

A Two-staged Survey on Release Readiness

S M Didar Al Alam
University of Calgary
Calgary, AB, Canada
smdalam@ucalgary.ca

Maleknaz Nayebi
University of Calgary
Calgary, AB, Canada
mnayebi@ucalgary.ca

Dietmar Pfahl
University of Tartu,
Tartu, Estonia
dietmar.pfahl@ut.ee

Guenter Ruhe
University of Calgary,
Calgary, AB, Canada
ruhe@ucalgary.ca

ABSTRACT

Deciding about the content and readiness when shipping a new product release can have a strong impact on the success (or failure) of the product. Having formerly analyzed the state-of-the art in this area, the objective for this paper was to better understand the process and rational of real-world release decisions and to what extent research on release readiness is aligned with industrial needs. We designed two rounds of surveys with focus on the current (Survey-A) and the desired (Survey-B) process of how to make release readiness decisions. We received 49 and 40 valid responses for Survey-A and Survey-B, respectively.

In total, we identified 12 main findings related to the process, the rational and the tool support considered for making release readiness decisions. We found that reasons for failed releases and the factors considered for making release decisions are context specific and vary with release cycle time. Practitioners confirmed that (i) release readiness should be measured and continuously monitored during the whole release cycle, (ii) release readiness decisions are context-specific and should not be based solely on quality considerations, and (iii) some of the observed reasons for failed releases such as low functionality, high cost, and immature service are not adequately studied in research where there is dominance on investigating quality and testing only. In terms of requested tool support, dashboards covering multidimensional aspects of the status of release development were articulated as key requirements.

Keywords

Software release engineering, release readiness, survey, empirical analysis, literature review

1. INTRODUCTION

Context: In the competitive commercial software market, software companies struggle with monitoring and controlling their software development and release processes so that they ensure competitiveness and business success. While the value of measuring and predicting release readiness is well acknowledged [21], the definition and scope of release readiness are not yet elaborated and were interpreted differently in separate studies.

Motivation: Release decisions are very complex and can cause significant damage (e.g. regarding product success, reputation, cash flow) if made incorrectly. A report published by the Clarrus Consulting Group [25] indicates that almost 50% of releases result in modest problems (or worse) with high or very high impact in terms of cost, morale, product quality and credibility. We aimed to understand the rationale underlying release decisions and identify the gaps between the perspectives of researchers and practitioners. The results of our study will help the research community to solve the right problems by understanding the real-world decision

process and thus will produce more applicable, useful and effective approaches for practitioners.

Methodology: To investigate the rationale behind release decisions in industrial practice, we designed a survey instrument consisting of two short questionnaires. We administered the surveys to product managers and senior developers in software industry in two subsequent rounds. The questionnaires consisted of 18 (Survey-A) and 11 (Survey-B) questions. The University of Calgary ethics board approved both survey questionnaires. We circulated email invitations among practitioners and advertised both surveys in a product management newsletter.

Results: Analysis of Survey-A provides insights about release decision processes actually used in practice. The comparison with research literature revealed interesting differences in concepts and approaches related to release readiness decision-making. To facilitate deeper understanding, we characterized survey participants based on their personal level of proficiency and the type of software products involved in their development. Preliminary results received from Survey-B are used to confirm some of our key findings from Survey-A and to shed light on how practitioners see release readiness decision-making in the future.

Structure of the paper: In Section 2 we briefly discuss approaches to assess release readiness proposed in the literature. In addition, we discuss literature reporting on release readiness in practice. Section 3 lists six research questions addressed in this paper. In Section 4, we describe our survey methodology and study design. Section 5 presents key demographics of our participants and introduces participant groups for analysis. Section 6 presents the results of Survey-A. In Section 7, results from Survey-B are directed towards consolidating former findings and to envision target process and tool support. Section 8 discusses limitations of our study followed by conclusions in Section 9.

2. RESEARCH QUESTIONS

A software release is well defined in the International Software Product Management Association (ISPMA) glossary. Therein, the term *software release* is defined as *an instance of the product that is delivered to customers, and maintained as part of product maintenance*" [8]. A number of informal definitions [21], [22], [27] exist of what is considered *release readiness*.

To investigate the rationale behind practitioners' decisions of when and why to ship a software release we analyzed release readiness from three perspectives, i.e., i) why releases fail, ii) how practitioners make release decisions (*status quo*), and iii) how practitioners think release decisions should be done (*to-be state*).

We address the following research questions:

RQ1: *What causes release failure?*

To define success, one must understand failure and its causes. We asked practitioners to rank four typical causes of failure identified in the literature.

RQ2: *Which factors influence release decisions?*

Even when there is a high risk of failure, product managers tend to deliver a version into the market. To understand what factors impact release decisions we asked participants to weight seven key factors with respect to their influence on release decisions.

RQ3: *How success and failure of a release is measured and who is relying the most on measurement?*

Extending RQ1 and RQ2 we were interested in understanding if, how, and to what extent, decision makers actually measure the risk of failure of a release. We asked participants to weight nine key measures based on their importance in release readiness decision.

RQ4: *How well correspond the practitioner's and the researcher's perspectives on release readiness?*

We analyzed the gap between state of the art and state of practice. We believe that this gap analysis will help the research community in focusing their research on the most urgent problems.

