechanism used to control the vocabulary. In our case, we use existing taxonomies as well as the one-sentence paper summaries as a basis for our custom taxonomy, which we operationalize by designing suitable regex;[11] the result is discussed with RE experts.[12] CS1–CS3 summarize the structure of the taxonomy, namely, its top-level dimensions, its lower-level terms and the relationships between the terms. Here, the top-level dimensions are the question phrases already highlighted above (→ 2.2.2), which are orthogonal to one another and each contain multiple, partially non-orthogonal facets (→ Tables 2.4–2.6, pp. 14–16).

| **RE Research Taxonomy** | | |
|---|---|---|
| *Planning* | | |
| P1 | Knowledge Area | Software Requirements |
| P2 | Subject Matter | RE Research: Full Conference Papers |
| P3 | Objective | Grouping RE Research Papers to Aggregate Practitioners' Research Paper Ratings |
| P4 | Taxonomy Structure | Facet-Based Tagging Scheme |
| P5 | Application Procedure | Regular Expressions Plus Manual Finishing |
| *Construction* | | |
| | *Terms* | |
| CT1 | Sources | Literature and One-Sentence Paper Summaries |
| CT2 | Extraction | Manual Synthesis of Concepts from the Sources |
| CT3 | Control | Discussion with RE Experts |
| | *Structure* | |
| CS1 | Top-Level Dimensions | *how*, *with whom*, *what* |
| CS2 | Lower-Level Terms | See Tables 2.4–2.6 (→ pp. 14–16) |
| CS3 | Relationships between Terms | Top Level: Orthogonal;<br>Lower Levels: Partially Non-Orthogonal (Tags) |
| *Application* | | |
| A1 | Guidelines | Contained in the Regular Expressions |
| A2 | Validation | Application to 435 RE Research Papers Based on Their One-Sentence Summaries |
| A3 | Maintenance | Updating or Restructuring Regular Expressions |

**Table 2.3:** RE Research Taxonomy: Fact Sheet

11 The regular expressions we employ are available in the online companion to this thesis.

12 In the present context, discussion does not mean that full agreement must be reached between the experts and the author.

Finally, A1–A3 address *taxonomy application*. A1 seeks to facilitate the application of the taxonomy—in our case, all necessary guidelines are contained in our regex. A2 tries to ensure that the taxonomy "works," which we ascertain by its application to our 435 research papers, well aware of the fact that the taxonomy is *designed* to fit our data.[13] A3 aims to move beyond the individual case by providing information on how the taxonomy might be maintained; for us, this largely comes down to a regex update (though such an update appears unlikely given the peculiarities of the taxonomy design that arise from the structure of our data).

Putting all the above into practice, we arrive at the taxonomy shown in Tables 2.4–2.6 (→ pp. 14–16). In all tables, the terms are sorted alphabetically on a per-level basis (this facilitates table lookup); terms that do not appear in the taxonomy application (e.g., due to flawed regex) are set in *italics*. The rationale behind the chosen terms and structures is explained in the right column;[14] here, "distinctions taken from NaPiRE" is a shorthand for "the distinctions used in the 2017 edition of the NaPiRE survey appeared to constitute an adequate synthesis of existing distinctions in the literature."

| **Level 1** | **Level 2** | **Level 3** | **Explanation** |
|---|---|---|---|
| how | engineering | methodology<br>reference<br>technology | how things *should be* `done` |
| | perspective | experience<br>opinion<br>philosophy<br>review | how things *can be* `seen` |
| | science | interrogation<br>intervention<br>observation | how things *are* `done` |
| withwhom | laypeople | others<br>students | subjects *without* RE expertise |
| | professionals | academics<br>practitioners | subjects *with* RE expertise |

**Table 2.4:** RE Research Taxonomy: Research Methods

13 For a discussion of the implications of this (unavoidable) choice, see chapter 4.

14 To ease the automatic processing of the taxonomy, terms consisting of multiple words are written without the interspersed whitespace characters.

| Level 1 | Level 2 | Level 3 | Level 4 | Explanation |
|---|---|---|---|---|
| what | challenge | content | all<br>completeness<br>consistency<br>feasibility<br>traceability<br>unambiguousness<br>understandability | dominant cues as derived from the data |
| | | context | regulation<br>uncertainty | |
| | | failure | | |
| | | people | collaboration<br>communication<br>skills<br>subjectivity | |
| | | problem | | |
| | | process | automation<br>deciding<br>formalization<br>improving<br>prioritization<br>standardization<br>visualization | |
| | documentation | artifacts<br>businessmodels<br>diagrams<br>featuremodels<br>goalmodels<br>naturallanguage<br>prototypes<br>statemachines<br>usecases<br>userstories | | distinctions taken from NaPiRE |

**Table 2.5:** RE Research Taxonomy: Research Contents (Part 1)

| Level 1 | Level 2 | Level 3 | Level 4 | Explanation |
|---|---|---|---|---|
| what | domain | organization | agile<br>distributed<br>lean<br>outsourced | distinctions derived from the data |
| | | sector | automotive<br>energy<br>health<br>it<br>media<br>mobile<br>nanotechnology<br>public<br>subsea<br>supplier | |
| | | systemclass | adaptive<br>bi<br>complex<br>embedded<br>*opensource*<br>safetycritical | |
| | general | framework<br>research<br>*terminology* | | meta-content |
| | information | architecture<br>functional<br>goals | | distinctions taken from NaPiRE |
| | | quality | all<br>*compatibility*<br>*maintainability*<br>performance<br>*portability*<br>reliability<br>safety<br>security<br>sustainability<br>usability | distinctions taken from NaPiRE<br>*added* |
| | | rules<br>scenarios<br>systembehavior | | distinctions taken from NaPiRE |
| | phase | analysis<br>elicitation<br>evaluation<br>management<br>specification | | RE life cycle phases<br>*verification/validation*<br>*adaptation, reuse, ...* |
| | region | *continent*<br>country | ...<br>... | geography |

**Table 2.6:** RE Research Taxonomy: Research Contents (Part 2)

## 2.3 Output: A Map of RE Research Supply

Applying the taxonomy shown in Tables 2.4–2.6 (→ pp. 14–16) to our 435 RE research papers, we are now able to map RE research supply.[15] Table 2.7–2.9 (→ pp. 17–19) show the number of times we assign the individual tags to papers in our sample, using the same structure and ordering as in Tables 2.4–2.6 (→ pp. 14–16).[16] We store the data presented in these tables in long-form `csv` files (one row per unique pair of paper and tag) for combination with our survey data, which enables us to analyze the results of the RE-Pract survey.

| Level 1 | Level 2 | Level 3 | Tag Count |
|---|---|---|---|
| how | engineering | methodology | 177 |
| | | reference | 5 |
| | | technology | 33 |
| | perspective | experience | 38 |
| | | opinion | 11 |
| | | philosophy | 1 |
| | | review | 14 |
| | science | interrogation | 45 |
| | | intervention | 39 |
| | | observation | 77 |
| withwhom | laypeople | others | 1 |
| | | students | 28 |
| | professionals | academics | 2 |
| | | practitioners | 30 |

**Table 2.7:** RE Research Map: Research Methods

Before we start this analysis, however, several comments are in order. In Table 2.7 (→ p. 17), the dominance of the `methodology` tag sorted under the `engineering` tag is due to the regex used during tag assignment, which were suggested by one of the authors of the RE-Pract survey. Whether the resulting third-level structuring in the engineering cat-

15 The statistics presented in the following are the result of a regular expressions run and a subsequent manual validation by the thesis author, the results of which were discussed with RE experts.

16 According to the authors of the RE-Pract survey, the regional information shown at the end of Table 2.9 (→ p. 19) should not have been included in the paper summaries. We retain the corresponding tags since the associated information *did* make its way into the survey, likely affecting participants' ratings of the papers in question.

egory is convincing, we leave for the critical reader to judge. More conveniently, there are roughly as many papers with the tag `withwhom:laypeople:students` as with the tag `withwhom:professionals:practitioners`, and also roughly equal numbers of papers with the `how:science:interrogation` and the `how:science:intervention` tags. We can exploit these regularities in our subsequent analyses.

| Level 1 | Level 2 | Level 3 | Level 4 | Tag Count |
|---|---|---|---|---|
| what | challenge | content | all | 9 |
| | | | completeness | 18 |
| | | | consistency | 13 |
| | | | feasibility | 2 |
| | | | traceability | 44 |
| | | | unambiguousness | 18 |
| | | | understandability | 15 |
| | | context | regulation | 27 |
| | | | uncertainty | 31 |
| | | failure | | 8 |
| | | people | collaboration | 13 |
| | | | communication | 24 |
| | | | skills | 26 |
| | | | subjectivity | 13 |
| | | problem | | 9 |
| | | process | automation | 48 |
| | | | deciding | 18 |
| | | | formalization | 8 |
| | | | improving | 35 |
| | | | prioritization | 18 |
| | | | standardization | 18 |
| | | | visualization | 11 |
| | documentation | artifacts | | 14 |
| | | businessmodels | | 1 |
| | | diagrams | | 7 |
| | | featuremodels | | 10 |
| | | goalmodels | | 7 |
| | | naturallanguage | | 48 |
| | | prototypes | | 1 |
| | | statemachines | | 1 |
| | | usecases | | 10 |
| | | userstories | | 4 |

**Table 2.8:** RE Research Map: Research Contents (Part 1)

| Level 1 | Level 2 | Level 3 | Level 4 | Tag Count |
|---|---|---|---|---|
| what | domain | organization | agile | 11 |
| | | | distributed | 3 |
| | | | lean | 2 |
| | | | outsourced | 1 |
| | | sector | automotive | 5 |
| | | | energy | 2 |
| | | | health | 5 |
| | | | it | 6 |
| | | | media | 6 |
| | | | mobile | 4 |
| | | | nanotechnology | 1 |
| | | | public | 4 |
| | | | subsea | 1 |
| | | | supplier | 3 |
| | | systemclass | adaptive | 6 |
| | | | bi | 1 |
| | | | complex | 1 |
| | | | embedded | 2 |
| | | | safetycritical | 7 |
| | general | framework | | 2 |
| | | research | | 8 |
| | information | architecture | | 7 |
| | | functional | | 14 |
| | | goals | | 18 |
| | | quality | all | 18 |
| | | | performance | 4 |
| | | | reliability | 2 |
| | | | safety | 6 |
| | | | security | 26 |
| | | | sustainability | 3 |
| | | | usability | 1 |
| | | rules | | 3 |
| | | scenarios | | 5 |
| | | systembehavior | | 3 |
| | phase | analysis | | 21 |
| | | elicitation | | 49 |
| | | evaluation | | 28 |
| | | management | | 76 |
| | | specification | | 44 |
| | region | country | china | 1 |
| | | | finland | 1 |

**Table 2.9:** RE Research Map: Research Contents (Part 2)

In Table 2.8 (→ p. 18), the relatively high counts for `challenge:content:traceability`, `challenge:process:automation`, and `documentation:naturallanguage` deserve special mention, as do the relatively large numbers for `information:quality:security` and `phase:management` in Table 2.9 (→ p. 19).[17] While the first four counts hint at particular foci of the RE research community (tackling traceability and process automation as challenges; handling requirements specified in natural language; dealing with security as a quality requirement), the last count might well be an artifact of our taxonomy design, which designates `management` as the process category comprising requirements evolution and adaptation. Finally, the internal validation of the data underlying Table 2.8 and Table 2.9 (→ pp. 18–19) is still ongoing, and the taxonomy presented here makes no attempt to be consistent with all terminological tastes in the RE research community. With our map of RE research supply in place, we now turn to analyzing RE research demand as expressed in the responses to the RE-Pract survey.

17 In this context, "relatively" means "when compared to the other items grouped under the same higher-level tag."

# 3 Survey: Exploring RE Research Demand

Most research in RE seeks to understand and ultimately improve RE practice, and the high number of methodology proposals in our sample of RE research papers (→ Table 2.7, p. 17) is a testament to this observation. But although the supply side seems willing to produce RE research that serves practitioners' demands, little is known about what research the RE industry actually needs. The RE-Pract survey aims to provide a window into RE research demand by asking practitioners to rate some of the 435 RE research papers mapped in the previous chapter, to give reasons for their ratings, and to highlight topics they would like to see investigated. Thus, we start our inquiry into the demand side of RE research by describing the survey in more detail (→ 3.1). We continue by analyzing how our respondents rate the RE research papers they are shown, aggregating our data by respondent characteristics and paper characteristics (→ 3.2). This leads us to a coarse-grained picture of RE research demand, which we refine by inspecting both the reasons practitioners give for their ratings and the research wishes they voice (→ 3.3). We generally use graphics, rather than tables, to present quantitative survey results, and provide tabular representations as well as details on the handling of free-text answers in the online companion to this thesis.[1]

## 3.1 Overview: The RE-Pract Survey

In this section, we give an overview of the RE-Pract survey. We first summarize the survey setup (→ 3.1.1) and then characterize our sample of respondents, the 154 RE practitioners who complete the survey, by its demographics (→ 3.1.2). Thereby, we establish the context in which we undertake all subsequent analyses.

1 The online companion will be made available at `https://github.com/napire/repract2018` and will be archived with *Zenodo*.

### 3.1.1 Setup

The RE-Pract survey seeks "to investigate practitioners' perceptions of the practical relevance of today's academic research in RE" [5]. In particular, the following research questions have been formulated by the survey authors [5]:

**RQ1** What is the relevance of RE research to practitioners in the industry?

**RQ2** What are the most highly rated research ideas?

**RQ3** What research problems do practitioners think are most important to be focused on by the RE research community?

**RQ4** Do papers with explicit ties to industry have higher practical relevance than other papers?

**RQ5** Do practitioners' perceptions and views differ in dependence on their roles?

These research questions allow for different interpretations;[2] their formulations blend phrases addressing reality (e.g., "is the relevance", "have higher practical relevance") with phrases addressing perceptions of reality (e.g., "do practitioners think", "practitioners' perceptions and views differ"). Additionally, as later sections demonstrate, not all of these questions can be fully investigated given the limitations of our data. Therefore, we derive the structure of our analysis in sections 3.2 and 3.3 from the data and address the research questions mentioned above where they fit in our context, referencing them as **RQ1** to **RQ5**.

To collect the data needed for answering their research questions, the authors design the core of the RE-Pract survey to consist of two parts. In the first part, each respondent is tasked with rating 15 one-sentence research summaries on a four-item Likert scale. The items on that scale are labeled `Essential`, `Worthwhile`, `Unimportant`, and `Unwise`. This choice might appear problematic, and we discuss the difficulties resulting from it in chapter 4, along with further methodological challenges. With the help of a standard survey administration tool, the summaries shown to individual practitioners are drawn randomly from the pool of 435 RE research paper summaries described in chapter 2. In the second part, participants are asked to explain the reasoning behind their assessments of two summaries which receive one of their best or worst ratings, respectively. Here, respondents are also given the chance to state which RE problems or topics they would like RE researchers to

2 For example, readers might have differing understandings of the concepts "relevance," "research idea," or "research problem."

address in the future. Thus, while the first part of the survey's core includes only closed-ended questions, its second part features open-ended questions exclusively.

To control response rates and sample composition, the RE-Pract survey is designed as an invitation-only online survey and distributed electronically to RE practitioners. According to information provided by the RE-Pract team, most invitees are selected from lists of personal contacts compiled by the survey authors. Further participants are recruited via the mailing list of *IREB e.V.*, a German association providing training and certifications in RE. Responses are recorded anonymously to lower barriers to participation. However, respondents are asked to provide information on their professional background at the end of the questionnaire.

### 3.1.2 Sample

The survey was open for participation from January 7 to March 27, 2018. In this period, 961 RE practitioners were invited to participate, with 410 invitees starting and 154 invitees finishing the survey. Thus, the *response rate* of the RE-Pract survey is $42.66\ \%$ and its *completion rate* is $16.03\ \%$. For completed questionnaires, the mean processing time is $22\mathrm{m}35\mathrm{s}$ while the median processing time is $13\mathrm{m}50\mathrm{s}$.[3]

The demographic information respondents provide at the end of the survey allows us to characterize the sample of participants who completed the questionnaire.[4] As evident from Figure 3.1 (→ p. 24), our data is dominated by answers from practitioners located in Germany, which is mentioned more frequently than the three runners-up—the United States, Brazil, and China—taken together ($39 > 14 + 10 + 10 = 34$). Notably, the four countries with the highest number of contributors lie on four different continents. Although most answers

3 These numbers suggest that the distribution of processing times is right-skewed. If the processing times are representative of the time our respondents actively spent working on the survey, this might indicate that some respondents put much more thought into their answers than others. However, a right-skewed distribution of recorded processing times could also be an artifact of the survey administration tool's time-tracking internals (e.g., if the tool does not distinguish between time spans with and without on-screen activity).

4 In the following, we briefly address all statistics generated by the demographics questions except for statistics resulting from answers to the following question: *To which extent do you actively engage in requirements engineering?* The answer options provided to respondents resulted in several responses which are internally inconsistent.

stem from countries in Europe and North America, all continents (with the exception of Antarctica) are represented in the sample. Thus, there is considerable geographic diversity, which suggests that different feedback cultures might have impacted our results.[5]

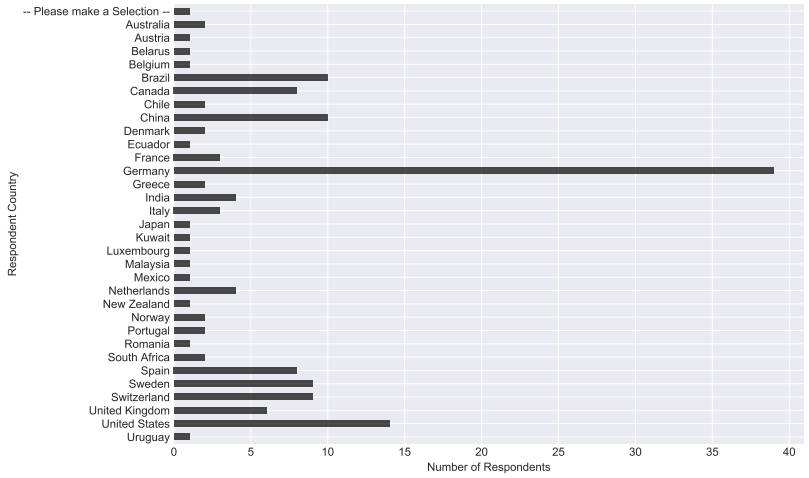



**Figure 3.1:** Number of Respondents per Country[6]

As Figures 3.2 and Figure 3.3 (→ p. 25) show, our respondents come from a variety of professional backgrounds. The dominant sectors are *information and communications technology* (ICT), *automotive*, and *financial services*, and the prevalent roles are *requirements engineer* and *business analyst*. From cross-tabulations, we can further infer that the (potentially overlapping) groups represented most strongly in our sample are people working in the automotive sector in Germany, requirements engineers located in Germany, and business analysts working in the financial services sector (12 respondents each).[7] Because different roles and sectors are likely associated with different everyday challenges, our respondents' assessments of RE research relevance might vary considerably with their professional background. This limits the conclusions we can draw from our analyses since we cannot control for all demographic factors due to our small sample size.

5 For example, some cultures (which might be more dominant in certain countries) might discourage giving feedback that could be considered rude (e.g., calling a research paper `Unwise`) more than other cultures.

6 Responses to the question *In which country are you located?*, shown at the very end of the demographics section. Participants are asked to select one option from a given list of countries. The option `-- Please make a Selection --` indicates that some respondents do not disclose their country information (whether accidentally or intentionally). Note that participants are asked for their *current location* rather than their country of birth or their citizenship.