RQ5: *How release readiness decisions should be done?*

We wanted to understand the practitioner's view on how release readiness decisions should be done. Based on a previous literature survey, we offered six alternatives for the key decision rationale.

RQ6: *What are key needs for tool support on release readiness decisions?*

Besides the key decision rationale, we wanted to understand the key needs for tool support.

3. RELATED WORK

This paper studies the state-of-the practice in release readiness and analyze the gap with existing research conducted in this field. We analyze the current and the intended status of release readiness from an industrial perspective. Related work on these two aspects is discussed in the following two subsections.

3.1 Release readiness

Deciding about the readiness to ship a product release is one of the key decision problems encountered in the context of release management. Measurement and various forms of data mining, in combination with analytical methods, are considered key drivers to facilitate release decision-making [17].

Readiness of a software release is a widely used concept but the definition is not yet well established. To understand the concept better, we performed a semi-systematic review of release readiness literature which is largely but not perfectly, following the established process defined by Kitchenham et al. [10]. From 23 selected articles, we identified a variety of approaches for analyzing release readiness. Each approach is related to one of the four major categories listed below:

Testing metrics: Release readiness indicators based on testing related metrics (e.g. test passing rate, defect find rate) [12], [18], [26], focusing exclusively on the testing. 43% of the studied articles fall under this category.

Defect prediction: Considers remaining defects as a proxy measure for release readiness and proposes defect prediction models

applying multiple techniques, e.g., neural networks [22], code change analysis [27]. 22% of the studied articles fall under this category. All of them exclusively focus on quality and thus limit the possibilities of a broader view on release readiness.

Checklists: Subjectively check a list of release readiness criteria at the end of the release cycle [13], [23]. 22% of the studied articles proposed this approach.

Multi-dimensional metrics: Evaluate a portfolio of metrics related to multiple measurement dimensions (e.g. quality, functionality, documentation) to measure release readiness [2], [24]. Ad hoc selection of metrics imposes the risk of misleading release readiness evaluation.

We observed that testing and quality were often used in measuring release readiness. These approaches refer to certain phases of the development cycles and mostly focus on *testing scope and status* based metrics [12], [22]. The majority of existing approaches also identifies release readiness towards the end of a release cycle [4], [12], [18], [21]. This often imposes a lack of continuity in monitoring release readiness.

In addition, many of the existing approaches require data that often is not available. For example, approaches proposed in [2] and [24] aggregate multiple release readiness attributes into one single measure although it is difficult to find the corresponding data. In such cases, the applicability of the approach highly depends on expert judgment.

Brousseau [25] performed a survey on "How software teams release their products". Although the survey was conducted with only 18 participants, it provided some interesting findings:

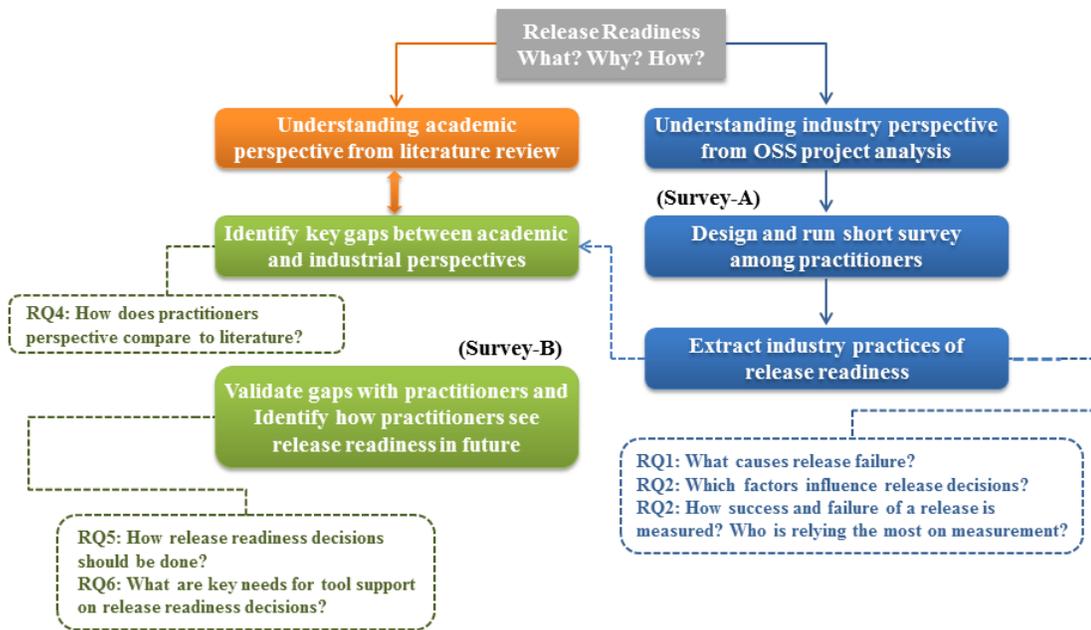
- The strongest driver for issuing a new product release is that "the promised release date has arrived";
- The decision to release and the evidence used to decide suggest that *when to release software* is a complex, multi-faceted problem;
- Almost 50% of the releases result in modest problems (or worse), and almost 25% of the releases result in significant or very severe issues.

Comparing the results of Brousseau's study with the current body of literature showed some mismatch. In particular, the key drivers affecting release decisions in practice and the evidence gathered by practitioners as an indicator of release readiness are different compared to what researchers proposed in literature. This existing gap motivated us to further investigate this topic.

3.2 Relating academic research to industrial needs

In software engineering, relating academic research to industrial needs has always been considered to be important. While many researchers have been working towards improving industry-academia collaborations, the number of successful collaborations is still low. To identify general challenges and best practices for such industry-academia collaborations Garousi et al. [6] conducted a systematic literature review study. The authors studied 33 articles between 1995 and 2014 and performed a grounded-theory based qualitative synthesis. As a result, they reported challenges and best practices that are hoped to facilitate success of industry-academia collaboration in future.

Figure 1: Main context and relationship of the research with its related research questions (RQ's).



Along the same lines, Nayebi et al. [16] studied the gap between academic approaches and industry needs with respect to the use of data analytics techniques in software engineering project management. A systematic mapping study comprising 115 papers was conducted. Comparing these results with a survey on analytics needs in industry [3] allows researchers to apply analytics more aligned with practice.

4. METHODOLOGY

In this section, we explain the main context and relationship of the research with its related research questions (RQ's).

4.1 Overview

Our proposed methodology is a combination of an analysis of existing research on release readiness, analyzing the state-of-the-practice and performing two rounds of surveys with an industrial audience. An overview of the process is given in Figure 1. Initially, a comprehensive literature review [1] was conducted to better understand how the concept of *release readiness* is presented in the research literature. Largely following the guidelines of [10], we identified 23 articles. We extracted metrics applied, proposed approaches and tools used.

In the next step, we tried to better understand the concept of release readiness in practice by analyzing some real world projects. We analyzed 34 open source projects hosted in GitHub [1]. The selected projects were equally distributed across the desktop and web-based software domains. We considered six established release readiness attributes. As a form of explorative case study research, we monitored their performance in retrospective over a two-year period for each project. As a result, we got an initial understanding of release practices at least for open source projects.

We designed two follow-up surveys. Survey-A asks practitioners about their experience on release readiness and aims at getting answers to RQ1, RQ2, and RQ3 to identify the actual release readiness practices, decision-making process, and monitoring of release readiness. Survey-A identifies the practitioners' perceived *as-is* status.

Next, we answered RQ4 and performed a gap analysis by comparing Survey-A findings with release readiness results obtained from the literature.

In the last step of our study, Survey-B was performed. We aimed at getting answers to RQ5 and RQ6 to validate some key findings from Survey-A and to ask practitioners i) how they would like to make release readiness decisions (*to-be*) and ii) what are their key needs for tool support.

4.2 Survey preparation and design

We designed our survey instrument for product managers and senior developers responsible for release decisions in industrial software organizations. Participation in both surveys was anonymous and voluntary.

Survey-A is a ten-minute survey with a total of 18 questions split into two branches. It follows the principles of a descriptive design [19] that explores and captures industry practices for release readiness measurement. Participants always had the option to skip a question except one. The mandatory question identifies the appropriate branch for follow-up questions. Survey-B is a five-minute survey with 11 questions. All questions were optional. In both surveys, participants could answer the majority of questions on a five-point Likert scale [11]. Both surveys are accessible via our website^{1,2}. They were approved by the University of Calgary

¹ <https://goo.gl/forms/KM3HY6epcAuV8sWU2>

² <https://goo.gl/forms/AiBjITu8wNH37jyh2>

ethics compliance board CFREB³. One software engineering graduate student, one practitioner and one software engineering faculty checked the validity [20] of the survey instrument. Based on their feedback we improved the understandability of the survey questions. We used Google Forms to create, distribute and manage our surveys.

4.3 Data collection and processing

We advertised both surveys in the product management newsletter *InnoTivum*⁴. In addition, we sent email invitations to 50 globally distributed product managers whom we knew. Survey participation was voluntary and participants had the option to withdraw anytime.

Survey-A was open for 5 weeks during the period October to November 2016. We received a total of 55 responses. We filtered six incomplete responses. We grouped the remaining 49 responses based on their characteristics related to the participant, corresponding product she is involved in and the underlying development process (see Section 5). All responses were analyzed across these categories (denoted as *participant groups*). We applied the Mann-Whitney U-Test [14] to verify statistical significance of the differences between participant groups.

Survey-B was open for two weeks in January 2017. The survey received 40 responses. While all closed question responses were complete, we had 15 missing responses for open questions.

5. DEMOGRAPHICS

We characterized participants of Survey-A with regards to five features related to both personal level of proficiency and type of software products developed. For each feature, we grouped the responses into two groups in order to have roughly balanced groups in the data analysis.

Experience: The number of years a participant had experience in software engineering. 38.5% of the participants had between one to five years of experience, while 61.5% of the participants had more than five years of experience.

Responsibility: The survey includes participants with different roles in a team. Responsibility represents a participant's level of involvement in release decisions. Participants had to respond on a five-point Likert scale. 43.6% of the participants stated very little to medium responsibility in release decisions, while 56.4% of the participants stated they had high or very high responsibility.