7 See the online companion for graphical and tabular representations allowing for these insights.

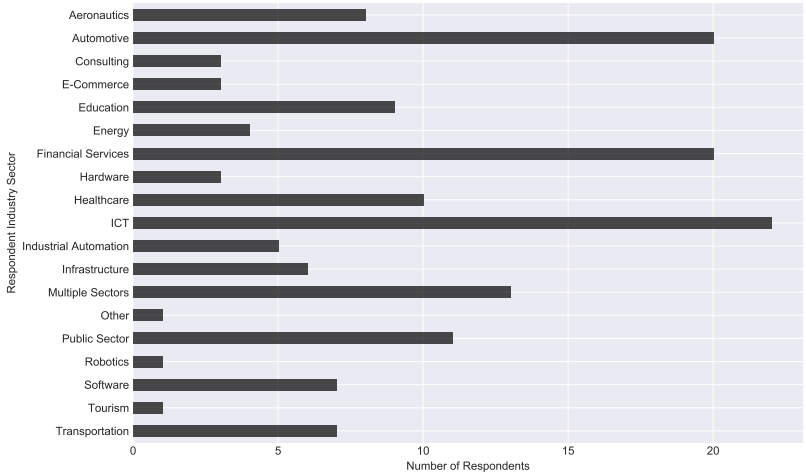



**Figure 3.2:** Number of Respondents per Sector[8]

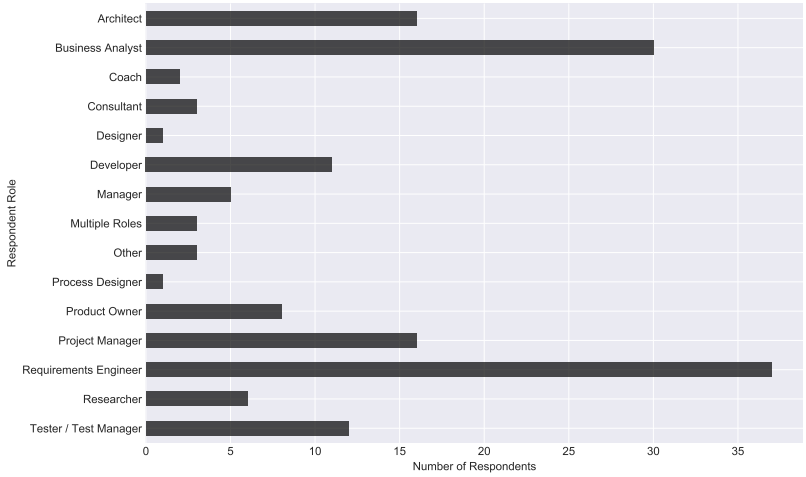



**Figure 3.3:** Number of Respondents per Role[9]

The majority of our respondents (117, 76.0 %) has a degree in computer science (CS) or a related field, but there are also 37 respondents (24.0 %) without such a degree.[10] While

8 Responses to the question: *What is the industry sector in which you are most frequently involved?* As participants are not shown any predefined options, they have to give short free-text answers. We code these answers semi-automatically, using regular expressions, drawing the sector categories from the data in an interactive coding process, and finally arriving at the statistics presented above.

9 Responses to the question: *Which of the following roles describes your primary working area best?* Participants can choose one of several predefined options or check `Other (please specify)` to give a short free-text answer. We code the free-text answers semi-automatically, reusing existing categories where possible, to enable aggregation and produce the statistics reported above.

10 Responses to the question: *Do you have a degree in Computer Science or a related field (such as computer engineering or information systems)?* This formulation leaves considerable room for interpretation (what is a "related field"?).

many respondents report up to ten years of RE experience (10.0 being the median of the response value distribution), our sample seems to include some RE veterans, too, as can be seen from Figure 3.4 (→ p. 26). Most of our respondents work in small- or medium-sized teams (→ Figure 3.5, p. 27), and almost half of our respondents develop (business) information systems in their projects (→ Figure 3.6, p. 27). As apparent from the graphics, the answers to our questions concerning respondents' team size and the type of system in scope of their projects are less fine-grained (by question design) than, e.g., the answers to the questions regarding respondents' sector or role (→ Figure 3.2 and Figure 3.3, p. 25). With these basic statistics regarding our respondents in place, we now turn to our research paper rating data, which is the focal point of the next section.

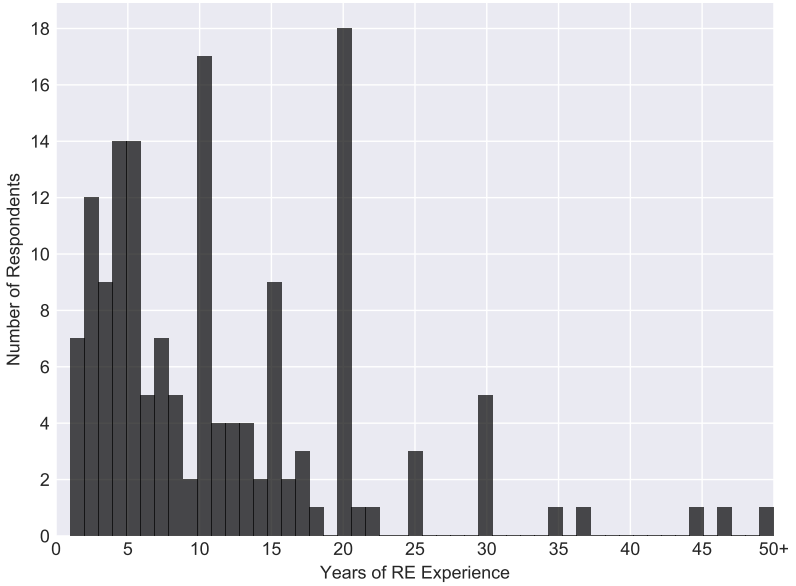



**Figure 3.4:** Respondents' Years of Experience in RE[11]

11 Responses to the question: *How many years are you working in your primary working area?* Participants entered their response into a free-text field; responses are coded using regular expressions and binned for the purposes of the graphic. Note that the precise wording of the question allows for different interpretations, which leads to some answers containing more than just a number; see the online companion to this thesis for how these cases are treated. Furthermore, some of the highest values appear almost implausible and might be the result of input errors (e.g., the omission of a dot between two digits) rather than correct representations of respondents' working experience.

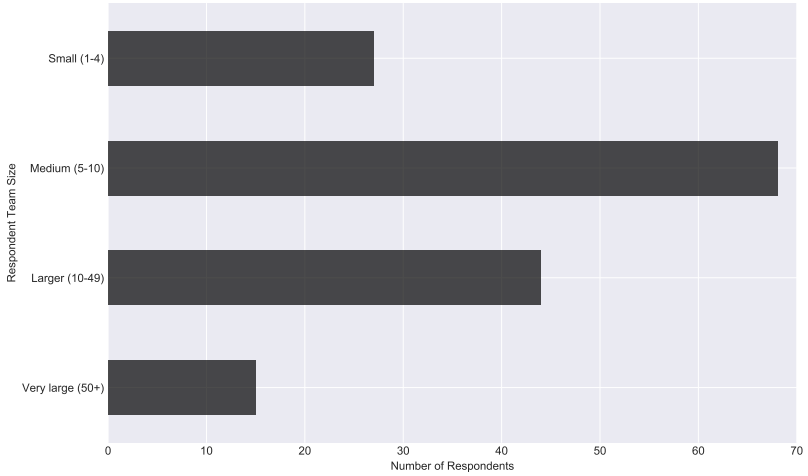



**Figure 3.5:** Number of Respondents per Team Size[12]

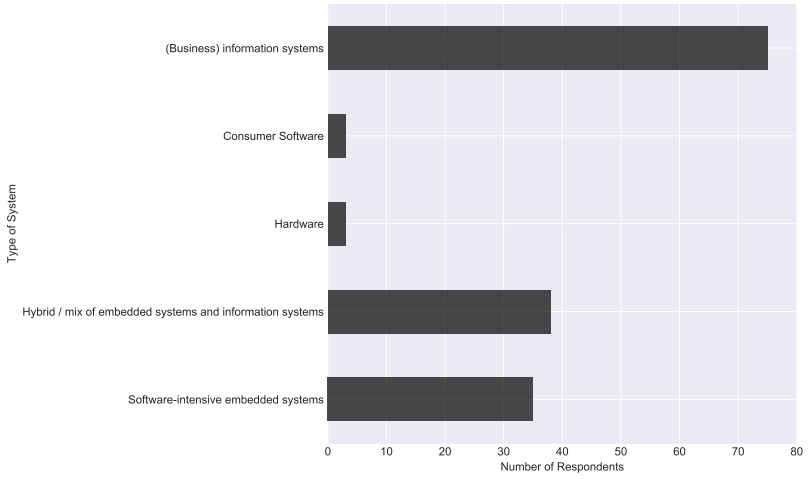



**Figure 3.6:** Number of Respondents per Type of System[13]

## 3.2 Rating: Practitioners' Judgments

So far, we have mostly been concerned with our respondents' characteristics, using the individual respondent as our unit of analysis. Now, we shift our focus to the central pieces of the RE-Pract survey, making the individual rating—a triple of type (`Respondent`, `Paper`,

12 Responses to the question: *What is the size of your team?* Participants are asked to select one of the options presented in the graphic.

13 Responses to the question: *What class of systems is in scope of your project(s)?* Participants are asked to select one of the three options with the highest counts but could also specify `Other` and give a free-text answer. We code the free-text answers manually to arrive at the statistics presented above.

`Rating`)—our unit of analysis. In the following, we refer to such triples collectively as our *rating data* and to individual triples as *ratings*.

From the 154 completed surveys, 2164 ratings can be gathered. They represent respondents' answers to the question: *In your opinion, how important are the following pieces of research?* Respondents can choose from the four-item Likert scale (`Essential`, `Worthwhile`, `Unimportant`, `Unwise`), and they have to base their answers on the one-sentence paper summaries presented to them.

While the majority of respondents cast the full 15 votes (113 participants, 73.38 %), there are also seven respondents (4.55 %) who cast between five and nine votes only, and even two respondents (1.30 %) who cast no votes at all. Since every participant rates a randomly drawn subset of the 435 selected RE research papers, the number of ratings collected per paper varies from 0 (six papers, 3.90 %) to 13 (one paper, 0.65 %). Figures 3.7 and 3.8 (→ p. 28) show the frequency distribution of paper ratings over respondents and over papers, respectively. Since we mostly have multiple ratings *by* the same respondent and multiple ratings *for* the same paper, our individual observations are not independent. Thus, it is barely possible to discern the causes of certain ratings: they might be driven by respondent characteristics, by paper characteristics, by a combination of or relationship between both, or by external factors not captured in the survey.

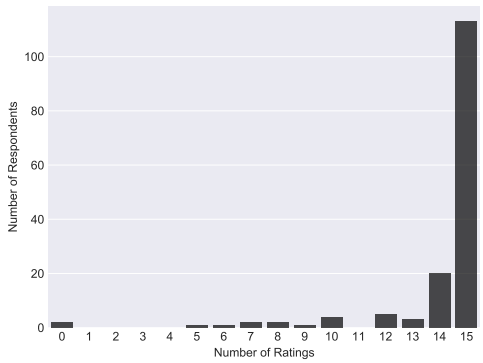



**Figure 3.7:** Distribution of Paper Ratings over Respondents

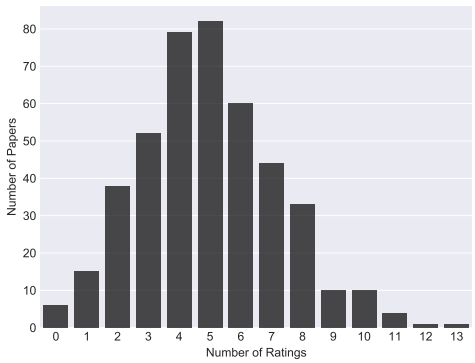



**Figure 3.8:** Distribution of Paper Ratings over Papers

The distribution of paper ratings over Likert categories is depicted in Figure 3.9 (→ p. 29). This figure allows us to give a first, high-level answer to **RQ1**, *What is the relevance of RE research to practitioners in the industry?*: From our 2164 ratings, roughly 1500 ratings (∼ 70 %) are positive, and almost 25 % are very positive. Thus, in the large majority of their ratings,

practitioners opine that the research summaries they were shown have some relevance to their practical context. At the same time, 30 % of practitioners' ratings are negative, and Figure 3.9 (→ p. 29) highlights that the group of `Essential` ratings is roughly the same size as the group of `Unimportant` ratings. Since the survey does not include a neutral category in the Likert scale, the overall positive impression should therefore be taken with a grain of salt: Even when genuinely indifferent towards the research to evaluate, respondents are forced to choose between `Worthwhile` and `Unimportant` (unless they decide to skip a research item entirely). In this case, it is likely that they opt for the socially more adequate alternative (`Worthwhile`) as they are effectively prevented from stating their true opinion.

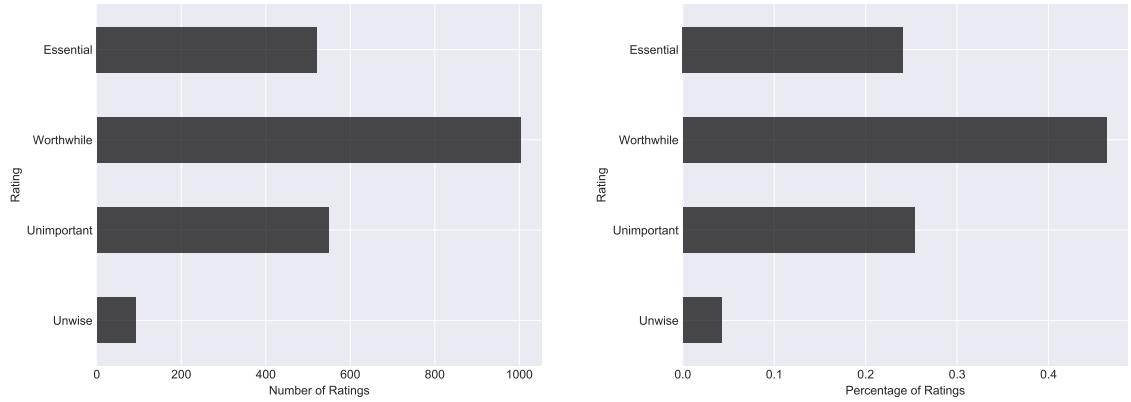



**Figure 3.9:** Distribution of Paper Ratings over Likert Categories

Due to the low numbers of ratings available both *from* each respondent and *for* each paper (→ Figures 3.7 and 3.8, p. 28), we need to group the ratings to gain further insights from the survey data. In the following, we therefore analyze our rating data by aggregating it on features pertaining to the first two elements of each rating triple, i.e., by respondent characteristics (→ 3.2.1), by paper characteristics (→ 3.2.2), or by respondent *and* paper characteristics (→ 3.2.3).[14] As we often have vastly different numbers of ratings available for each of the resulting groups, we usually present both our count and our percentage data to lower the risk of misinterpretation. We use the same color map, ranging from dark blue (`Essential`) to dark red (`Unwise`), in all of our rating graphics.

14 In all subsequent analyses, it is generally understood that our results might be driven by factors we could not control for, and that our observations might well be the results of statistical fluctuations or quirks in the pairing of respondents and papers. See chapter 4 for a discussion.

### 3.2.1 By Respondent Characteristics

Since we have the demographic data summarized in section 3.1.2 for all of our respondents, we can derive statistics for our rating data sliced by any demographic feature. In the following, we demonstrate the potential of this procedure for two features, respondents' business sectors and respondents' roles. Using the code made available in the online companion, similar analyses can be performed for all other demographic features collected in the survey.

**Respondent Sector** Aggregating the ratings by respondent sector, we arrive at the count data shown in Figure 3.10 (→ p. 30). This figure first and foremost highlights the uneven distribution of ratings over sectors, reflecting the uneven distribution of respondents over sectors (→ Figure 3.2, p. 25), but it is necessary to contextualize the percentage data depicted in Figure 3.11 (→ p. 30). For example, the extremely negative ratings from the tourism sector are based on very few ratings—in fact, if we go back to Figure 3.2, they are based on just one respondent (→ p. 25).

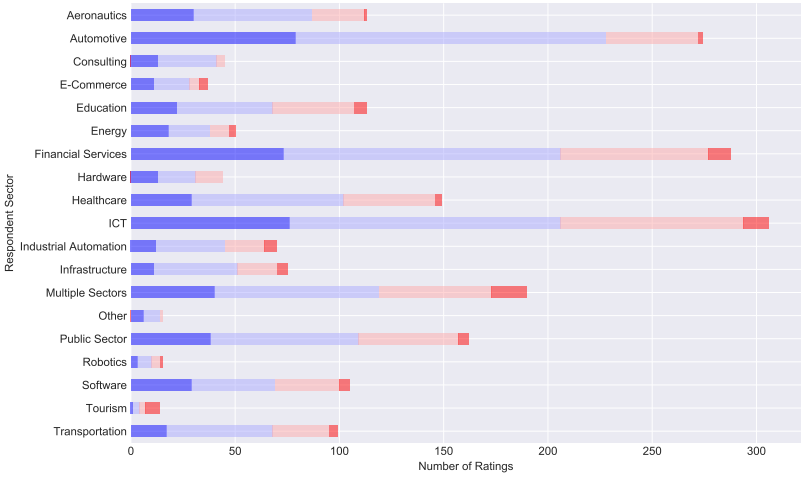



**Figure 3.10:** Ratings by Respondent Sector (absolute)

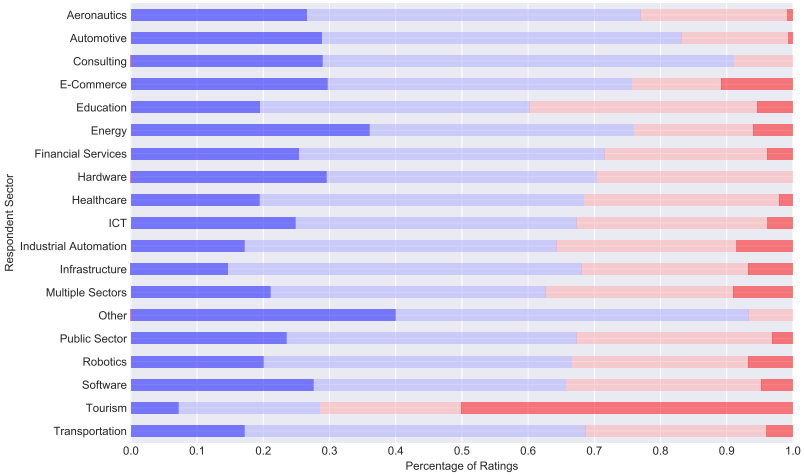



**Figure 3.11:** Ratings by Respondent Sector (relative)

To mitigate the differences in the numbers of respondents and ratings for different sectors, we focus our discussion of the percentage data (→ Figure 3.11, p. 30) on the five sectors with ten or more respondents: *automotive*, *financial services*, *ICT*, *multiple sectors*, and *public sector*. Comparing the ratings between these sectors, we find that the relatively highest satisfaction with RE research is expressed by participants from the automotive sector. Satisfaction levels of the financial services, the ICT, and the public sector are fairly similar; the relatively most critical ratings come from respondents marked "Multiple Sectors". If one expects practitioners with this sector label to share problems and experiences from the multiple sectors they work in, this statistic might appear surprising. However, one interpretation could be that practitioners working in multiple sectors have specific problems which are not sufficiently addressed in the literature.

**Respondent Role** Grouping our rating data by respondent roles, we can provide a first, high-level answer to **RQ5**: *Do practitioners' perceptions and views differ in dependence on their roles?* Again, we are dealing with an uneven distribution of ratings over roles (→ 3.12, p. 32)—and again, we focus our discussion of the percentage data (→ 3.13, p. 32) on the roles with ten or more respondents (→ Figure 3.2, p. 25): *architect*, *business analyst*, *developer*, *project manager*, *requirements engineer*, and *tester / test manager* (in the following: *tester*).

Based on our rating data, testers appear to be relatively most satisfied with existing RE research; they rated no research summary as `Unwise`, over 80 % of their ratings are positive, and a comparatively large share of their ratings is `Essential`. Two (compatible) explanations come to mind. First, RE research might cater more to testers than to RE practitioners in other roles. Second, people who are testers might generally rate RE research more favorably. We examine the first explanation more closely in section 3.2.3.

There, we also come back to the observation that *overall*, architects, business analysts, developers, project managers, and requirements engineers exhibit similar relative rating profiles—i.e., practitioners' perceptions and views, in the aggregate, do not appear to differ in dependence on their roles—although developers seem a little more satisfied in general, and project managers appear to avoid judging research as `Unwise`. However, the more nuanced picture provided in section 3.2.3 relies on data aggregation not only on respondent characteristics but also on paper characteristics. Taking one step at a time, we therefore continue by looking at our rating data when aggregated *solely* on paper characteristics.