Number of major releases: Participants responded to our survey questions based on one of the products she is currently involved in as a product manager or senior developer. 43.6% of the participants reported less than five major releases, while 56.4% participants reported five or more major releases.

Release frequency: 47.3% of the participants' stated that they ship new releases for their product monthly, bi-weekly, weekly or even more frequently. 52.7% of the participants had quarterly or yearly releases.

Maturity: To measure maturity of product management, we asked participants to use a five point Likert scale that resembles the CMMI maturity level definitions [5]. 39.4% of the participants stated the achieved maturity of product management is either

between maturity levels one or two. 60.6% of the participants placed the maturity of product management higher than level two.

In Survey-B, our objective was to learn the general view of the practitioners on how release readiness should be measured. No demographic information was collected, but in cases where we could influence the selection of survey participants, the preference was on experienced product managers.

6. ANALYSIS RESULTS FROM SURVEY-A

In this section, we discuss the findings related to RQ1, RQ2, RQ3 and RQ4.

6.1 What causes release failure (RQ1)?

All of the participants faced failure and realized a wrong release decision after product shipment. Figure 2(c) shows the majority of participants "often" or "rarely" realized that they made a wrong release decision after shipment. To gain these result, we asked the participants how often they faced a failure and then they should evaluate the severity of the damage. Both questions had to be answered on a five point Likert scale [11]. Release failure occurred rarely but when it happened 40% of the participants said the extent of damage (i.e. regarding product success, reputation, cash flow) was "much" or "very much" (cf. Figure 2(d)).

Talking about the factors causing a failure, the participants considered *low functionality* (expected functionality not offered), *poor quality* (expected quality not achieved), *immature service* (wrong functionality offered), and *high cost* (pricing being higher than what customer is willing to pay) all as equally frequent reasons for failure. Figure 2(b) presents the frequency of encountering release failure due to different reasons of failure. We did not observe any significant difference between the frequencies of these factors.

Finding 1: There is no dominant but multiple reasons considered relevant for release failure in practice.

The lack of significant difference between the frequencies of the four failure factors made us to investigate whether there are differences in significance considering participants' demographics. However, we could not find any interesting significant difference using the Mann-Whitney U test between participant groups.

While investigating release failure rate among participant groups, Mann-Whitney U test showed, more matured organizations have a lower release failure rate compared to less mature organizations. In a correlation analysis among four failure factors, extent of damage and realizing release failure rate (cf. Figure 2(a)), we found significant correlation between the frequency of encountering failure because of *poor quality* and *immature services*.

Finding 2: More matured organizations have a lower release failure rate compared to less mature organizations.

6.2 Which factors influence release decisions (RQ2)?

We investigated the most influential factors on release by asking participants to weight seven key factors we gathered from literature. We selected these factors considering the literature study on 23 articles reported in [1] and a former survey results on release

³<https://www.ucalgary.ca/research/researchers/ethics-compliance/cfreb>

⁴<https://www.innotivum.com/>

practices [25]. Participants had to assess each factor independently on a five point Likert scale by considering the influence of it on release decisions. The Likert scale ranged from “very little” (= 1) to “very much” (= 5) impact.

For each of the factors, Figure 3 presents the distribution of participants over the five offered choices. Implementing a *new functionality* and *achieving the targeted quality* are found to be the most influential factors in a release decision. *Customer and time pressure* as well as *experience* were considered equally the second

most important with *median = 4*. *Competitor releases* were considered the least influential factor for releasing a version while *social events* were considered the second least important.

Finding 3: *New functionality* and *achieved quality* are the most influential factors on release shipping decisions.

Finding 4: *Customer pressure*, *time pressure* and *experience* are about equally influential on release decisions.

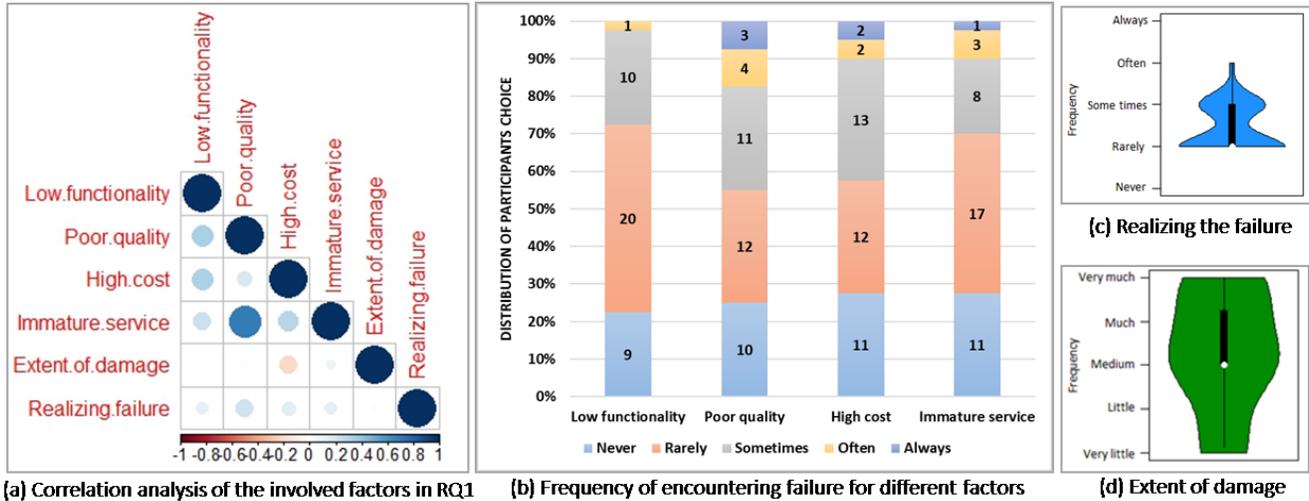


Figure 2: Results of RQ1- What causes release failure?

Considering the demographic information of participants, we did not find significant differences between participant groups. However, the more experienced participants have the tendency to rely more on *experience* and *social events* comparing less experienced participants. We also observed that *customer pressure* is more influential on release decisions than the other factors if participants stated the maturity of their product management to be low (maturity levels one or two).

frequencies of factors influencing a release decision (such as *new functionality* or *achieved quality*).

Finding 5: *Achieved quality* is more influential on release decisions for the case that observed frequency of release failure is high.

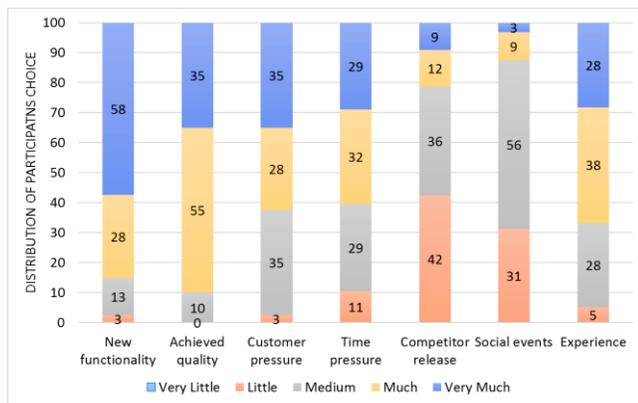


Figure 3: Participants choices across key influential factors.

Applying the Mann-Whitney U test between participant groups, we found that the influence of *achieved quality* significantly (p-value = 0.044) depends on the frequency of realizing release failure. There was no correlation between the frequencies of factors causing a release failure (such as *poor quality* or *low functionality*) and the

6.3 How success and failure of a release is measured and who is relying the most on measurement (RQ3)?

We assumed that release decisions are impacted by measurement. Therefore, we asked the participants to what extent their release decisions are influenced by measurement. In addition, we asked participants to weight nine release measures with respect to their influence on release decisions. In our literature study, we found these measures are influential on release decisions. Moreover, they cover a large portion of the commonly used release metrics in industry [23]. As shown by the grouping in Table 1, these measures represent three measurement dimensions, i.e., *implementation status*, *testing scope and status*, and *source code quality*.

Participant performed the weighting of nine release measures on a five-point Likert scale (i.e. very low=1 to very high=5 importance). Figure 4(top) presents a stacked bar chart showing participant choices with respect to the importance of release measures in release decision-making. *Feature completion rate*, *change request completion rate*, *bug fix rate*, and *defect find rate* are the measures that were considered most influential. The average importance of most of the measures was between 3 and 3.5 (close to category “medium”). Only *code complexity* had an average weight below 3.

Finding 6: *Feature completion rate, bug fix rate, defect find rate and build success rate* are the most important measures used for making release readiness decisions.

We compared relative ranking of measures among groups with high and low failure rate in Figure 4 (bottom). The lines between the two bar charts join the same measure under two participant groups. *Bug fix rate* and *feature completion rate* are the most influential measure respectively for groups with high and low failure rate. The largest difference in relative ranking was found in *build success rate* and *continuous integration rate*. The second largest difference exists in *bug fix rate, feature completion rate* and *estimated and actual efforts ratio*.

Applying the Mann-Whitney U test between participant groups we found that participants who had more major releases of their product in past have a tendency to rely more on *change completion rate* (p-value = 0.006) in contrast to the participants who had less major releases of their product in past.

Finding 7: Participants who had products with higher number of major releases rely more on *change completion rate*.

The success of a release also depends on the effectiveness of applying measurement in decisions. 89% of our participants indicated they perform measurement. 41% of the participant indicated they heavily (i.e. much/very much) rely on measurement for release decisions.

Table 1: List of nine release measures considered in this study along with their definition and acronym.

| Measurement dimension | Release measure | Measurement definition |
|--------------------------|-----------------------------------|---|
| Implementation status | Feature completion rate | # implemented features till week (k) / # requested features till week (k) |
| | Change request completion rate | # implemented changes till week (k) / # requested changes till week (k) |
| | Continuous integration rate | # of continuous integrations performed per week (k) |
| Testing scope and status | Bug fix rate | # resolved bugs till week (k) / # requested bugs till week (k) |
| | Defect find rate | # identified defects in week (k) / # identified defects till week (k) |
| | Test coverage | # of software units covered by test code till week (k) / # total implement software units till week (k) |
| Source code quality | Build success rate | # of successful builds in week (k)/# of total builds in week (k) |
| | Code complexity | # of linearly independent path is a software unit |
| | Estimated and actual effort ratio | the ratio of estimated and actual effort |