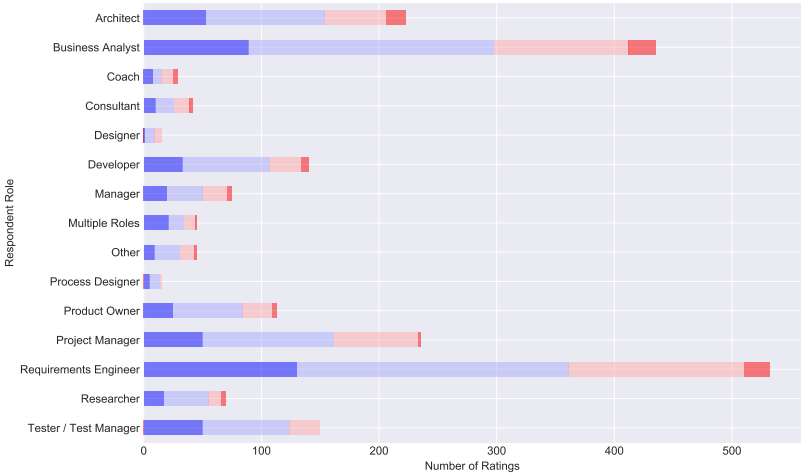



**Figure 3.12:** Ratings by Respondent Role (absolute)

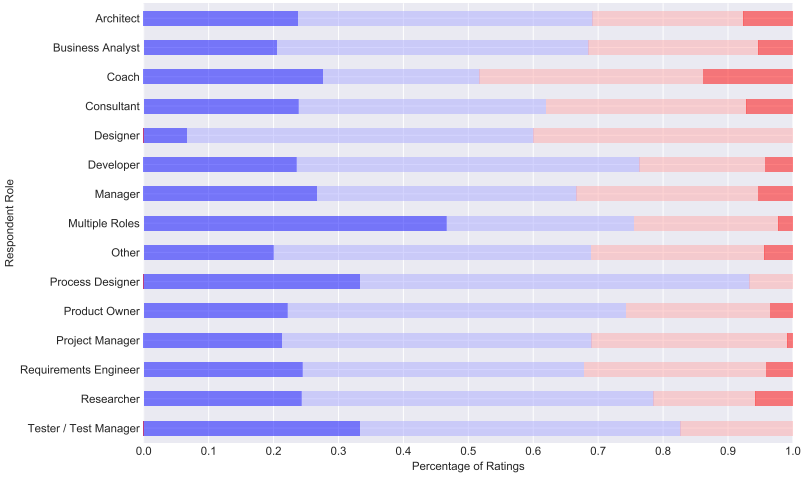



**Figure 3.13:** Ratings by Respondent Role (relative)

### 3.2.2 By Paper Characteristics

Our options to group the ratings by paper characteristics are more diverse than the options for grouping by respondent characteristics. First, we can aggregate on features available in our paper metadata; this type of aggregation is similar to the aggregation on demographic data performed in the previous section. Second, we can aggregate our rating data on features engineered in the second chapter—i.e., information regarding the methods and contents mentioned in the one-sentence summaries—by leveraging the tags we assign to the 435 papers in our mapping of RE research.

### Paper Metadata

We start by aggregating on features contained in our paper metadata. As explained in section 2.1 and summarized in Tables 2.1 and 2.2 (→ p. 4), we have access to a paper's *publication year*, *publication venue*, and *conference track* (whether *academic* or *industry*), as well as to its *author affiliations* (*academic*, *industry*, or *mixed*). In the following, we (briefly) highlight some observations made possible by aggregating our rating data on these features.

**Publication Year** As Figure 3.14 (→ p. 33) shows, our ratings are much more evenly distributed over publication years than over the respondent features previously examined. Against this background, the marked decline of perceived research relevance from 2010 to 2016 might appear worrying. The shift from around twenty to around thirty percent of research deemed `Unimportant` seems to be a macro-level trend as it cannot be explained by the decrease in perceived relevance of research published at a particular venue.[15]

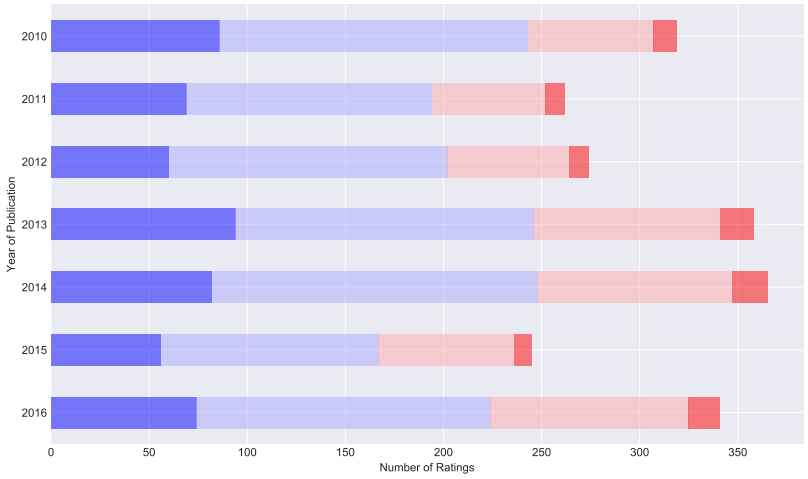



**Figure 3.14:** Ratings by Publication Year (absolute)

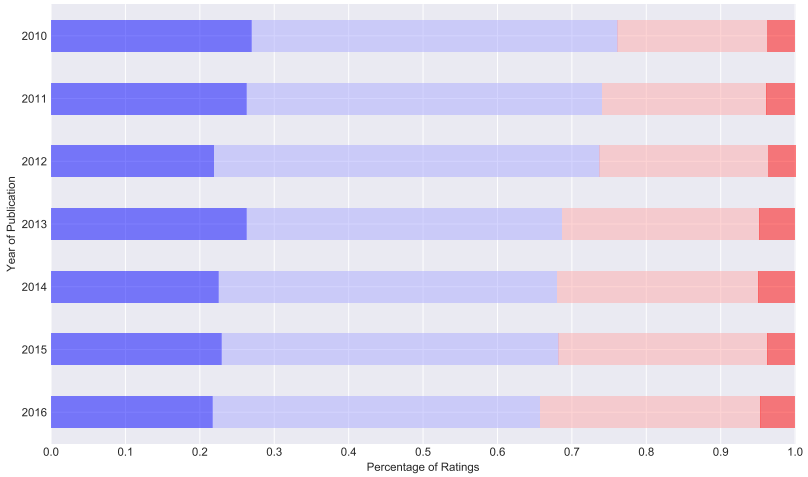



**Figure 3.15:** Ratings by Publication Year (relative)

15 See the online companion to this thesis for a view of the data as aggregated by year *and* publication venue.

**Publication Venue** Looking at the ratings aggregated by publication venue, the dominance of the *RE* and the *REFSQ* conferences in the rating counts (→ 3.16, p. 34) is unsurprising given their focus on RE. In the percentage data, REFSQ compares favorably against all other conferences, and views on ESEC/FSE and FSE (when held alone) appear relatively polarized. Notably, with REFSQ and RE, the two specialized conferences enjoy the highest relevance ratings from RE practitioners.

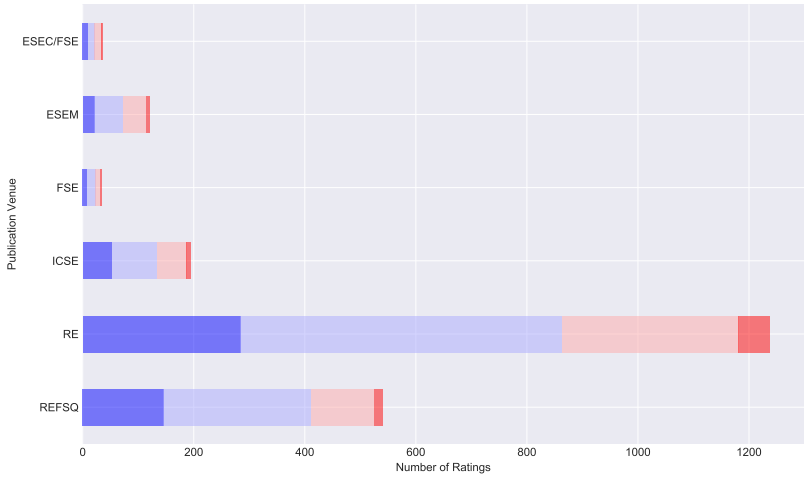



**Figure 3.16:** Ratings by Publication Venue (absolute)

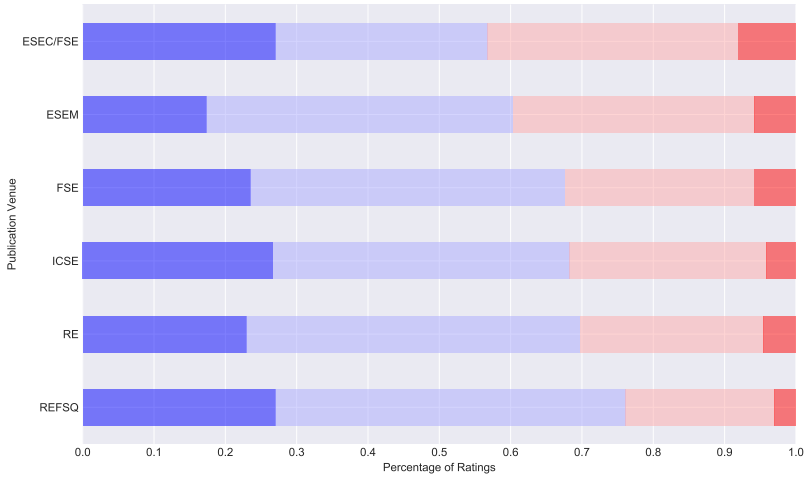



**Figure 3.17:** Ratings by Publication Venue (relative)

**Author Affiliation** Grouping by author affiliations, we reach the realm of **RQ4**: *Do papers with explicit ties to industry have higher practical relevance than other papers?* Our rating data is dominated by authors with exclusively academic affiliations (→ 3.18, p. 35), which calls for a cautionary interpretation of the percentage data (→ 3.19, p. 35). From that data, two tendencies might be discerned: First, research by authors with only industry affiliations

seems relatively more likely to be rated `Essential` by RE practitioners. Second, research by authors with some kind of industry affiliation (mixed or exclusive) appears relatively more likely to receive a positive rating than research by authors with only academic affiliations.

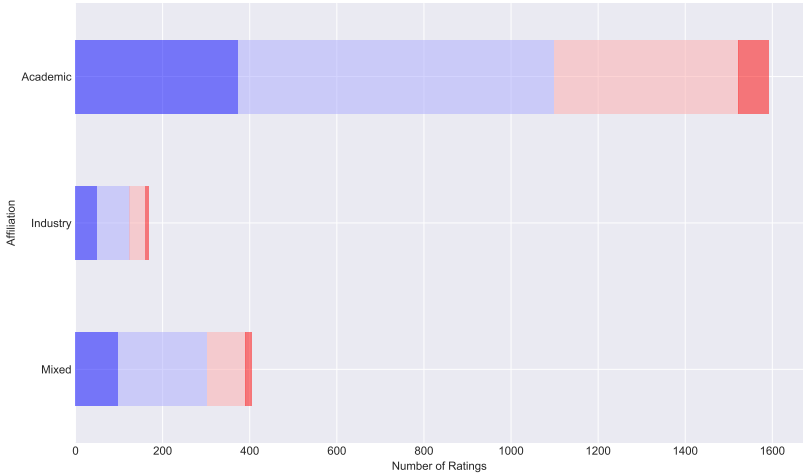



**Figure 3.18:** Ratings by Author Affiliation (absolute)

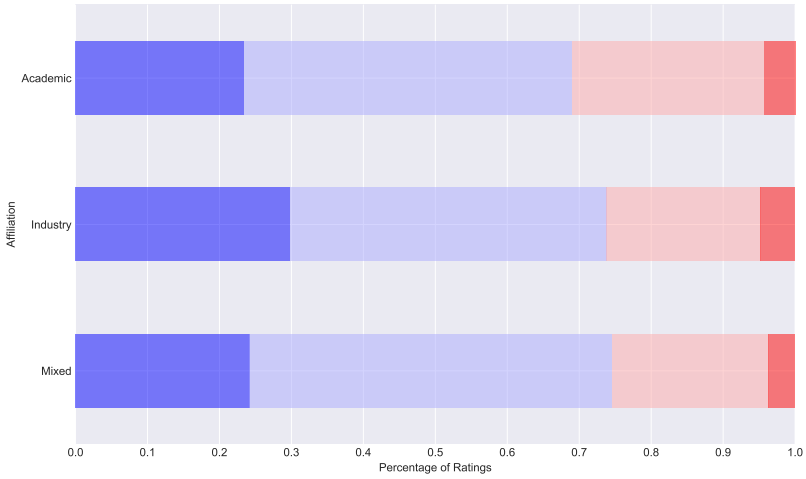



**Figure 3.19:** Ratings by Author Affiliation (relative)

One possible explanation is that author affiliations are correlated with choices of paper topics (methods and contents): Authors from industry are more likely to be familiar with the industry's most pressing problems and, consequently, more likely to address these problems in their research using methods popular with practitioners (otherwise, they might hardly spend time on publications); teams with mixed authorship are likely to address problems of at least some relevance to industry (otherwise, there would be little incentive for academia and industry to collaborate). However, since the number of papers (and ratings) in the *academic* category is so much larger than the number of papers in the other two categories, the tendencies found should not be overinterpreted.

**Conference Track** From our analysis of author affiliations, the preliminary answer to **RQ4** is: *Yes, papers with explicit ties to industry have higher practical relevance than other papers*—at least as judged by the RE practitioners in our sample. This statement is further substantiated by the percentage data we obtain when grouping our rating data by the Boolean feature *Industry Track* (→ Figure 3.21, p. 36). Again, however, the number of ratings available for papers *not* on an industry track is much larger than the number of ratings available for papers on an industry track (→ Figure 3.20, p. 36) so that the statistics should be compared with caution. When interpreting the figures, we must also keep in mind that not all conferences offer a dedicated industry track, which might cause *technically* industry track research to be presented outside an industry track, thus ending up in the *Industry Track: No* category for the purposes of the RE-Pract survey.

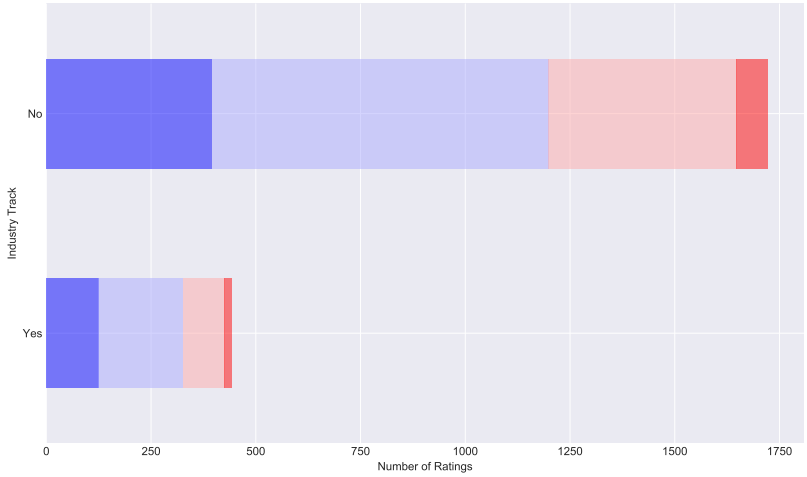



**Figure 3.20:** Ratings by Conference Track (absolute)

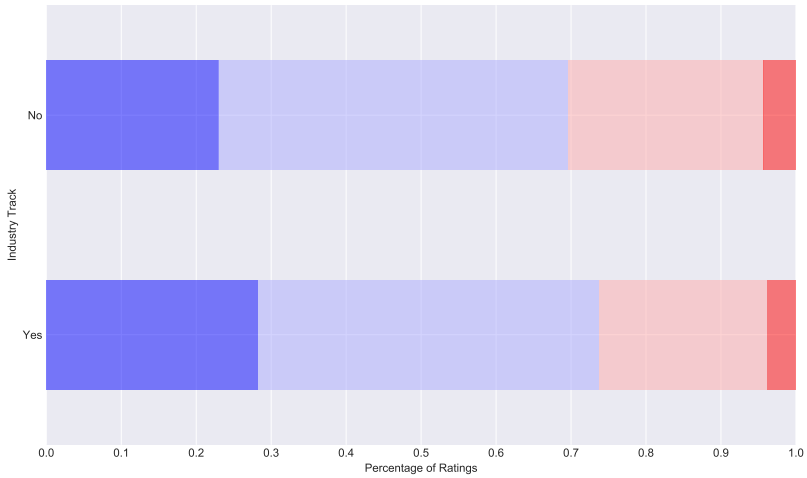



**Figure 3.21:** Ratings by Conference Track (relative)

**Author Affiliation and Conference Track** We can add one more nuance to our assessment of **RQ4**, *Do papers with explicit ties to industry have higher practical relevance than other papers?*, by combining author affiliation and conference track information. As Figure 3.22 (→ p. 37) demonstrates, unsurprisingly, most of the ratings we gathered refer to papers by academic authors not presented in an industry track, and fewest ratings refer to papers by industry authors presented in an industry track. Apart from these extremes, the numbers for other combinations of affiliation and industry track values are roughly comparable.

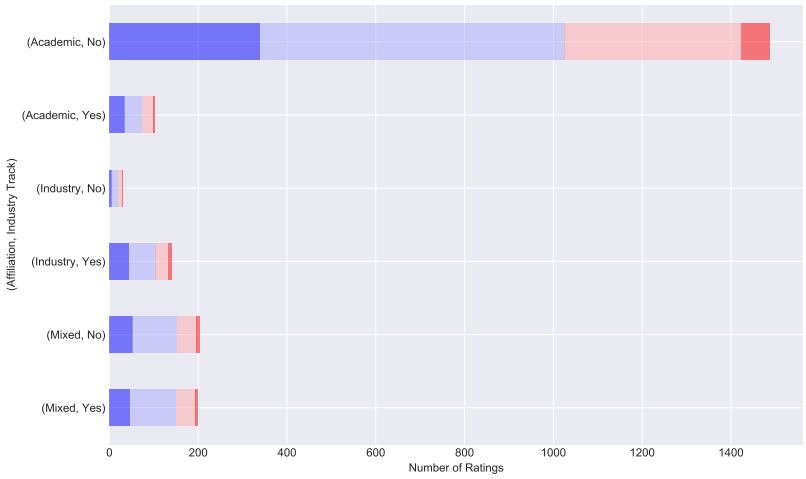



**Figure 3.22:** Ratings by Author Affiliation and Conference Track (absolute)

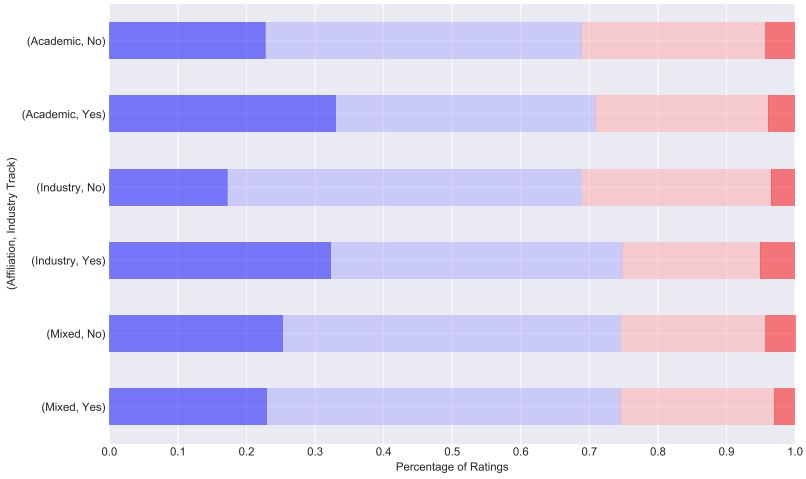



**Figure 3.23:** Ratings by Author Affiliation and Conference Track (relative)

Here, the percentage data (→ Figure 3.23, p. 37) provides interesting insights. First, the percentage of ratings in the `Essential` category is noticeably higher for ratings pertaining to papers published on an industry track either by academics or by industry practitioners—both when compared to research published by the same group outside of an industry track

and when compared to all other combinations of affiliation and industry track values. Second, while a higher fraction of `Essential` ratings is available for industry track research when the authors are either from academia or from industry, the same relationship does not hold for mixed-affiliation teams. Third, the least relevant research—as perceived by our respondents—seems to stem from industry practitioners publishing outside of their natural habitat, the industry track.