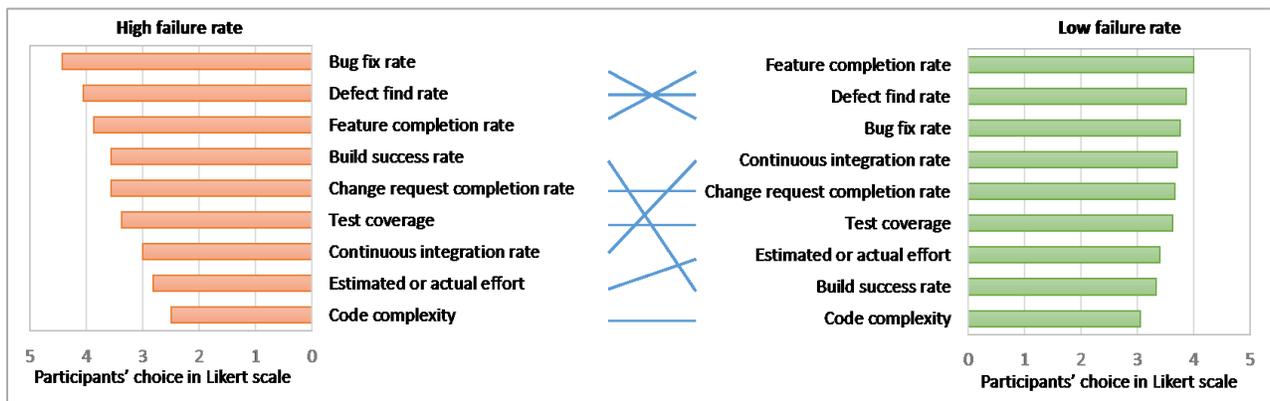
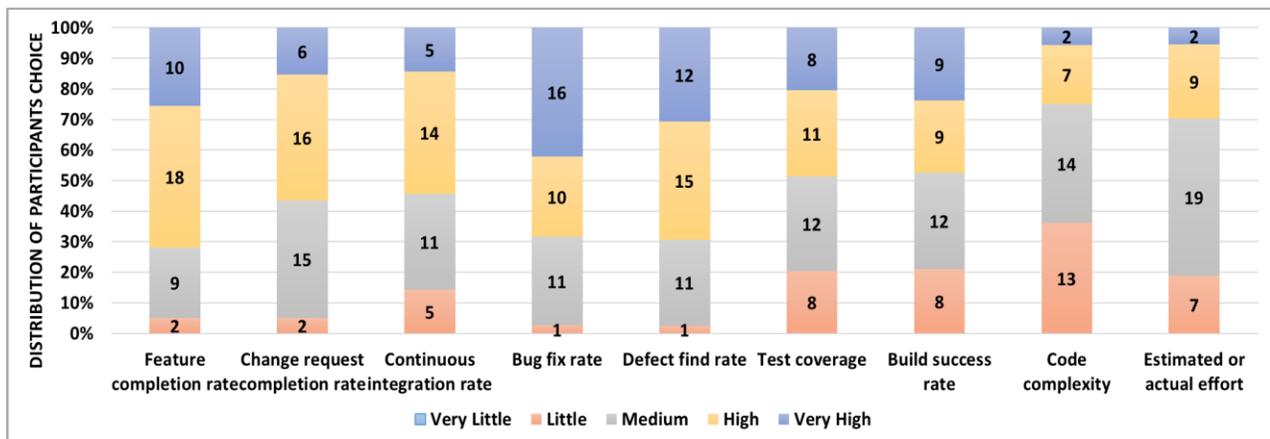


Figure 4: Participants' preferences with respect to measures of importance in release readiness decision-making for products with high (left) vs. low (right) failure rate.

Applying the Mann-Whitney U test we found that participants with five or more years of experience (p-value = 0.005) and participants with high responsibility (p-value = 0.001) in release decisions are the most reliant on measurement.

Finding 8: Participants with five or more years of experience rely more on measurement compared to less experienced participants.

6.4 How well correspond the practitioner’s and the researcher’s perspectives on release readiness (RQ4)?

Release readiness literature proposed various unique methods for measuring release readiness. As shown in Section 3, these approaches vary in their choice of metrics and the development phase focused for measurement. Despite the variation in metrics selection, all existing approaches heavily rely on measurement. Along the same line, Survey-A shows, 89% of participants apply measurement and 41% participants heavily rely on measurement to support their release decisions. Therefore, we conclude that measurement based solutions proposed in the literature would be useful to industry.

We identified four categories in our studied approaches based on checklists (22% articles), testing metrics (43% articles), defect prediction (22% articles), and multi-dimensional metrics aggregation (11% articles). Majority (87%) of studied approaches identifies release readiness towards the end of a release cycle [4], [12], [18], [21]. In contrast, Survey-B shows, 58% participants consider release readiness should be measured continuously from the beginning of the release not only at the end of the release.

Figure 9: Release readiness should be monitored continuously from the beginning of a release instead of the end of the release cycle.