Collectively, our observations might be interpreted as follows. For authors with only academic or only industry affiliations, the choice of track (whether industry or not) is indicative of the choice of topic (higher or lower relevance to RE practitioners). For mixed teams, however, the fact that the team is mixed is more indicative of topic choice than the publication track chosen. This interpretation is in line with pragmatic considerations concerning how author teams choose the track in which they present their work. For purely academic or purely industry teams, there is a "natural" track choice, whereas for mixed teams, there is not. Academics need to actively decide (and choose their topic) to present in the industry track, whereas practitioners need to actively decide (and choose their topic) *not* to present in the industry track. The track decision of mixed teams might hinge on a conglomerate of (partially arbitrary) factors such as who is available to present, who has the lead in the project, or who is the primary audience.

In summary, therefore, our answer to **RQ4**, *Do papers with explicit ties to industry have higher practical relevance than other papers?*, should not be a confident *Yes* but rather a cautious *It depends*: Whether papers with explicit ties to industry have higher (perceived) practical relevance than other papers depends on the type of industry affiliation (*industry* or *mixed*) and on the publication track chosen (*Industry Track: Yes or No*).[16] As demonstrated above, we can fit the details of this finding into a sensible explanation. This explanation, however, assumes that what drives the relevance of RE research as perceived by RE practitioners is really the topic of the paper to be rated as defined by its methods and contents. This assumption is plausible *a priori* because the one-sentence summaries shown to our respondents provide cues to methods and contents but refrain from including the paper metadata. To provide further support for our assumption, and to tackle **RQ2**, we now turn to analyzing our rating

16 Note that in our use of the word *depend*, dependence does not imply causality.

data as aggregated on the paper characteristics added by our map of RE research: the RE research paper tags.

### Paper Tags

One of the main research questions the RE-Pract survey sets out to answer is **RQ2**: *What are the most highly rated research ideas?* Due to the RE research paper map developed in chapter 2, we are now in the position to investigate this question, providing some building blocks of an answer. Because of the low number of ratings available, we can only look at method and content factors separately—if we group by a combination of both, our data become too sparse to derive meaningful results. Since the one-sentence paper summaries presented to respondents mix method and content cues, this limits the conclusions we can draw from our analyses. With the necessary caution, however, we can still arrive at some interesting insights. To facilitate our analysis, the rows in the graphics that follow are sorted by the number (percentage) of positive ratings and—in case of a tie— the number (percentage) of `Essential` ratings.

**Method** In our map of RE research, two first-level tags refer to a paper's methods: the `how` tag and the `withwhom` tag. We examine these tags separately, aggregating our rating data by second-level and third-level paper tags.

As Figure 3.24 (→ p. 40) shows, our count data for the `how` tag are dominated by *engineering* papers concerning *methodology*. This is a consequence of the tag assignment strategy favored by one of the authors of the RE-Pract survey (already identified as problematic in section 2.3). Also, there are hardly any ratings on philosophical perspective papers, therefore, the comparatively high fraction of `Unwise` ratings in this category visible in Figure 3.25 (→ p. 40) should not be overrated. Similar caveats apply for the interpretation of the top spot for the percentage data as the number of ratings available for engineering reference papers is also relatively low. Among the categories with comparable numbers of ratings, the position of the three science categories—*interrogation*, *intervention*, and *observation*—in the sorted percentage data merits closer inspection. Here, it appears that RE practitioners have a strong preference for interrogation or observation research (e.g., surveys or case studies) over intervention research (e.g., experiments). Notably, intervention research not only ranks last

among all categories but also features the highest fraction of `Unwise` ratings (if we exclude the philosophy ratings due to the small number of observations they are based on).

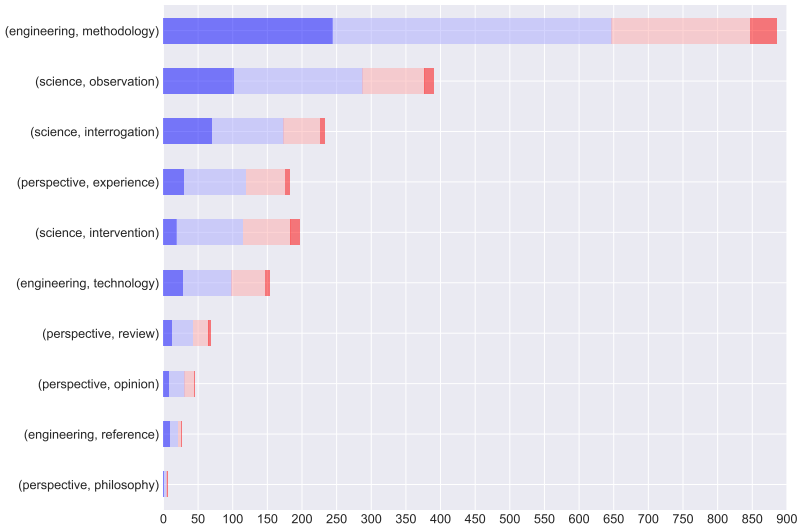



**Figure 3.24:** Ratings by First Level Paper Tag `how` (absolute)

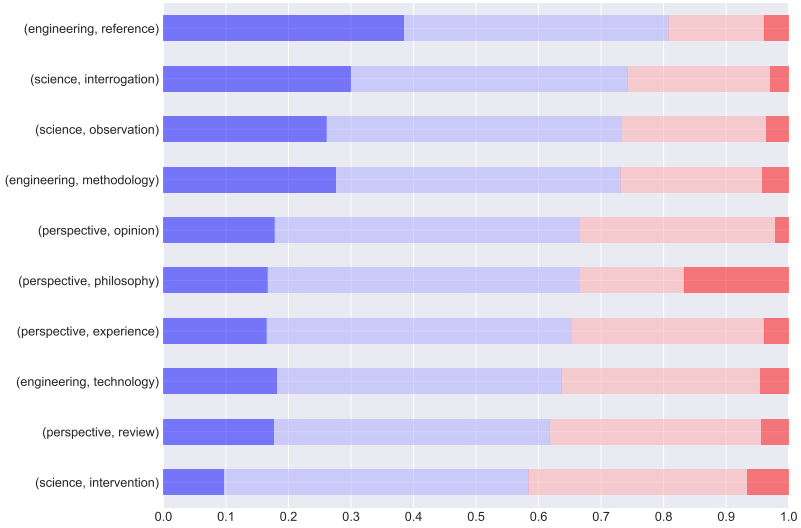



**Figure 3.25:** Ratings by First Level Paper Tag `how` (relative)

One possible explanation for the relative unpopularity of intervention research is directly related to the `withwhom` tags. Intervention research, by definition, includes all experiments, and many experiments are done with students. This begs the question: *Do practitioners dislike research with student participants?*

If we inspect Figures 3.26 and 3.27 (→ p. 41)—again, excluding the two categories with very few ratings—, the answer is likely: *yes.* The share of `Essential` ratings is much higher for research involving practitioners than for research involving students, and the share of

`Unimportant` and `Unwise` ratings is much higher for research involving students than for research involving practitioners. This result is perhaps unsurprising since the transferability of results obtained in a student laboratory to the real world is subject to longstanding discussions in many empirical disciplines. RE practitioners, it appears, have sided with the critics in that debate. Thus, the method part of an answer to **RQ2** might be framed in the negative: Highly rated research ideas do *not* involve student participation.

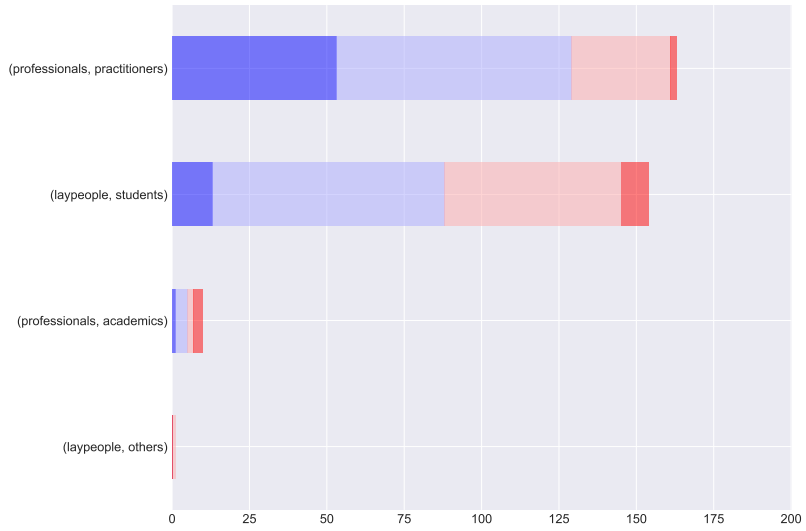



**Figure 3.26:** Ratings by First Level Paper Tag `withwhom` (absolute)

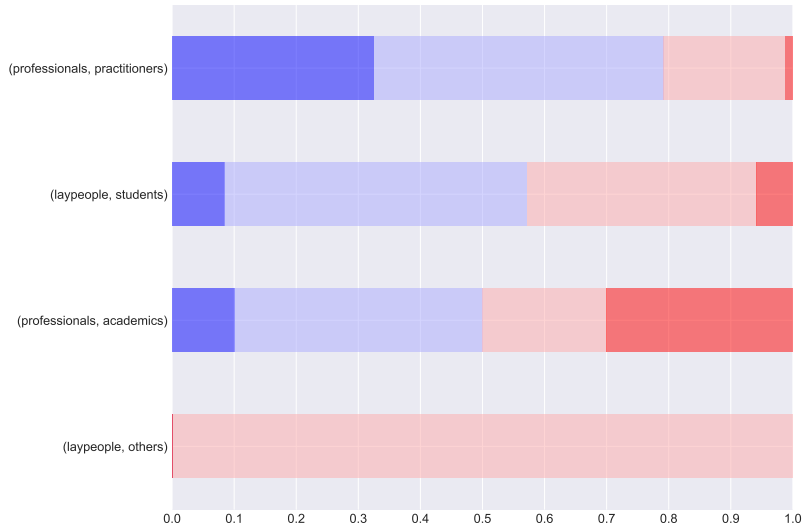



**Figure 3.27:** Ratings by First Level Paper Tag `withwhom` (relative)

**Content** Searching for the content part of our answer to **RQ2**, *What are the most highly rated research ideas?*, we can finally leverage our the paper tags under the first-level paper tag `what`. As Figure 3.28 (→ p. 42) shows, the taxonomy we are dealing with is too fine-grained to look at all lower-level tags simultaneously. This is no limitation, however, since

the tags are designed to characterize RE research content from different perspectives. We therefore look at the tags within selected second-level and third-level categories, aggregating by third-level and fourth-level tags as appropriate. The second-level content tags we examine are *information* (referring to the content of requirements), *documentation*, and *challenge*, where for the *challenge* tag, we inspect the third-level tags *content* (referring to attributes of requirements contents) and *people*.


**Figure 3.28:** Ratings by First Level Paper Tag `what`

As a baseline for comparisons, we retain from the percentage data in Figure 3.28 (→ p. 42) that most categories witness roughly $15-30\,\%$ of `Essential` ratings and $3-8\,\%$ of `Unwise` ratings. We also note that the tags under the first-level tag `what` are controversial amongst the authors of the RE-Pract survey, and that the risk of a skew in our results due to imperfections in the tag assignments is larger here than for the method tags discussed above.

**Information** If we restrict our analysis to ratings concerning papers with at least one second-level *information* tag, our attention when examining Figure 3.29 (→ p. 43) is first drawn to the high number of ratings available for research addressing the quality requirement *security*. Again, we have low numbers of ratings for several tags, and we focus our analysis of the percentage data (→ Figure 3.30, p. 43) on tags with $25$ or more ratings. Among these tags, the third-level *architecture* tag enjoys the highest percentage of positive ratings but research concerning the quality requirements *reliability*, *safety*, or *security* is rated `Essential` in a larger fraction of cases. If we exclude the *rules* tag (because of its small sample size), the highest percentage of `Essential` ratings is achieved by the *scenarios* tag. This can be taken to underscore the importance of the stakeholder perspective in RE, and it stands in stark contrast to the low percentage of `Essential` ratings obtained by research

with the *goals* tag, which also features one of the highest fractions of `Unwise` ratings. Therefore, one part of the content answer to **RQ2** might be: Highly rated research ideas may address scenarios, architectural requirements, or safety requirements.

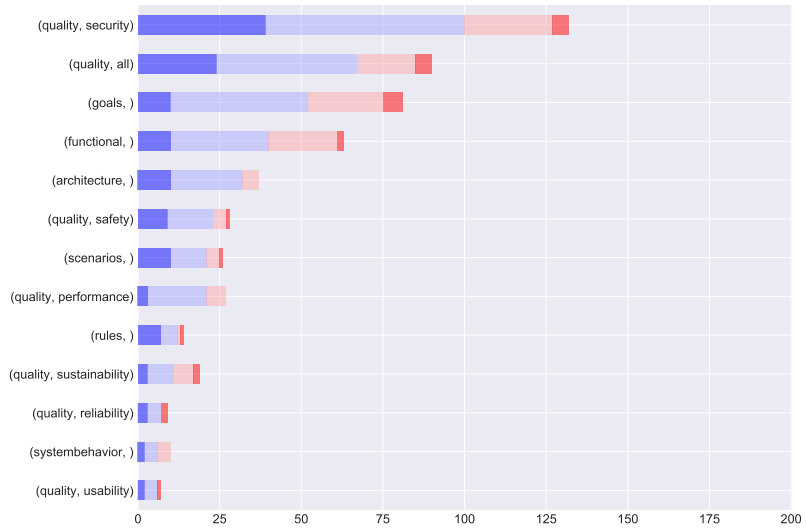



**Figure 3.29:** Ratings by Paper Content: Information (absolute)

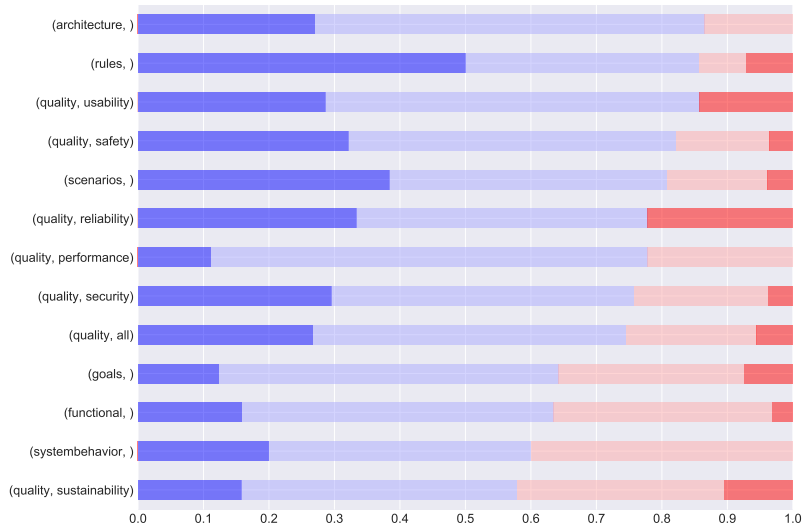



**Figure 3.30:** Ratings by Paper Content: Information (relative)

**Documentation** Turning to ratings for papers with at least one second-level *documentation* tag, Figure 3.31 (→ p. 44) once again reminds us of the prominent role of natural language in research addressing requirements documentation. Of the tags for which we have more than 25 ratings, research on artifacts ranks first and research on goal models ranks last according to the fraction of positive ratings, as evident from Figure 3.32 (→ p. 44). In particular, the low fraction of `Essential` ratings and the high fraction of `Unwise` ratings for goal mod-

els is striking, although it is in line with the comparatively weak position of research on goals under the second-level *information* tag discussed above. Only research on use cases sees a comparably high fraction of `Unwise` ratings, yet this fraction is counterweighted by a decently high fraction of `Essential` and overall positive ratings. The ratings profile for research on natural language is somewhat similar to that of categories which are much less represented in our sample, e.g., diagrams and feature models. Thus, the extraordinarily large number of papers concerning natural language is not met with extraordinarily high interest among RE practitioners. In sum, if we were to choose one aspect from our observations to include in the content part of our answer to **RQ2**, we could again add a negation: Highly rated research ideas do *not* concern goal models.

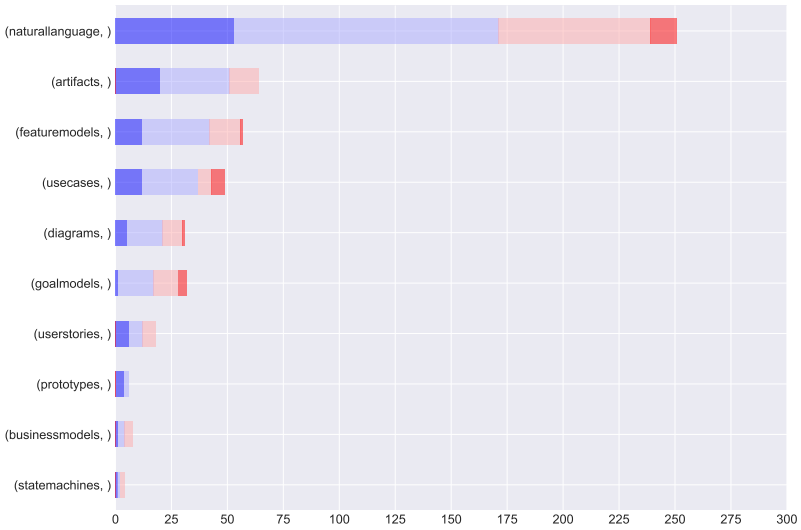



**Figure 3.31:** Ratings by Paper Content: Documentation (absolute)

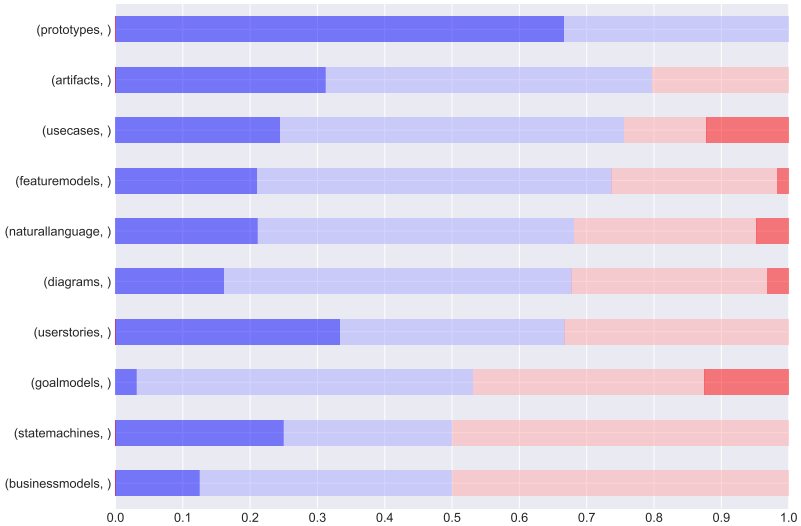



**Figure 3.32:** Ratings by Paper Content: Documentation (relative)

**Challenge** To wrap up the inquiry into our rating data as aggregated by our content tags, we examine two third-level tags under the second-level tag of *challenge*, namely *people* and *content*. The fourth-level tags grouped under the *people* tag refer to four human-related challenges in RE. Of these challenges, higher numbers of ratings are available for research addressing human communication and skills than for research on collaboration or subjectivity in human perception (→ Figure 3.33, p. 45), but the numbers of ratings in all categories are large enough to include them in our analysis of the percentage data (→ Figure 3.34, p. 45).

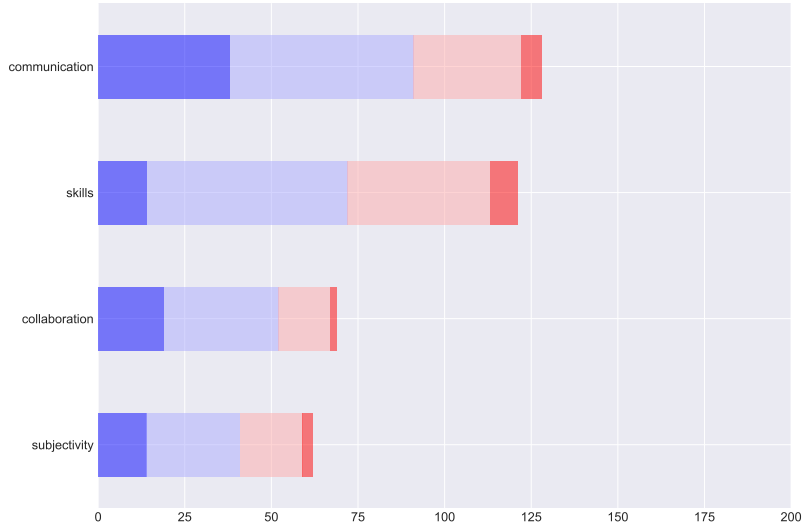



**Figure 3.33:** Ratings by Paper Content: Challenge — People (absolute)

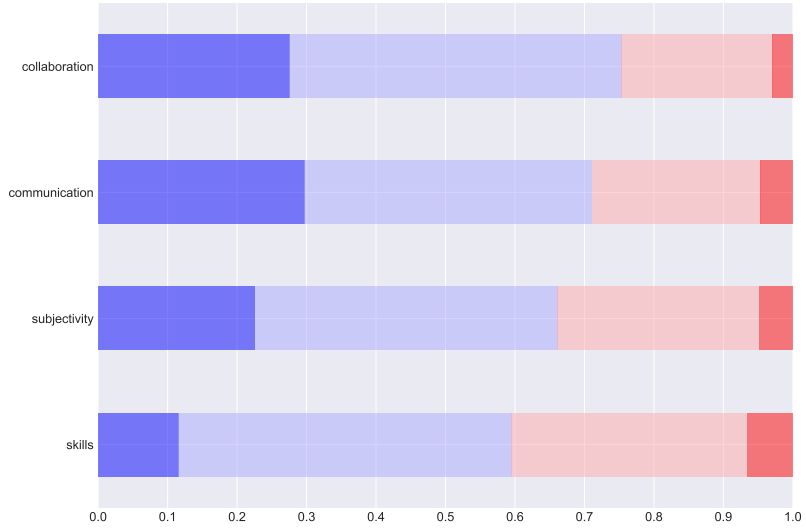



**Figure 3.34:** Ratings by Paper Content: Challenge — People (relative)

Here, the practitioners in our sample show a preference for research concerning the problems of collaboration and communication, although when compared to our baseline of $15-30\,\%$ `Essential` and $3-8\,\%$ `Unwise` ratings, the values we observe are not extraordinary. The preference for research on collaboration and communication appears remarkable

only when compared with research addressing the skills people might need in RE practice. Two factors might contribute to this result. First, RE practitioners likely experience communication and collaboration problems on a daily basis, which increases the chance they will rate research addressing these problems as important. Second, RE practitioners are unlikely to doubt that they have the skills necessary for their job, and they are even less likely to value academics telling them what skills they may need, which decreases the chance they will rate research addressing skill problems as important. Thus, we can again add a negation to the content part of our answer regarding **RQ2**: Highly rated research ideas do *not* concern people's skills.

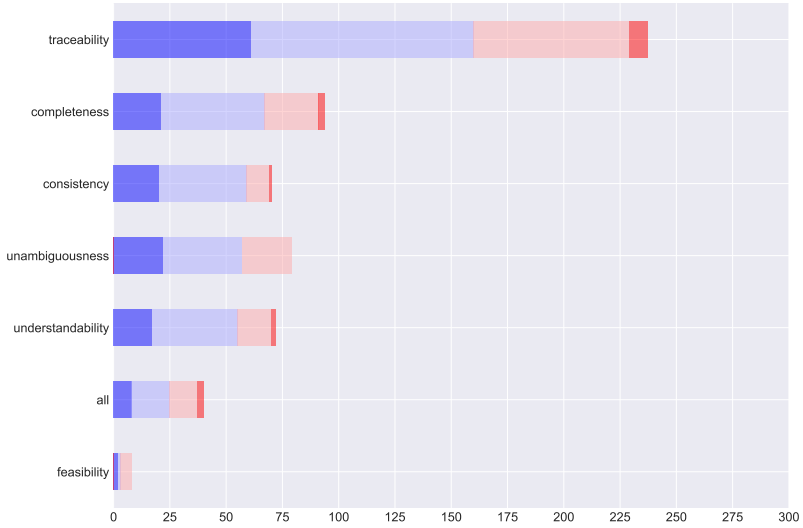



**Figure 3.35:** Ratings by Paper Content: Challenge — Requirements Contents (absolute)

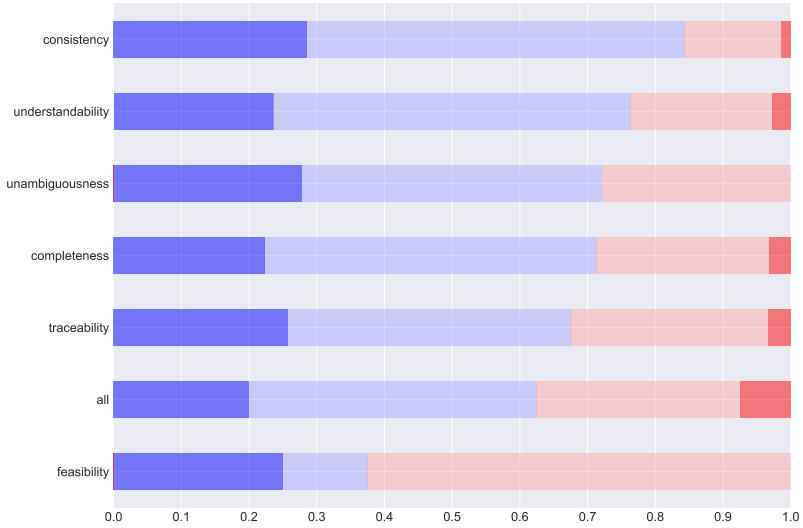



**Figure 3.36:** Ratings by Paper Content: Challenge — Requirements Contents (relative)

The fourth-level tags assembled under the *content* tag refer to content-related characteristics of requirements that are generally considered desirable, e.g., consistency, traceability, and

understandability. As Figure 3.35 (→ p. 46) shows, most ratings for research with a third-level *content* tag relate to papers on traceability, but completeness, consistency, unambiguousness, and understandability also feature a decent amount of ratings. The rating profiles based on the percentage data for these contents are also remarkably similar (→ Figure 3.36, p. 46), with consistency and unambiguousness seeming a little more popular than understandability, completeness, or traceability (especially from the percentage of `Essential` ratings). Since the number of negative ratings for research regarding consistency is particularly low, we might make one last addition to the content part of our answer to **RQ2**: Highly rated research ideas may address requirements consistency.

Collecting the answer pieces we derived in section 3.2.2, we can now state our full tentative answer to **RQ2**, *What are the most highly rated research ideas?*: From the 2164 ratings collected in the RE-Pract survey, the most highly rated research ideas seem to be ideas that:

— do not involve students *and*

— do not concern goal models *and*

— do not concern people's skills *and*

— may address scenarios, architectural requirements, or safety requirements *and*

— may address requirements consistency.

### 3.2.3 By Respondent and Paper Characteristics

So far, we have aggregated our rating data on *either* respondent characteristics *or* paper characteristics in order to address many of the research questions listed in section 3.1.1. Now, we pick up where we left off in section 3.2.1 to provide a more detailed answer to **RQ5**: *Do practitioners' perceptions and views differ in dependence on their roles?* We do so by grouping our rating data by respondent role and paper content tags simultaneously. Thus, we give one concrete example of how data aggregation on *both* respondent characteristics *and* paper characteristics can be used to obtain further insights from our ratings.

In the following, we ask how our respondents' ratings differ in dependence on their roles, focusing in reverse order on the tags we examined in the content tag part of the previous section: Starting with the *content* and the *people* tag under the second-level *challenge* tag, we move on to discuss the *documentation* tag and end with an inspection of the *information* tag.

For the sakes of brevity and comparability, we only present the percentage data for the six most prominent roles and order the rows alphabetically.[17]

**Respondent Roles and Requirements Challenges** Figures 3.37–3.42 (→ p. 48) show the content challenge data from Figure 3.36 (→ p. 46) broken down by respondent roles.

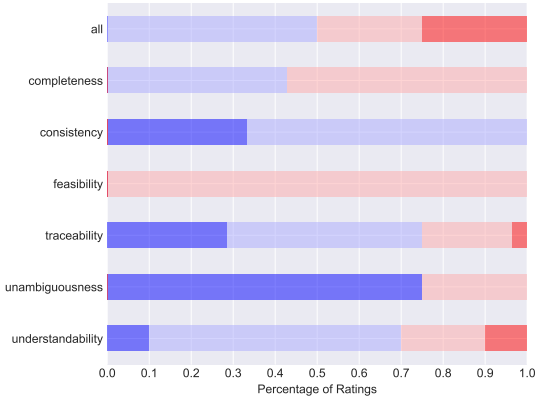



**Figure 3.37:** Architect

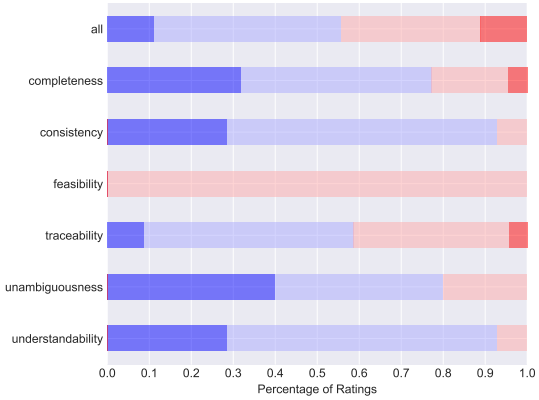



**Figure 3.38:** Business Analyst

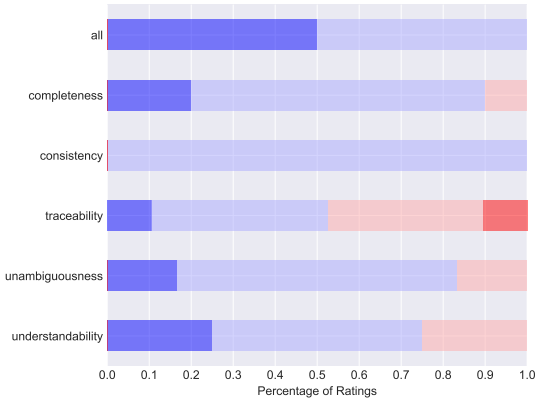



**Figure 3.39:** Developer

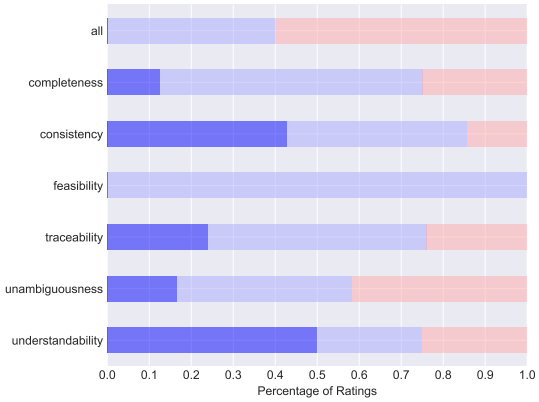



**Figure 3.40:** Project Manager

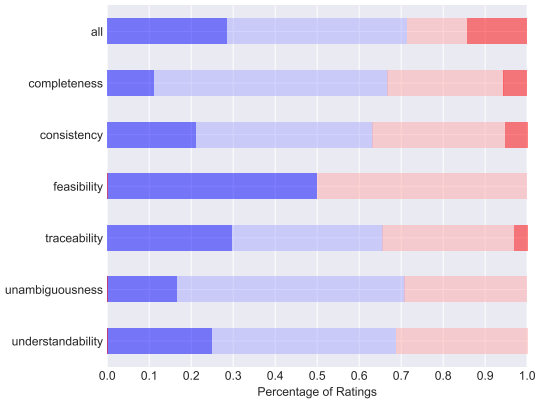



**Figure 3.41:** Requirements Engineer

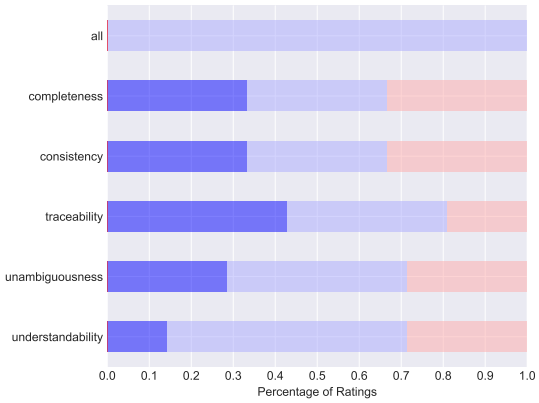



**Figure 3.42:** Tester / Test Manager

17 Since not all content tags feature ratings by all roles, some graphics have fewer rows than others. Graphics for the count data and for the less prominent roles can be generated using the code provided in the online companion to this thesis.

We remark that since the number of ratings in each category without grouping by respondent roles is already small (→ Figure 3.35, p. 46), the number of ratings underlying each category after grouping by roles is even smaller. This applies to all of the content features we investigate below; therefore, all conclusions we might draw from the observations noted here take the form of *potential directions for further research*.

For our analysis of the *content* tag under *challenge*, we again exclude *all* and *feasibility* from our comparisons. Looking at Figures 3.37–3.42 (→ p. 48), the following propositions appear plausible:

— Architects value research on unambiguousness but show comparatively little interest in research addressing completeness.

— Project managers value research on understandability and consistency.

— Testers value research on traceability.

— Business analysts and developers have relatively little interest in research on traceability.

— Requirements engineers have no clear preferences concerning research on challenges associated with requirements contents.

While these propositions may be in line with our intuition regarding the roles in question, they rest on a very narrow empirical basis and thus require further investigation.

Similar caveats apply to the propositions we can derive from Figures 3.43–3.48 (→ p. 50), which break down the people challenges data from Figure 3.34 (→ p. 45) by respondent role. Here, we might propose:

— Of all prominent roles, testers exhibit the strongest interest in research on collaboration.

— Of all prominent roles, developers are least interested in research on communication.

— Of all prominent roles, architects and developers are most interested in research on subjectivity.

— Business analysts and project managers find research on communication particularly essential.

Again, requirements engineers show a remarkably balanced preference profile, which might reflect the diversity of challenges associated with their central role in the RE process.

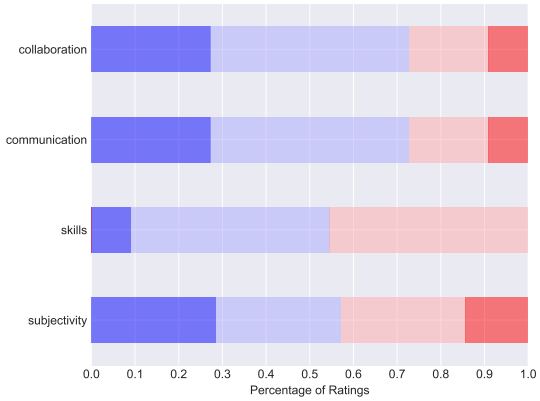



**Figure 3.43:** Architect

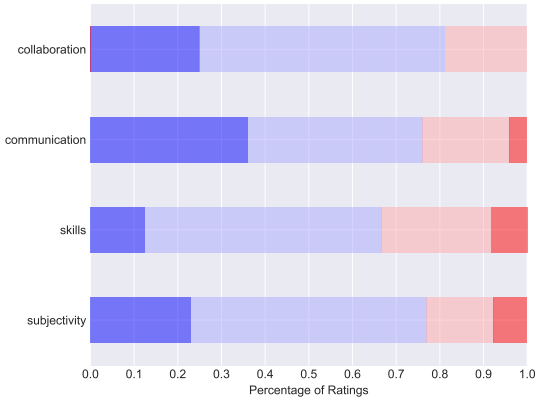



**Figure 3.44:** Business Analyst

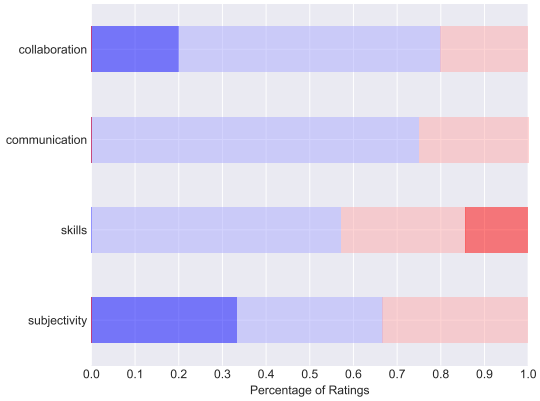



**Figure 3.45:** Developer

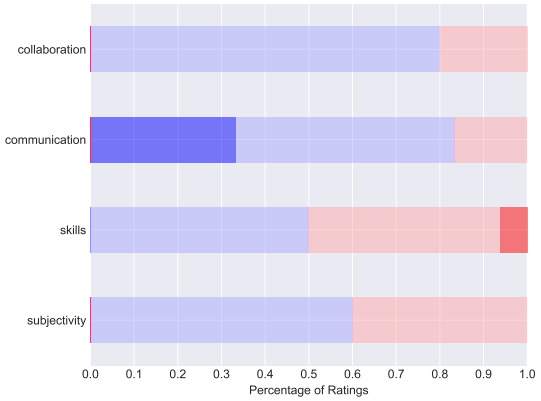



**Figure 3.46:** Project Manager

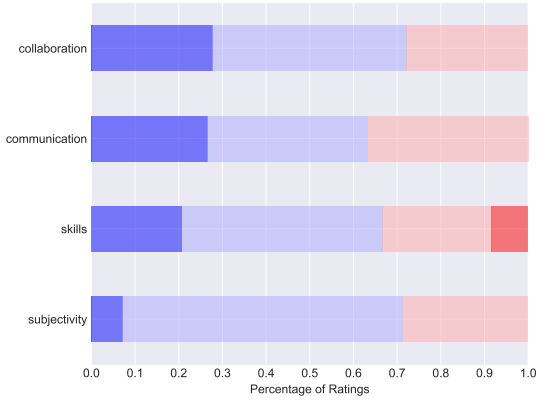



**Figure 3.47:** Requirements Engineer

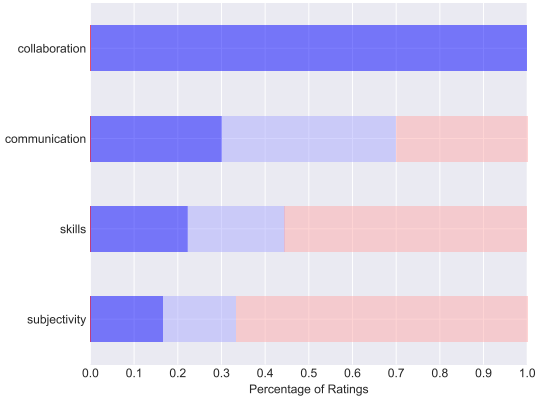



**Figure 3.48:** Tester / Test Manager

**Respondent Roles and Requirements Documentation** Figures 3.49–3.54 (→ p. 51) break down the requirements documentation data from Figure 3.32 (→ p. 44) by respondent role. Here, we exclude *state machines*, *business models*, *prototypes*, and *user stories* from our evaluation based on their small sample size (→ Figure 3.31, p. 44). From these figures, and with the same caveats as before, we might derive the following propositions:

— Architects value research on feature models and show comparatively little taste for research on goal models.

— Developers value research on diagrams and use cases.

— Testers are particularly interested in research on artifacts.

— Of all prominent roles, project managers show the highest levels of strong interest in research on goal models.

— Of all prominent roles, requirements engineers show the least interest in research on feature models.