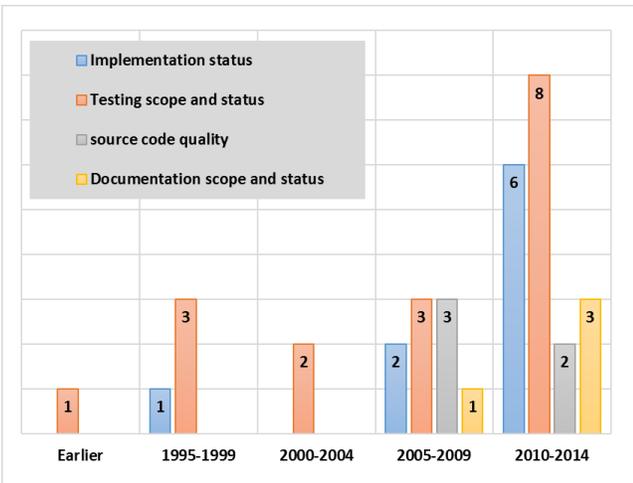


Figure 5: Frequency of publications applying metrics correspond to different measurement dimensions.

In addition, we observed numerous measures are proposed related to *achieved quality* along with *new functionality* both in literature and practice. For example, in Microsoft release checklist [23] we identify multiple measures related to *quality* and *functionality* of a release. But rarely we see metrics available for measuring *customer pressure, time pressure or experience*. However, practitioners considered these three factors equally the second most influential factors on release decision in Section 6.2.

We observed some mismatch between the factors studied in the research literature for release decisions, and the factors practitioners are relying upon. For example, the results presented in Section 6.1 indicate that all four reasons for release failure are equally frequent in industry, whereas over 65% of the release readiness literature considered *poor quality* as the most frequent reason for failure over all other reasons.

Some mismatch seems to exist between the focus of measurement in literature and practice. To better understand this mismatch, we illustrated all metrics used in studied approaches and broadly identified four measurement dimensions i.e. *implementation status, testing scope and status, source code quality, and documentation scope and status*. Figure 5 shows the frequency of publications applying metrics representing different measurement dimensions. 73% of publications considered release measures from the *testing scope and status* dimension, 40% considered measures from the *implementation status* dimension, and only 20% considered measures from the *source code quality* and *documentation scope and status* dimensions. Measures representing the *testing scope and status* dimension are prioritized over other dimensions. In contrast, Section 6.3 indicated that measures from different measurement dimensions are equally important in release decisions. Among the top four measures we found two measures representing the *testing scope and status dimension* and two measures representing the *implementation status* dimension.

Findings 10: *Quality and testing* measures alone are insufficient for making release readiness decision. Instead, readiness should be decided comprising multiple measures.

7. ANALYSIS RESULTS FROM SURVEY-B

We discuss the findings related to RQ5, RQ6 in this section primarily based on Survey-B analysis results.

7.1 How release readiness decisions should be done (RQ5)?

Based on the results of Survey-A (reflecting the status quo), we also wanted to better understand the target rational for release readiness decisions: To what extent release readiness decisions should be based on one of the following options:

- Gut-feeling (Q1)
- Customer & time pressure (Q2),
- Implemented new functionality (Q3),
- Achieved quality (Q4),
- Tradeoff between functionality and quality (Q5), or
- Combination of new functionality, achieved quality, customer and time pressure (Q6).

The evaluation was performed on a five-point Likert scale (with strongly disagree = 1 to strongly agree = 5). Following [15], we grouped our responses into four categories for ease of our analysis: i) Agree (agree, strongly agree), ii) Disagree (disagree, strongly disagree), iii) Neutral, iv) No response.

As can be seen in Figure 6 (left), the majority of the participants agreed that release readiness decisions should be made based on *implemented new functionality* (85%) and *achieved quality* (95%) of the release. 70% participants said it is a trade-off between *implemented new functionality* and *achieved quality*. 65% participants also considered including *customer and time pressure* in release readiness measurement. Above mentioned observations support our conclusions made in RQ2 and RQ4. It confirms that i) release readiness is a multi-dimensional concept, and ii)