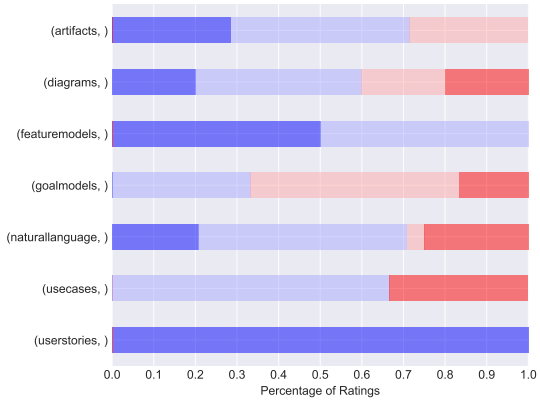



**Figure 3.49:** Architect

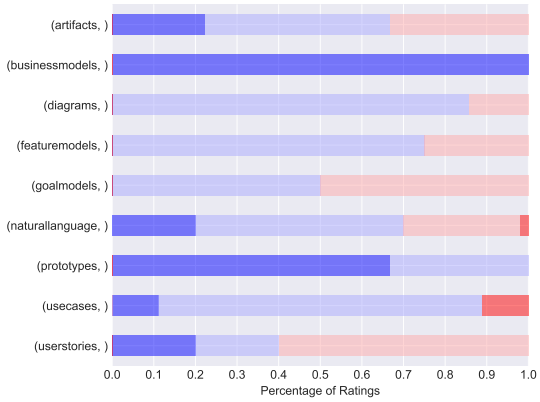



**Figure 3.50:** Business Analyst

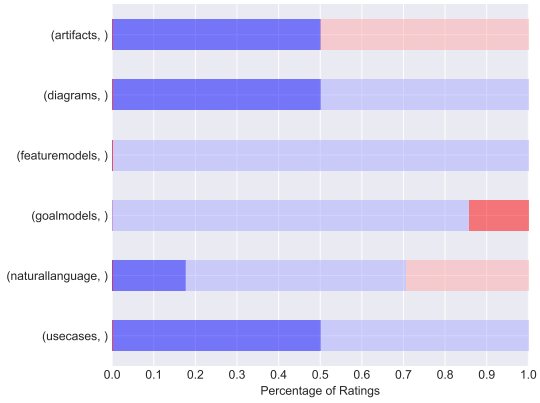



**Figure 3.51:** Developer

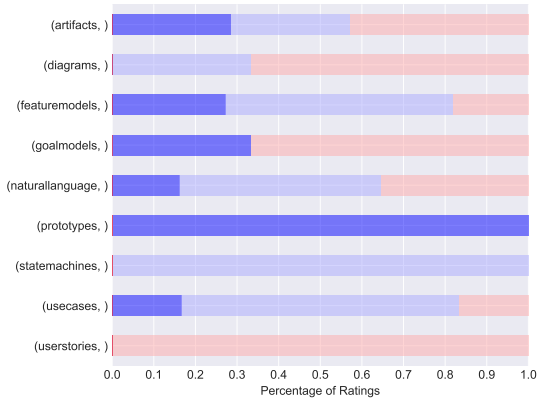



**Figure 3.52:** Project Manager

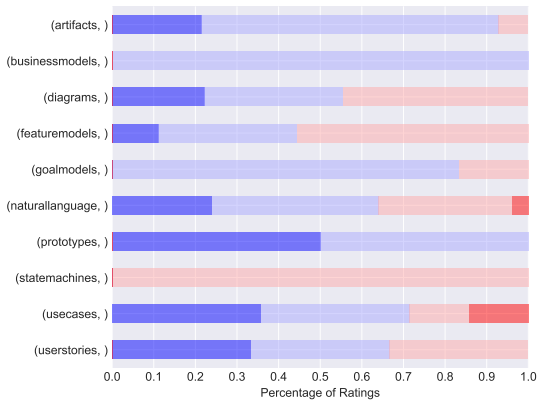



**Figure 3.53:** Requirements Engineer

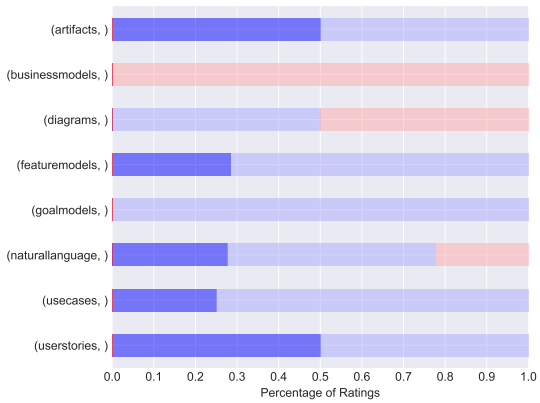



**Figure 3.54:** Tester / Test Manager

**Respondent Roles and Requirements Information** Finally, we slice our data on the information potentially captured in requirements (→ Figure 3.30, p. 43) by respondent roles to arrive at Figures 3.55–3.60 (→ p. 52). Here, we exclude *rules*, *system behavior*, and the quality requirements *reliability*, *usability*, and *sustainability* from our deliberations due to their small sample size (→ Figure 3.29, p. 43).

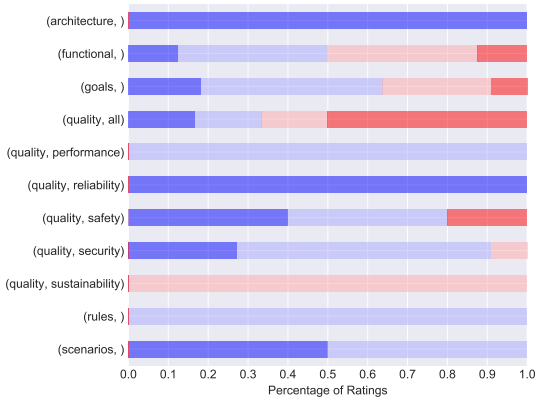



**Figure 3.55:** Architect

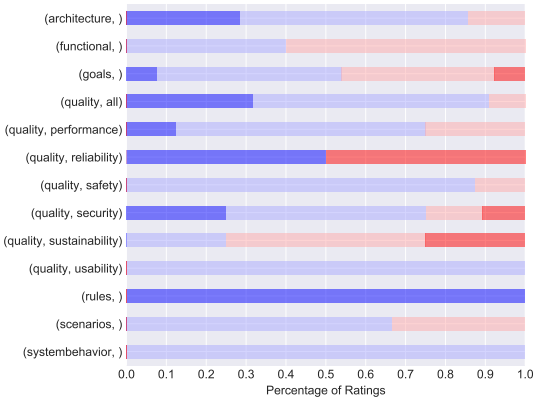



**Figure 3.56:** Business Analyst

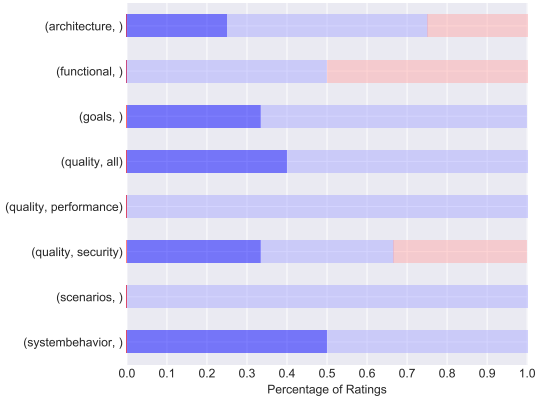



**Figure 3.57:** Developer

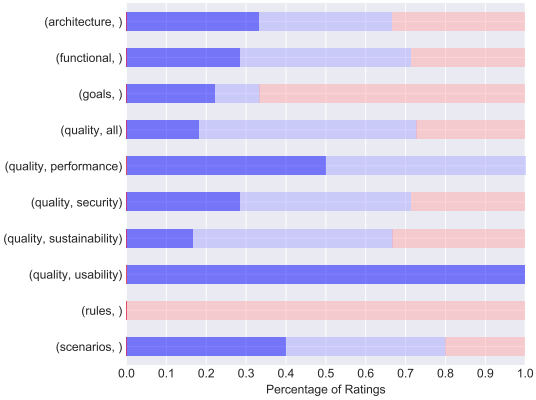



**Figure 3.58:** Project Manager

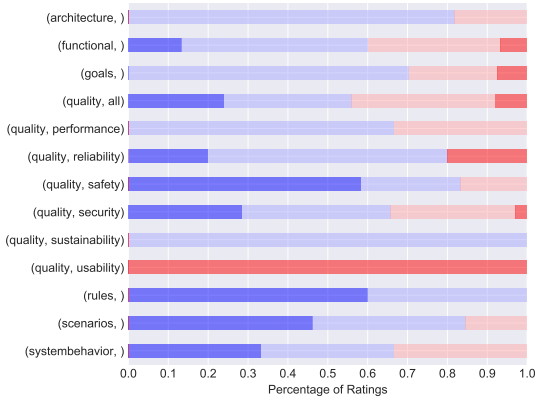



**Figure 3.59:** Requirements Engineer

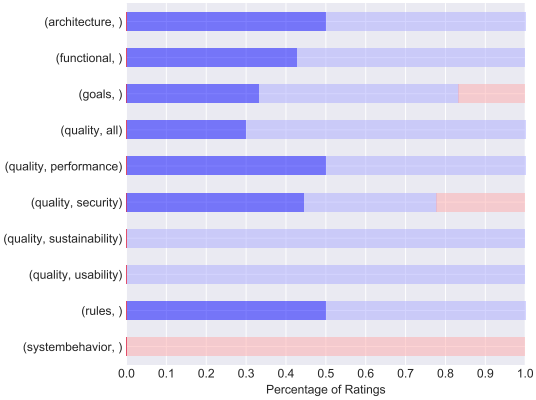



**Figure 3.60:** Tester / Test Manager

From these graphics, we might distill the following cautious propositions:

— Of all prominent roles, architects value research on architectural requirements most.

— Of all prominent roles, project managers and testers show the greatest interest in research on performance.

— Of all prominent roles, requirements engineers are most interested in research on safety.

Once again, we underline that the statistics presented in this section are based on very small numbers of ratings. Therefore, our results are likely not robust against statistical fluctuations or differences in opinion regarding the tag assignments. Consequently, our more nuanced answer to **RQ5**, *Do practitioners' perceptions and views differ in dependence on their roles?*, might be phrased as follows:

*Practitioners' perceptions and views **might** differ in dependence on their roles, and our data is more suggestive of some differences than of others. However, we do not have enough data to draw even preliminary conclusions.*

Thus, the rigorous investigation of our propositions and the eventual response to **RQ5** must be left for future research.

## 3.3 Reasoning: Practitioners' Thoughts

Our analysis of the rating data in section 3.2 has been based on the first part of the RE-Pract survey. In this section, we turn to the second part of that survey: the reasons practitioners give for particularly high ratings (→ 3.3.1) and particularly low ratings (→ 3.3.2) as well as the research wishes they voice (→ 3.3.3). This provides additional context and details for our answers to the research questions from section 3.1.1 we have already addressed, and it allows us to tackle **RQ3**: *What research problems do practitioners think are most important to be focused on by the RE research community?* In the following, all one-sentence summaries and practitioners' statements are reproduced verbatim (typos included).

### 3.3.1 Positive Rating Explanations

In the first question belonging to the second part of the RE-Pract survey, respondents are presented with the following prompt: *Please provide a brief explanation for why you provided one of the higest* [sic] *ratings to the following piece of research.* The research item shown to individual participants depends on their ratings in the first part of the survey and is selected automatically by the survey administration tool.

From our 154 respondents, 117 entered one or more characters in the free-text answer field, and responses are available for 103 of the 435 research papers included in the sample. Using regular expressions similar to those employed when creating our map of RE research in chapter 2, we can semi-automatically assign tags to practitioners' free-text statements. This allows us to quantify the reasons practitioners gave for their ratings. Table 3.1 (→ p. 54) gives examples of practitioners' explanations for their positive ratings and the tags they are assigned, along with the RE research paper summaries to which the ratings refer.

| **Summary** | **Reasoning** |
|---|---|
| A method for reasoning about likely sources of uncertainty in dynamically adaptive systems in order to apply the right adaptation strategies | Because uncertainty is where the trouble begins and the right adaptation strategies can reduce some sources of uncertainty which will always make the system better.<br>*reason:plausibility; reason:relevance* |
| A case study for validating a multi-level approach for planning and managing variability and reuse across independent product ranges in a product family | the title sounds like advanced RE and further problem-solving -> only few articles/literature exist, especially in this combination<br>*reason:originality* |
| A method for systematically and repeatedly exchanging requirements between manufacturers and suppliers. | Personal experience shows that this exchange is both extremely helpful and rarely implemented systematically.<br>*reason:plausibility; reason:relevance; source:opinion* |
| An analysis on the integration of non-functional requirements into model-driven development processes in order to include this type of requirements into such processes | non-functional reqs have the biggest impact on architecture and may not be left aside<br>*reason:relevance* |
| A case study on collaboration networks between small and medium-sized software companies in order to investigate the impact to software product management and requirements engineering practices. | In my opinion we need to share and exchange more information related to the practices that we have been using in order to increase our performance and maturity level.<br>*reason:plausibility; source:opinion* |
| | Communication (and so culture) is a key concern in requirements engineering and such a study could help understand interesting aspects in this field<br>*reason:originality; reason:plausibility; reason:relevance* |

**Table 3.1:** Examples of Reasons for Positive Ratings

As Table 3.2 (→ p. 55) highlights, the overall diversity in high-level explanations for positive ratings is limited. Many respondents make some reference to the perceived relevance of the research item they are shown, providing further details as to what aspect of the research is responsible for their judgment (see, e.g., the first item in Table 3.1). Similarly, participants often state that the research approach presented appears plausible to them in one way or another (see, e.g., the last item in Table 3.1). In contrast, only few respondents reference originality as the driver behind their positive evaluations (see, e.g., the second item in Table 3.1). Most participants do not explicitly mention a source for their relevance, plausibility, or originality estimate. Those who do specify a source make reference to their personal experience (see, e.g., the third item in Table 3.1) or to their own opinion (see, e.g., the last item in Table 3.1).

| Level 1 | Level 2 | Tag Count |
|---|---|---|
| NotAnswered | | 1 |
| reason | originality | 5 |
| | plausibility | 57 |
| | relevance | 75 |
| source | experience | 10 |
| | opinion | 12 |

**Table 3.2:** Tags of Reasons for Positive Ratings

Overall, the vagueness in many responses indicates that practitioners face difficulties pinpointing what exactly causes them to rate certain research summaries as important. This supports the view that survey respondents react intuitively to cues in the text of the one-sentence summaries rather than evaluate consciously and carefully the merits of the research these summaries describe. Against this background, the reasoning questions in the second part of the RE-Pract survey appear to ask practitioners for *post-hoc* rationalizations of gut decisions—with limited success in the case of positive evaluations. Therefore, the only element of an answer to **RQ3** we can defensibly derive from the set of positive rating explanations might be—redundantly—stated as follows:

*Practitioners would like researchers to focus on research problems that practitioners deem important.*

### 3.3.2 Negative Rating Explanations

In the second question belonging to the second part of the RE-Pract survey, respondents are presented with the following task: *Please provide a brief explanation for why you provided one of the lowest ratings to the following piece of research.* Here, responses are available from 117 survey participants for 103 different research summaries. Again, we can use a semi-automatic tagging procedure involving regular expressions to structure and quantify our responses. Table 3.4 (→ p. 57) showcases some answers, along with the summaries they reference and the tags they were assigned, and Table 3.3 (→ p. 56) displays the tag counts.

| Level 1 | Level 2 | Tag Count |
|---|---|---|
| NotAnswered | | 1 |
| reason | notconvincing | 37 |
| | notefficient | 10 |
| | notimportant | 20 |
| | notinteresting | 6 |
| | notoriginal | 4 |
| | notrealistic | 24 |
| | toocomplicated | 5 |
| | toospecialized | 12 |
| | toosubjective | 1 |
| | toovague | 4 |
| rejection | questionnotunderstood | 3 |
| | ratingnotnegative | 8 |

**Table 3.3:** Tags of Reasons for Negative Ratings

As the number of different tags already indicates, the rationales provided for negative ratings are much more diverse than the rationales provided for positive ratings. Interestingly, while perceived relevance is the most-stated reason for positive ratings, it is not perceived irrelevance but rather a perceived lack of plausibility that is mentioned most frequently as a reason for negative ratings. Answers marked with the *notconvincing* and the *notrealistic* tags both fall into this category. However, irrelevance as an explanation is still frequently cited; the *notimportant* and *notrealistic* indicate variants of this rationale. According to our tag counts, overspecialization and inefficiency (especially of solution proposals) rank third and fourth on our respondents' research blacklist. Finally, the non-negligible number of question rejections highlights a problem with the survey administration tool: Even if re-

spondents give positive ratings to all of their research items, they are still asked to provide reasons for one of their (relatively) lowest ratings.

| Summary | Reasoning |
|---|---|
| An experience report on the development of a methodology and tools for the formalization and subsequent validation of specifications in a project for the public administration | no one cares 'how' you developed, just that you have your specifications. there are also too many personalities that no one would use it anyway<br><br>*reason:notconvincing; reason:notrealistic; reason:toospecialized* |
| An online-survey on factors that prevent business analysts from applying their requirements analysis knowledge in practice. | As a practioner, I prefer learning about the answers than learning about the problems we all know we have.<br><br>*reason:notinteresting* |
| An experiment with students for investigating the relationship between time pressure and efficiency in test case development and requirement review | Students do not yey have the skills to contribuye meaningfully.<br><br>*reason:notrealistic* |
| A solution for determining the quality of trace links and detecting unacceptable deviations in order to support the systematic assessment of a project's traceability. | Traceability is not a high-priority activity in Requirements Engineering.<br><br>*reason:notimportant* |
| A modeling language for representing and analysing requirements for Self-Adaptive Systems in order to make them readable by non-engineering stakeholders | In my experience, most modeling language research never goes beyond the PhD lab. It may be valuable eventually, but this needs application before asserting its utility.<br>*reason:notrealistic* |
| A method for identifying single- and multi-word terms that have a particular significance in a given domain in order to characterize the most salient features of the document in which they appear | I'm not sure how I would apply this research as I don't understand what the outcome would be.<br><br>*reason:toocomplicated* |
| A method for automatically deriving conceptual data models from user stories written in natural language in order to create a holistic view of the requirements specification, | The research is too academy oriented, it mixes so many aspects, I doubt that it can lead to practical results, other than the ones that can be obtained in a controlled environment with a very reduced set of requirements.<br>*reason:notrealistic* |

**Table 3.4:** Examples of Reasons for Negative Ratings

From the patterns observed in the explanations practitioners provided for negative ratings, it is difficult to augment our answer to **RQ3**. First, most rationales that do not quote irrelevance refer to solutions rather than to problems. Second, as with the reasoning for positive ratings, practitioners are again asked for a *post-hoc* rationalization of their intuitive decisions. Thus, although the *post-hoc* procedure works better here than for positive ratings, the only guideline for RE researchers we might identify in our respondents' explanations for their negative ratings is:

*Whatever research problems researchers focus on, their solutions had better survive a reality check.*

### 3.3.3 Research Wishes

So far, we have only approached **RQ3**—*What research problems do practitioners think are most important to be focused on by the RE research community?*—indirectly, leveraging practitioners' rationales for particularly positive and negative ratings to uncover their research preferences. In this section, we tackle the question directly, sourcing information from the answers practitioners provided to the last question in the second part of the survey's core: *Suppose that you could provide guidance to a team of requirements engineering researchers. What topics / problems should they focus on first to support you in your primary work area?* 103 of our 154 respondents entered one or more characters to answer this question, although again, some answers contain no information (e.g., ".", "-", and "No").

Although responses vary greatly in content, style, and volume, their overall verbosity highlights that many RE practitioners are eager to share ideas with the RE research community. From the feedback collected, five research wish themes appear particularly popular:

— *practical RE research*, closing what is perceived as a gap between RE research and RE practice;

— *guiding RE research*, pointing to solutions directly applicable in RE practice;

— *research on agile RE*, tackling the perceived difficulties when traditional RE meets agile or other non-waterfall practices;

— *research on human factors in RE*, addressing classical challenges of human communication and collaboration; and

— *research on requirements management*, especially supporting requirements verification, validation, and evolution.

Table 3.5 and Table 3.6 (→ pp. 59–60) show selected answers referring to these topics, along with respondents' demographic data.

| **Practical RE Research** |
|---|
| I think that the community is to focused in producing short term results for scientific publications. In general, experiments are fictitious and irrelevant, and conducted in very controlled environments. I think that research shall move towards conducting large experiments in the industry and present the results of the application of methods and notations in more real non-controlled environments. I currently see the RE community as a very closed family in which novel ideas are not welcomed. Same authors on same topics year after year, with very little chance for real experiences to be presented.<br>*— business analyst, 15 years experience, ICT sector, Ecuador (larger team, hybrid systems)* |
| The main issue is to make research around practical methods and techniques for doing requirements elicitation, requirements specification, and requirements validation. Focus within Software Engineering (Requirements Engineering) is on a level that is far to high and abstract (frameworks, philosophy, theory). The industry needs a toolbox with techniques and practices that work. It is often about the small things: how to conduct an interview with a customer; where to find raw information (organizational charts, service descriptions, product documentation); how to validate requirements (through documents or through triangulation); how to document the results of a customer meeting (Minutes of Meeting). The industry is completely void when it comes to having a toolbox with techniques and methods such as these. It is too much theory and high-level talk. You have all these techniques in research (this questionnaire, open-ended interviews, structured interviews, observations), but the industry use post-its.<br>*— developer, 22 years experience, aeronautics sector, Sweden (small team, [business] information systems)* |
| They should first understand the context of my work area. It seems to me (just a feeling) that a lot of the work on requirements comes from a context where legal aspects are important (e.g. aviation) and/or where the development model is inspired by a mechanical history (e.g. cars). In other words not relevant to a lot (maybe most?) of software development done in the world.<br>*— tester / test manager, 13 years experience, ICT sector, Sweden (very large team, software-intensive embedded systems)* |
| **Guiding RE Research** |
| Provide some guidelines and set of best practices that can/should be applied in projects with different dimensions.<br>*— tester / test manager, 8 years experience, ICT sector, Portugal (medium-sized team, [business] information systems)* |
| - advanced methods - Problem solving methods - complex methods - methods for complex products<br>*— requirements engineer, 14 years experience, automotive sector, Germany (larger team, software-intensive embedded systems)* |
| Standard documentation, communication methods, consistent releases<br>*— business analyst, 11 years experience, public sector, Canada (medium-sized team, hybrid systems)* |

**Table 3.5:** Examples of Research Wishes (Part 1)

| **Research on Agile RE** |
|---|
| Integration of your techniques and models with Agile development methodologies and Agile architecture methods and tools. In the real-world, teams are organized in small agile teams orchestrated by architecture work. At both levels requirements are very relevant, but the tools used to specify and validate them IMHO are very diferent to what classical academic research on RE is concerned with. Therefore, analyse those methods and tools and try to provide solutions that are realistic to their problems; as I said, in agile development and agile architecture methods and tools.<br>*— architect, 10 years experience, tourism sector, Spain (small team, [business] information systems)* |
| Main focus of my work is on working in an agile development environment but having a big set of requirement requests from the product requester and sponsor. Managing this difficult situation between agile development and waterfall-like requirement setting is key for my work.<br>*— product owner, 5 years experience, infrastructure sector, Germany (medium-sized team, [business] information systems)* |
| **Research on Human Factors in RE** |
| I think one of the main areas of concern is hearing the voice of the user, overcoming the problem of the user's failing to adequately articulate their needs. This would include interdisciplinary research in the fields of cognitive psychology, decision analysis (e.g., Decision Trees), software design (problem definition).<br>*— project manager, 15 years experience, energy sector, United States (small team, [business] information systems)* |
| I think it is important that how to integrate the requirements of complicated stakeholders. And it is also important to derive real demands from statements of stakeholders. In recent years, production technology (how to make a system) has developed and achieved high productivity, but we still trial and error about what to make .<br>*— requirements engineer, 8 years experience, multiple sectors, Japan (medium-sized team, [business] information systems)* |
| First of all I'll ask a factive way to meet both client's requests and dev team's requests. Requirements are often written in a business oriented view only, but a good requirement engineer has to keep in mind that he/she has to interface also with a dev team, an architect and some testers. So, first of all we have to find a factive way to present requirements for all groups in a clear and consistent manner. Secondly, we are people that work together. Psicology and communication styles are really important.<br>*— tester / test manager, 4 years experience, e-commerce sector, Italy (larger team, [business] information systems)* |
| **Research on Requirements Management** |
| Focus on Change and Version Control (Configuration Management) associated with changing requirements. Modifications to requirements can have an enormous impact on the design, verification, cost and schedule aspects of a project and this impact is often not adequately communicated because of poor change and version control processes.<br>*— requirements engineer, 20 years experience, ICT sector, Australia (very large team, hybrid systems)* |
| How to validate requirements (or specifications, prototypes, ...) with actual users to understand/validate the actual value they would have. Many requirements are implemented *before* getting feedback on their actual value meaning resources are wasted.<br>*— manager, 5 years experience, software sector, France (medium-sized team, [business] information systems)* |

 **Table 3.6:** Examples of Research Wishes (Part 2)

In the light of respondents' answers to the last question in the second part of the RE-Pract survey, we can summarize our answer to **RQ3**, *What research problems do practitioners think are most important to be focused on by the RE research community?*, as follows:

*Generally, RE practitioners would like RE researchers to give guidance on how to solve the most pressing problems of RE practice. In particular, they would like RE researchers to focus on the problems of agile RE, human factors in RE, and requirements management (including verification, validation, and evolution).*

Beyond **RQ3**, the wording of practitioners' research wishes hints at how they see the relationship between RE research and RE practice. Most testimonials, despite their diversity, seem to share the assumption that RE research can support RE practice primarily by offering solutions (e.g., methods, tools, or technologies). In part, this might be due to the wording of the survey question, which explicitly asks what researchers should focus on *to support* respondents in their primary work area. But our respondents' formulations might also be reflective of an implicit consensus amongst RE practitioners that RE research *serves*, rather than *observes*, RE practice.[18] Thus, one last insight from our respondents' research wishes might be summarized as follows:

*RE practitioners expect RE researchers to offer solutions for their problems.*
*They want engineering, not science.*

18 This view of the relationship between RE research and RE practice would be remarkably similar to a widespread view of the relationship between legal research and legal practice. An in-depth investigation of the parallels between RE and law, however, lies beyond the scope of this thesis.

# 4 Discussion: Perceiving the Gap

In this chapter, we gauge the implications of our results for the RE research community. Having sketched both the supply side and the demand side of RE research in the previous chapters, we start our assessment with a comparison between the two, thereby constructing what we call the *apparent gap* between RE research and RE practice (→ 4.1). We continue by inspecting our research approach for sources of potential errors, highlighting that our apparent gap is also a *fragile gap* (→ 4.2). Finally, we scrutinize the consequences of the identified error sources for the interpretation of our results, arriving at a more nuanced account. This leaves us with an opportunity for further research: the *uncharted gap* between RE research and RE practice (→ 4.3).

## 4.1 The Apparent Gap: Comparing Supply and Demand

RE is a heterogeneous, interdisciplinary field. Therefore, RE research supply and RE research demand may appear to match or mismatch in many ways, depending on our perspective. In the following, we highlight five aspects that might lead us to perceive a gap between RE research supply and demand, based on our results from the previous chapters.

In our inquiry into the RE research supply side in chapter 2, we map seven years of RE research as presented at the field's most prominent conferences, assigning tags to one-sentence summaries of 435 full papers to characterize their methods and contents. We find, *inter alia*, that (1) a large fraction of the examined research proposes methodologies (→ Figure 2.7, p. 17), and that (2) many research items address questions of requirements management, whereas (3) comparatively few research items tackle the challenges of RE in agile environments (→ Figure 2.9, p. 19). Furthermore, our sample comprises (4) similar fractions of interrogation and intervention research, and (5) roughly equal fractions of studies carried out with practitioners and with students.

On the RE research demand side, which we investigate in chapter 3 using the data collected in the RE-Pract survey, we observe (1) practitioners' calls for more guidance from research that addresses practical problems by suggesting plausible solutions (→ Tables 3.3–3.5, pp. 56, 57, 59). In their research wishes, practitioners express their perception of both (2) a need for more research on aspects of requirements management and (3) a need for more research on agile RE (→ Figure 3.6, p. 60). From the paper ratings, we also discern (4) a preference for interrogation research over intervention research and (5) a preference for research with practitioners over research with students (→ Figures 3.24–3.27, pp. 40–41). Table 4.1 (→ p. 64) contrasts our supply-side with our demand-side observations.

| | **RE Research Supply** | **RE Research Demand** |
|---|---|---|
| 1 | Much research proposing methodologies | Calls for more practical guidance from RE research |
| 2 | Much research on aspects of requirements management | Calls for more research on aspects of requirements management |
| 3 | Little research on agile RE | Calls for more research on agile RE |
| 4 | Comparable amounts of interrogation and intervention research | Preference for interrogation research over intervention research |
| 5 | Comparable amounts of research with practitioners and with students | Preference for research with practitioners over research with students |

**Table 4.1:** RE Research Supply vs. RE Research Demand

Upon closer inspection, there appear to be two types of gap between RE research supply and RE research demand, which might also occur cumulatively. The first type of gap is the *quantity gap*. The quantity gap is characterized by disagreements between research and practice regarding the sheer *amount* of research needed. Here, disagreements can be both *absolute* (e.g., little supply of research on a certain topic meets high demand by practitioners; item 3 in Table 4.1) and *relative* (e.g., same supply sizes for research on two topics meet different demands [as indicated by ratings or rationales]; items 4 and 5 in Table 4.1).

The second type of gap between RE research supply and RE research demand is the *quality gap*. The quality gap is defined by dissension between research and practice regarding the *kind* of research needed. Here, disagreements may concern both the *core* of the research process, namely, how problems are identified, approached, and—ideally—solved, and the *communication* of the results from the research process. The quality gap might be rooted in the differing socializations (and, correspondingly, experience horizons) of researchers and

practitioners: The research community tends to reward novel or elegant solutions while practitioners often need to "just make things work;" researchers cherish deliberations while practitioners long for directions.

In our case, observations 1 and 2 in Table 4.1 (→ p. 64) might be interpreted as examples of the quality gap: There is a lot of research proposing methodologies but—apparently—the solutions frequently fail to address practitioners' needs (observation 1). Similarly, despite the large amount of research on aspects of requirements management, practitioners still wish for more, which suggests that RE research frequently works past the needs of RE practitioners (observation 2)—especially when read in conjunction with observation 1.

As the discussion above shows, it is at least possible to interpret our results as pointing towards a gap between the supply side and the demand side of RE research. More precisely, we can identify not one but several gaps, both quantitative and qualitative, between what RE practitioners seek and find: our *apparent gap* is really a landscape of lacunae. However, our results are based mostly on descriptive statistics for data generated in the context of the RE-Pract survey. To assess the validity of our results, we thus need to confront ourselves with the sources of error in our data generation process.

## 4.2 The Fragile Gap: Exposing Sources of Error

At first glance, the counts and percentages we present in chapters 2 and 3 might emanate objectivity due to their nature as numbers. In reality, however, these numbers are only as objective as the process underlying their generation. Here, we therefore inspect the RE-Pract data generation process for sources of error that could have impacted our results.

Figure 4.1 (→ p. 66) gives an overview of the RE-Pract data generation process. Three sources of subjectivity are identified (*blue ellipses*): the RE-Pract team, the respondent, and the author of this thesis. Furthermore, only three information items are associated with a "ground truth" that could enable (somewhat) objective verification (*green rectangles*): the individual RE research paper, its metadata, and the metadata describing the respondent. The rest of the information items involved in the process are all subjective constructions (*gray hexagons*). They are created by humans (*bold arrows with numbers*) on the basis of data inputs (*dashed arrows*) and can only be validated on the basis of plausibility or inter-subjective agreement (*thin arrows*). As the graphic shows, our statistical analyses are based on inputs

from five different sources (*yellow stars*). Some ground truth exists for only two of these five sources (paper metadata and respondent metadata), and only two of the sources (ratings and respondent metadata) are directly derived from our survey observables (*bold frames*).

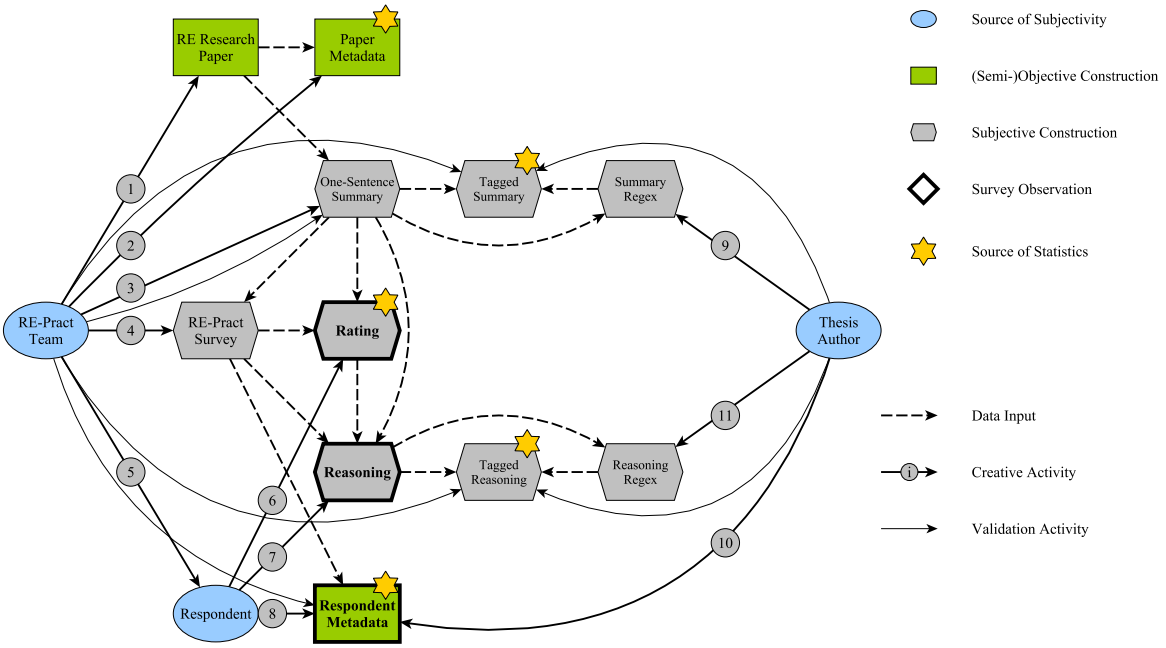



**Figure 4.1:** Overview of the RE-Pract Data Generation Process

The potential for errors in the RE-Pract data generation process can be estimated by analyzing the arrows that lead to our sources of statistics. While the dashed input arrows only represent the propagation of errors introduced in previous steps, the bold, numbered arrows stand for creative human activities (emanating from sources of subjectivity) that might act as original sources of errors. The graphic identifies eleven such error sources, numbered by the approximate order in which they occur in the RE-Pract research pipeline. Five error sources result from the undertakings of the RE-Pract team, three are introduced by respondents' activities, and three stem from efforts made by the author of this thesis. In the following, we describe each of the creative tasks associated with our error sources and delineate the error potentials—and potential biases—they imply.

1. *Select papers to be included in the sample.*

   The sample of RE research papers to be included in the RE-Pract survey is defined by the survey authors. It is based on a combination of different criteria whose application is not always straightforward: the paper must be a *full paper* having *some connection to RE* and published at a certain conference in a certain year. Therefore, inconsisten-

cies in the application of the selection criteria may have distorted our results from the start. Further, the criteria themselves reflect value judgments by the RE-Pract survey authors. For example, had the authors decided to include short conference papers and journal articles in their sample of RE research, our results would have probably been quite different.

2. *Collect and code paper metadata.*

   The metadata associated with the papers included in our sample is collected by the RE-Pract team, which might make mistakes in the collection process. Moreover, some of the variables that are recorded, such as an author's affiliation (*academic*, *industry*, or *mixed*), require further research and human judgment. This might lead to additional mistakes in our data, e.g., when the categorization of an institution as *academic* or *industry* is unclear or when an author is classified as having academic affiliations only when in reality, an industry affiliation is also present. Such errors are realistic especially since academic institutions have very different attitudes towards their academic personnel's industry connections. This could cause some authors to emphasize their academic affiliations in their publications, concealing their industry affiliations by omission.

3. *Draft one-sentence paper summaries for all papers in the sample.*

   In the RE-Pract survey, practitioners are asked to state how they perceive the importance of RE research papers *as represented by their one-sentence summaries*. These summaries are drafted by the authors of the RE-Pract survey. Although they are eager to use consistent terminology, these authors do not use a formally controlled vocabulary. Furthermore, they base their one-sentence summaries mostly on the papers' abstracts, rather than on an in-depth reading of the papers themselves. While pragmatic given the resource constraints of the RE-Pract survey, the one-sentence summaries produced might fall victim to three sources of inaccuracy: first, the paper authors' creativity in formulating the abstract (which often functions as an advertisement rather than as a statement of fact); second, the survey authors' interpretation of the abstract (which might be influenced by their own perspectives on RE); and third, terminological inconsistencies in the summary representations introduced by the survey authors (due to the lack of a controlled vocabulary).

4. *Design RE-Pract survey questions.*

   One of the most important steps in the RE-Pract survey is the design of the survey questions, again effected by the RE-Pract team. Here, each part of the RE-Pract survey demands different design decisions, many of which can drastically alter survey results. To address a particularly problematic example: If practitioners are to rate one-sentence summaries, the rating scale is a critical choice. The authors of the RE-Pract survey opt for a Likert scale with the ordered categories `Essential`, `Worthwhile`, `Unimportant`, and `Unwise`, following similar approaches in the replicated studies [14, 4].

   While this choice facilitates comparisons with results from prior work, it is questionable for at least two reasons. First, the labels of the Likert scale items refer to concepts that vary in more than one dimension. For example, the negative connotation of the label `Unwise` is categorically different from that of `Unimportant`: Lacks of wisdom are (directly or indirectly) attributed to people, lacks of importance are attributed to matters; wisdom refers to a largely static property, importance refers to a largely dynamic state. Second, the four-item scale eliminates the possibility to state an absence of opinion without negative connotation, which might lead respondents to rate research as `Worthwhile` despite genuine indifference. Thus, a rating scale eliciting more accurate results might have been `Very Important`, `Somewhat Important`, `Neither Important Nor Unimportant`, `Somewhat Unimportant`, and `Very Unimportant`. Unfortunately, the rating scale adopted in the RE-Pract survey, which is omnipresent in our statistical analyses, could have reduced the accuracy of our survey results: The perceptions of importance expressed by our respondents are probably more positive than their true perceptions.

5. *Select and invite practitioners to participate in the survey.*

   The last error source in the RE-Pract data generation process that is directly introduced by activities of the RE-Pract team is the selection of practitioners to ask for participation in the survey. This selection is mostly based on lists of personal contacts compiled by the survey authors, and survey invitations are sent out with the aim to control response rates and sample composition [5]. Thus, due to *convenience sampling*, there is a natural—even intended—selection bias in our sample of invitees, which leads to a sample of respondents that is heavily skewed towards German RE practitioners from

certain industries. This skew is likely aggravated by the survey authors' decision to recruit additional participants via a mailing list of the German *IREB e.V.* (→ section 3.1.1). Therefore, the selection process used in the RE-Pract survey further reduces the representativity of our results. However, these results might still lend support to existing hypotheses regarding practitioners' perceptions of RE research, and they might serve as a basis for the development or refinement of theories to be investigated in future research.

6. *Assign ratings to one-sentence summaries.*

   RE practitioners invited to participate in the RE-Pract survey introduce three sources of error in the data generation process, one in each part of the survey. The first of these error sources lies in practitioners' ratings of the one-sentence summaries. Notably, practitioners rate neither the RE research papers included in the sample nor the meaning of the one-sentence summaries as understood by the RE-Pract team. Rather, they rate their interpretation of the one-sentence summaries, which might be very different from the interpretation expected by the summary authors. Since the sentences presented to practitioners are often lengthy, we can assume—as discussed in chapter 3—that respondents base their research ratings on an intuitive reaction to verbal cues concerning a paper's methods or contents rather than on a thorough understanding of the paper summary. Furthermore, respondents must express their importance perceptions on the four-item (not-so-)Likert scale previously identified as problematic, and it is not far-fetched to suppose that the standards used to assign labels to intuitions vary amongst respondents. Since heterogeneity in these standards might well result from heterogeneity in respondents' socialization or cultural norms, ratings by RE practitioners with very different backgrounds might not be as comparable as we assume for the purposes of our statistics.

7. *Explain ratings for one-sentence summaries and provide guidance for RE researchers.*

   The second error source in the RE-Pract data generation process that is directly connected to our respondents is rooted in the reasoning tasks these respondents are given. Practitioners are asked to give explanations for one particularly positive and one particularly negative rating—but only on a new survey page and only after they have finished the ratings section of the survey. Thus, while they might have assigned their

ratings intuitively, they are not asked to reflect on them until after the fact, which begs for potentially biased *post-hoc* rationalization.

Additionally, answering the question on research guidance involves much more (cognitive and physical) effort than the multiple choice questions practitioners start with. This might cause busy (or lazy) practitioners to skip the question or answer only very briefly, which creates the risk that our responses to the research guidance question are dominated by people who feel strongly about certain topics or people who happen to have enough time to think carefully about their answer. Therefore, it cannot be excluded that we fail to detect the priorities of a "silent majority." In fact, we cannot even assume that the individual responses to the guidance question are representative of the individual respondent's priorities: When working through a survey, humans in a hurry tend to forget even problems they usually deem important unless they are primed to remember them, e.g., by a list of multiple-choice options.

8. *Share personal background information.*

   The third and last error source introduced by respondents concerns the demographic information they are asked to share towards the end of the survey. Although it is unlikely that respondents answer deliberately untruthfully, they might well make errors entering numbers (e.g., do we really have one respondent with more than $50$ years of experience?) or give objectively inaccurate answers (e.g., when respondents working for a firm producing mostly civil aircrafts enter "Defense" as the sector they work in). Respondents could also misunderstand the question, which might lead to accidentally untruthful or internally inconsistent answers (as in the case of the question on involvement in RE). If many respondents misunderstand a question, however, the dominant error source is likely the question design (see $4.$ above).

9. *Design regular expressions for paper summaries.*

   To analyze the data collected in the RE-Pract survey, the author of this thesis designs regular expressions to tag the one-sentence summaries produced by the RE-Pract team with terms characterizing paper contents and methods. This decision, while necessary to enable data analysis in the present context, introduces an additional source of error in the RE-Pract data generation process. The regular expressions are designed after manual inspection of the one-sentence summaries, partially reverse-engineering the structure originally put in by the summary authors. The expressions are likely too

broad in some ways and too narrow in others, and the taxonomy underlying their construction is also debatable (and, at the point of writing, still debated within the RE-Pract team). To put it bluntly, the work invested in taxonomy building and paper tagging would have been better placed at the start of the data generation process, right before the design of the one-sentence summaries. That way, the tags could have served as a controlled vocabulary in the formulation of the summary sentences. This could have improved the consistency of the tag assignments—and it might have made the statistics using the tagging data more accurate.

10. *Code respondent metadata.*

    By design (see 4. above), many demographics questions in the survey expect free-text answers where other surveys use numeric fields (e.g., *years of experience*) or multiple-choice options (e.g., *respondent sector*). Also, even multiple-choice questions usually include an option *Other (please specify)*, which allows respondents to provide a free-text answer in case they find all proposed answers unfitting. To enable the aggregation of our rating data by respondent metadata features, the free-text answers must be coded.[1] This task is performed manually by the author of this thesis, with partial support by regular expressions. The coding activity introduces a source of error mainly for three reasons. First, the coding scheme developed might be flawed; second, the coding scheme might be applied inconsistently by the thesis author; and third, the coding of free-text answers elaborating on the choice *Other (please specify)* in a multiple-choice question might lead to the addition of categories that would have been selected by other respondents had they been available in the survey.

11. *Design regular expressions for respondent reasoning.*

    Finally, we aggregate not only our rating data but also the reasoning data available from respondents' explanations of particularly positive and particularly negative ratings. Here again, the author of this thesis crafts a tagging scheme following the manual perusal of all responses and designs regular expressions for applying the tagging scheme to the available answers. This yields a further source of error for reasons similar to the ones explained in the context of the paper tagging activity (see 9. above). When tagging the answers to the reasoning questions, the fact that each response

1 Here, *coding* refers to the representation of free-text answers by at least one of several categories as it is commonly practiced (e.g., by social scientists) in qualitative data analysis.

comes from a different person leads to complications not present in the tagging of the one-sentence summaries (which are all written by the RE-Pract team). However, since the tagging scheme used is much coarser than in the paper mapping, the overall impact of inaccuracies in the tagging procedure is likely smaller here than in the case of the one-sentence summaries.

For an assessment of the overall error potential affecting our sources of statistics, we might return to Figure 4.1 (→ p. 66). To estimate the degree of subjective judgment underlying each of our apparently objective arrays of numbers, we identify and count the labeled arrows leading—directly or indirectly—to our sources of statistics. Table 4.2 (→ p. 72) summarizes the results.

| **Source of Statistics** | **Error Sources** | $\Sigma$ |
|---|---|---|
| Paper Metadata | $1, 2$ | $2$ |
| Rating Data | $1, 3, 4, 5, 6$ | $5$ |
| Respondent Metadata | $4, 5, 8, 10$ | $4$ |
| Tagged Summary Data | $1, 3, 9$ | $3$ |
| Tagged Reasoning Data | $1, 3, 4, 5, 6, 7, 11$ | $7$ |

**Table 4.2:** Error Sources of RE-Pract Statistics

Although not all sources of error are made equal, the sheer count of judgments involved in the generation of our data sources is daunting, and the error sources listed might well have impacted our sources of statistics in different ways. If we further consider that in some of our analyses, we aggregate our rating data ($5$ error sources) by features in our respondent metadata ($4$ error sources), our paper metadata ($2$ error sources), our tagged summary data ($3$ error sources), or even our respondent metadata *and* our tagged summary data ($4$ *and* $3$ error sources)—all of that given an already small number of ratings and respondents—, it seems unavoidable to conclude that our *apparent gap* identified in the previous section is really a *fragile gap*: We can construct it from the RE-Pract data but it could disappear, or look very different, if we aggregated our data differently. Thus, while our data *suggests* that there is a gap between RE research demand and RE research supply, that RE research could do more to cater to the needs of RE practice, the results of the RE-Pract survey should be taken as an input to, rather than an endpoint of, the discussion about the practical relevance of RE research and the proper relationship between RE research and RE practice.

## 4.3 The Uncharted Gap: Qualifying the Survey Results

So far, we have analyzed the potential causes of distortions in our statistics from a micro-level perspective, evaluating the subjectivity involved in the individual steps of our research process. Now, we adopt a macro-level perspective and scrutinize the potential effects of such distortions on our perception of what we have already identified as a *fragile gap* between RE research and RE practice. More precisely, we seek to estimate the validity of the RE-Pract survey, i.e., the *trustworthiness* of our results.

We frame our discussion using the categorization of validity types suggested for case studies by Wohlin et al. in [20], which we adapt to the survey research at hand. In their description of case studies, the authors distinguish between four types of validity: *construct validity*, *internal validity*, *external validity*, and *reliability* [20, pp. 68–69].[2] In the following, we define each of these validity types in the context of survey research and assess to what extend the error sources in our data generation process might have impaired the validity of our results.[3]

1. *Construct Validity.*

   In the context of case studies, construct validity is concerned with the relationship between *observation* and *theory*: Does what is measured really reflect what researchers think and what they claim to investigate in their research questions?

   This definition can also be used to evaluate survey research: Does what is measured by the survey really reflect what researchers intend to measure? In the case of the RE-Pract survey, the answer is likely: not entirely.

   The arguments in support of this view are manifold, and many of them follow from the discussions in previous sections. Two constraints—discussed by Wohlin et al. in the context of experimental studies [20, pp. 108–109]—seem to be particularly responsible for decreasing the construct validity of our research. The first constraint is *inadequate preoperational explication of constructs.* This constraint results from the finding that the constructs of interest are insufficiently defined before they are operationalized. In our

2 This classification is similar to but not identical with the classification offered for experimental studies in [20, pp. 102–112], where *reliability* is replaced with *conclusion validity* and the meaning ascribed to each of the categories is tailored to the experimental setting of treatments and outcomes.

3 All descriptions of validity categories presented below that explicitly reference the context of case studies are based on [20].

case, the most problematic constructs are *perceived relevance* and *research idea*. Eager to replicate previous work in an only slightly amended setting [5], the authors of the RE-Pract survey seem to have skipped the definition of these constructs altogether. In fact, they forward the construct definition task to respondents by asking them for an assessment of research *importance*—where the interpretation of the term is left to respondents and the semantic relationship between perceived importance and perceived relevance remains unclear—on the basis of the *one-sentence summary* of a research paper.

The second constraint that impairs the construct validity of our research is *mono-method bias*. This constraint results from the use of a single type of measure that gives a measurement bias. As discussed in previous sections, most questions in the RE-Pract survey ask practitioners to rate the importance of one-sentence research summaries on a four-item Likert scale. The design of this scale, with its labels `Essential`, `Worthwhile`, `Unimportant`, and `Unwise`, makes the core of the survey prone to measurement bias. The resulting validity threat is partially mitigated by the reasoning questions following the ratings section, but answers to these questions are often too fluffy and too few for the bias potential to be offset.

2. *Internal Validity.*

   In the context of case studies, internal validity is concerned with the type of relationship between *observations* and *non-observations*: If researchers investigate causal relationships, are there potentially unmeasured confounding factors?

   This definition applies directly only to explanatory survey research, i.e., survey research seeking causal relationships. But it can easily be adapted to fit descriptive and exploratory survey research: When painting a picture of the domain under study, do researchers include all potentially relevant factors? In the case of the RE-Pract survey, we again tend to answer: no.

   Here, the primary problem lies with the design of the one-sentence paper summaries. Although one intention of the survey authors is to extract the most highly rated research ideas, due to the missing preoperational explication of the idea construct discussed above, they draft the sentences to include verbal cues concerning research methods as well as research contents (and sometimes even research goals, e.g., using the phrase “in order to”). Thereby, they effectively conflate some distinctions that might be necessary to paint even a preliminary picture of the RE research landscape

and its perception through the eyes of our respondents. Our research paper mapping attempts to disentangle some of these factors but it cannot compensate for the lack of direct observations. Thus, most of our more detailed analyses are based on shaky empirical grounds. This especially challenges our more concrete findings, and it limits the internal validity of our research.

3. *External Validity.*

   In the context of case studies, external validity is concerned with the relationship between the *sample* and the *population of interest*: To what extend can the findings of the research be generalized?

   This definition is directly applicable to survey research: To what extend can the survey results be generalized? In the case of the RE-Pract survey, we must admit: to a very limited extend.

   The primary reason for the lack of generalizability is the *selection bias* resulting from the procedure used to recruit survey participants, which is also evident in the statistics presented in chapter 3. As discussed previously, the RE-Pract survey is designed as an invitation-only online survey. Here, a first layer of selection bias is introduced by the decision to select most invitees from a list of personal contacts compiled by the survey authors. A second layer of selection bias is added by the resolution to choose from the list based on explicit considerations regarding response rate control and sample composition. Finally, a third layer of selection bias stems from social factors in the sphere of the invitees: Since survey participation is not only voluntary but also costly, it is likely that the set of invitees who complete the survey is systematically different from the set of invitees who do not complete the survey: Respondents are likely to have a closer relationship with the survey distributors (directly or indirectly, e.g., through their supervisors or colleagues), a more favorable attitude towards survey research, or simply more time than invited non-respondents. Since our sample, i.e., invited RE practitioners who complete the survey, is hardly representative of the target population, i.e., RE practitioners around the globe, our results are hardly generalizable. Thus, the external validity of our results is further constrained.

4. *Reliability.*

   In the context of case studies, reliability is concerned with the relationship between the *observation* and the *observer*: If other researchers conducted the same study, would the results be the same?

   This definition is also suitable for survey research, although the guiding question might be phrased more precisely as follows: If other researchers conducted the same survey *or* analyzed the same data, would the results be the largely same? In the case of the RE-Pract survey, the response is: probably not.

   More precisely, we identify two sets of reliability constraints affecting our research. The first set consists of problems associated with the subjectivities highlighted in Figure 4.1 (→ p. 66). For example, as far as survey setup and administration are concerned, other researchers might select a different sample of papers to include in the rating questions, write the one-sentence summaries differently, or recruit a different crowd of respondents. In the realm of data analysis, other researchers might craft a different taxonomy or design different regular expressions to tag the one-sentence summaries and the responses to the reasoning questions, or they might take different decisions when coding the respondent metadata.

   The second set of reliability constraints comprises mostly aspects that are commonly discussed under the heading of *conclusion validity* in the context of experimental research [20, pp. 105–106]. This includes an overall *low number of observations*, which leads us to observe patterns that are based on "small $n$" and thus likely artifacts of statistical fluctuations; the limited *reliability of measures* due to our flawed Likert scale and the subjectivity involved in the coding of free-text answers; as well as *random irrelevancies in the response setting*, i.e., factors in our respondents' environments (e.g., in their office) that might keep them from filling out the survey truthfully. Perhaps most importantly, the second set of reliability constraints also includes *fishing*: Both in the formulation of the survey questions and in the analysis of the survey data, the RE-Pract team or the author of this thesis might—consciously or unconsciously—search for specific results. This is especially evident in the choice of features to aggregate on in chapter 3: Because the possible combinations are too numerous to investigate all of them, we focus our attention on the factors deemed interesting by the researchers

involved in the survey. Since we can only find what we seek, this procedure—though hardly avoidable in exploratory studies—further limits the reliability of our research.

In summary: From our analysis of the data generated in the context of the RE-Pract survey, there is an *apparent gap* between RE research and RE practice, between RE research supply and RE research demand. The statistics underlying our analysis, however, are the product of many subjective decisions which, if taken differently, might have altered our results. Thus, our *apparent gap* is also a *fragile gap*. Finally, our validity assessment shows that the conclusions we may draw from our results are limited because our research suffers from numerous validity constraints; in particular, most concrete findings are likely invalid. Therefore, we are left with some data supporting the proposition that there is a cleft between RE research supply and demand, but we have little information on the shape of that cleft. What remains is the *uncharted gap* between RE research and RE practice.

# 5 Conclusion: Bridging the Gap

This thesis has investigated the relationship between RE research and RE practice as a relationship between RE research supply and RE research demand. In chapter 2, we have addressed the supply side, producing a map of RE research as presented at the field's most prominent conferences over a period of seven years. Chapter 3 has inquired into the demand side, analyzing the data collected from RE practitioners by means of the RE-Pract survey. This has allowed us to examine, in chapter 4, the gap between RE research supply and demand, which we have characterized as *apparent* (from our data), *fragile* (due to our data generation process), and *uncharted* (due to the validity constraints affecting our research). Our data suggests, *inter alia*, that practitioners value research which does not involve students, does not concern goal models or people's skills, and may address scenarios, architectural requirements, safety requirements, or requirements consistency. The limitations of our study, however, might question the value of our results.

*What does this mean for the RE research community?*

First, we need further inquiries into the practical relevance of RE research before deriving desiderata for the RE research community. The RE-Pract survey, just like any other *single* empirical study, can only contribute to the body of evidence resulting from a research program that practices RE research as an *empirical*, rather than *normative*, discipline—a research program as laid out for SE by Kitchenham, Dyba, and Jørgensen in [12] and translated to RE by initiatives such as NaPiRE [15, 16]. In this context, the propositions developed from our data in chapter 3 might serve as starting points for additional explorations.

Second, future work on the practical relevance of RE research should attempt to minimize threats to validity, rather than replicate prior work that suffers from severe validity constraints. If research paper summaries are to be rated, the use of a more balanced Likert scale that includes a neutral option seems advisable, and paper summaries should be constructed using a controlled vocabulary based on a preestablished taxonomy of RE research methods

and contents. To isolate the causes of certain ratings, it might be expedient to present summaries of hypothetical, rather than real, research items. These items could be constructed from the RE research taxonomy to contain only a certain number of clearly distinguishable cues, which could help RE researchers disentangle the reasons behind practitioners' rating decisions.

Third, we might want to study not only the practical relevance of RE research but also, on a more general level, the adequate relationship between RE research and RE practice. Should RE research seek to be relevant for RE practice—as the reader might have implied—and if so, to what extend? Does RE research exist to *serve* or to *observe* RE practice? How can the RE research community collectively strike the balance between the *theoretical*, the *empirical*, and the *normative*, between *perspective*, *science*, and *engineering*?

The data generated in the context of the RE-Pract survey provides input for discussions addressing all of these questions. Therefore, we end with a graphical summary of the RE-Pract survey (→ p. 80), which allows us to conclude on a positive note:

*The debate on researching the right thing is still in full swing.*

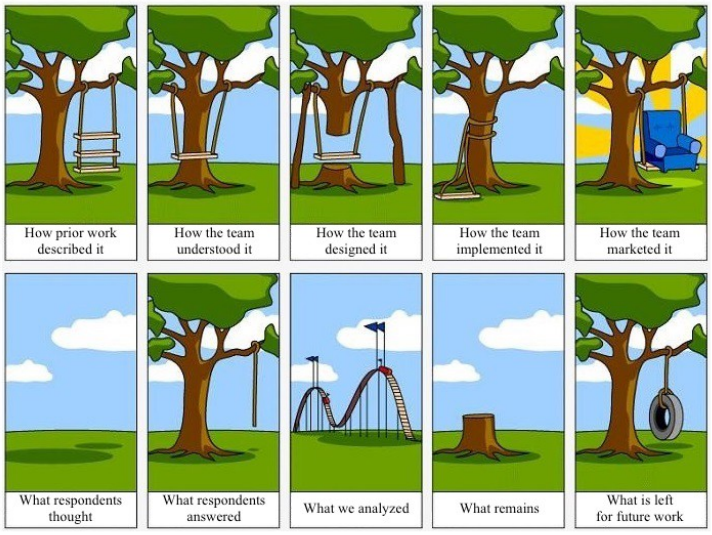



**Figure 5.1:** Graphical Summary[1]

1 Adapted from a picture series found at: https://medium.com/omarelgabrys-blog/requirements-engineering-introduction-part-1-6d49001526d3.

# Bibliography


[1] Mikio Aoyama, Takako Nakatani, and Shinobu Saito. "REBOK Manifest: Towards a Requirements Engineering Body of Knowledge". In: *18th International Requirements Engineering Conference (RE)*. 2010, pp. 383–384.

[2] Mikio Aoyama et al. "A Model and Architecture of REBOK (Requirements Engineering Body Of Knowledge) and Its Evaluation". In: *Asia-Pacific Software Engineering Conference*. 2010, pp. 50–59.

[3] Mikio Aoyama et al. "Requirements Engineering Based on REBOK (Requirements Engineering Body Of Knowledge) and Its Practice". In: *Asia-Pacific Software Engineering Conference*. 2013.

[4] Jeffrey C. Carver et al. "How Practitioners Perceive the Relevance of ESEM Research". In: *ESEM '16: Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. 2016.

[5] Xavier Franch et al. "How Do Practitioners Perceive the Relevance of Requirements Engineering Research? An Ongoing Study". In: *IEEE 25th International Requirements Engineering Conference (RE)*. 2017, pp. 382–387.

[6] IREB e.V. *Syllabus IREB Certified Professional for Requirements Engineering - Foundation Level - Version 2.2.2*. 2017.

[7] IREB e.V. and Martin Glinz. *A Glossary of Requirements Engineering Terminology*. 2014.

[8] ISO/IEC 12207:1995(E). *Information Technology — Software Life Cycle Processes*. 1995.

[9] ISO/IEC 122707:2008(E). *Systems and Software Engineering — Software Life Cycle Processes*. 2008.

[10] ISO/IEC/IEEE 12207:2017(E). *Systems and Software Engineering — Software Life Cycle Processes*. 2017.

[11] ISO/IEC/IEEE 29148:2011(E). *Systems and Software Engineering — Life Cycle Processes — Requirements Engineering*. 2011.

[12] Barbara A. Kitchenham, Tore Dyba, and Magne Jørgensen. "Evidence-Based Software Engineering". In: *ICSE '04 Proceedings of the 26th International Conference on Software Engineering*. 2004, pp. 273–281.

[13] Gerald Kotonya and Peter Sawyer. "Software Requirements". In: *SWEBOK v3.0. Guide to the Software Engineering Body of Knowledge*. Ed. by Pierre Bourque and Richard E. Fairley. IEEE Computer Society, 2014.

[14] David Lo, Nachiappan Nagappan, and Thomas Zimmermann. “How Practitioners Perceive the Relevance of Software Engineering Research”. In: *ESEC/FSE 2015: Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*. 2015, pp. 415–425.

[15] Daniel Méndez Fernández and Stefan Wagner. “Naming the Pain in Requirements Engineering: A Design for a Global Family of Surveys and First Results from Germany”. In: *Information and Software Technology* 57 (2015), pp. 616–643.

[16] Daniel Méndez Fernández et al. “Naming the Pain in Requirements Engineering: Contemporary Problems, Causes, and Effects in Practice”. In: *Empirical Software Engineering* 22.5 (2017), pp. 2298–2338.

[17] Birgit Penzenstadler et al. “The Requirements Engineering Body of Knowledge (REBoK)”. In: *21st International Requirements Engineering Conference (RE)*. 2013, pp. 377–379.

[18] Muhammad Usman et al. “Taxonomies in Software Engineering: A Systematic Mapping Study and a Revised Taxonomy Development Method”. In: *Information and Software Technology* 85 (2017), pp. 43–59.

[19] Roel Wieringa et al. “Requirements Engineering Paper Classification and Evaluation Criteria: A Proposal and a Discussion”. In: *Requirements Engineering* (2006), pp. 102–107.

[20] Claes Wohlin et al. *Experimentation in Software Engineering*. Heidelberg: Springer, 2012.

[21] Pamela Zave. “Classification of Research Efforts in Requirements Engineering”. In: *ACM Computing Surveys* 29.4 (1997), pp. 315–321.