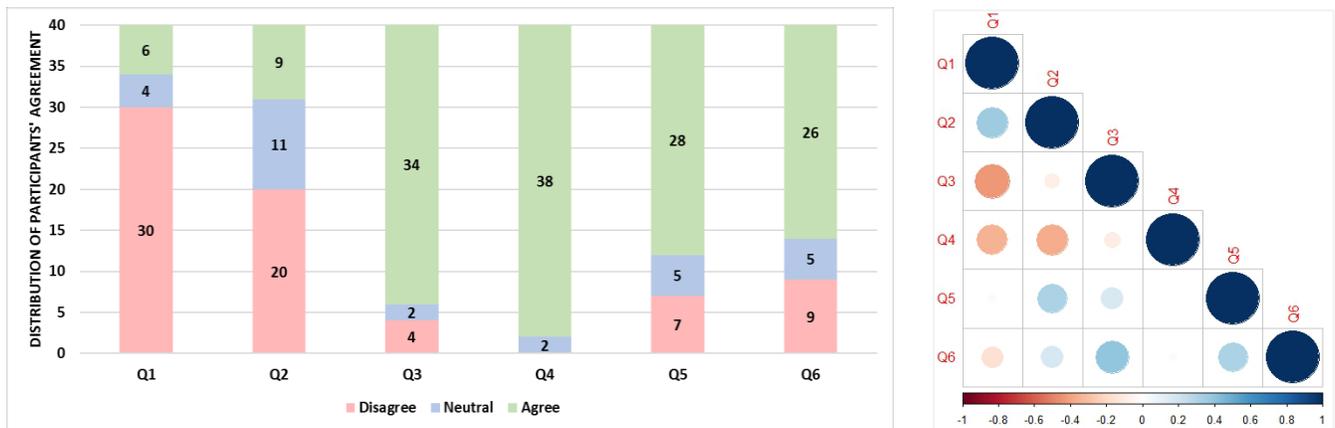


Figure 6: Distribution of participants' agreement on how decisions should be done (left) and related correlations (right).

Table 2: Key barriers that participants faced in release readiness decision-making (top) and key tool features requests stated (bottom).

| | Dendrogram | Category | Content |
|--------------|------------|---|--|
| Key barriers | | Barriers related to internal development process (57%) | Quality and functionality issues (23%) |
| | | | Lack of control due to automation (21%) |
| | | | Improper estimation (13%) |
| | | Barriers related to lack of tools and techniques (33%) | Lack of market information (13%) |
| | | | Dependency on external environment (20%) |
| | | Barriers related to external influence / dependency (10%) | Lack of tools (5%) |
| | | | Lack of key performance indicators (5%) |
| Key features | | Summarizing measures in a dashboard (43%) | Collecting business info (21.5%) |
| | | | Presentation in dashboard (21.5%) |
| | | Identify and gather completeness from multiple dimensions (35.5%) | Multidimensional completeness (21.5%) |
| | | | Supporting test completeness (14%) |
| | | | Supporting existing tools (21.5%) |

implemented new functionality and achieved quality are considered the most important factors in release readiness decisions.

Further correlation analysis presented in Figure 6(right) among measurement scopes identified that i) participants agreeing on achieved quality (Q4) and new functionality (Q3) based release decision have a high negative correlation with gut-feeling based (Q1) release decisions, and ii) participants agreeing with new functionality (Q3) based release decision have a positive correlation in inclusion of customer and time pressure (Q6) in release readiness measurement.

7.2 What are key needs for tool support on release readiness decisions (RQ6)?

Practitioners apply multiple code and issue management tools to understand readiness of individual software development phases. 85% participants in Survey-B indicated they require tool support

for release readiness measurement. To better understand this need we asked about

- key barriers in release readiness measurement,
- key features required in a release readiness tool.

We summarized responses using the open card sorting [7] technique. Open card sorting technique is a 'generative' method. It allowed us to derive taxonomies by identifying and organizing topics (denoted as 'cards') from open responses. For key barriers and key features we initially identified 47 and 20 cards respectively. These cards were further organized in hierarchical categories based on their relative meaning in context of release readiness decision making. We applied the Uxsort⁵ tool for open card sorting analysis.

The top levels of extracted categories are presented in Table 2 along with the underlying categories and the percentage of cards under each category. Using dendrogram visualization, we presented the

⁵ <http://www.uxsort.com/>

relationships between extracted categories. For space limitation, we skip presenting individual cards and presented all card categories in the dendrogram.

Finding 11: Practitioners consider *internal software development* (57%), *lack of tools and technique* (33%) and *external dependency and influences* (10%) as key barriers for decision-making.

Finding 12: Participants categorized tool requirements into three groups which includes i) *summarizing information into a dashboard* (43%), ii) *identifying completeness* (35.5%) *from multiple dimensions*, and iii) *supporting existing code, issue and build management systems* (21.5%).

8. THREATS TO VALIDITY

This section lists key threats to validity of this study and report the mitigation actions taken to minimize each threat.

We prepared both surveys for product managers and senior developers with experience in release decisions. To achieve variation in participants, we circulated the invitation among practitioners with varying experience, and varying products and development processes. However, due to the design and content of the survey, we have the risk of self-selection bias. That means, only a limited group of practitioners from certain types of products who actually do measurement may have participated. This can impose external validity threat on this study and limit generalizability of the results. We received 56 and 40 responses for Survey-A and Survey-B, respectively, which is a moderate sample size. In Survey-A we don't find a heavy impact of this threat as participant demographics is well distributed. For Survey-B we skipped collecting demographics information and thus the impact of self-selection bias is unknown. However, a high impact is acceptable to us as Survey-B exclusively investigates i) key needs for tool support and ii) future of release readiness. Experienced product managers are certainly the key focus for participation in Survey-B.

Responses were collected subjectively in a five-point Likert scale. Subjective opinions can vary across products and impose threats to internal validity of the survey. To reduce this threat, each participant was asked to consider one of the products she is currently involved in for responses. In addition, we consider participant, product, and process related characteristics in our result analysis.

Our responses consist of non-normally distributed ordinal data achieved from Likert scale choices. For ease of analysis, we introduced participant groups based on demographic information of participants. We consistently tested our null hypothesis for one independent variable between two participant groups. We choose the Mann-Whitney U test for our analysis [14]. The risk of achieving unbalanced groups imposes conclusion validity threats. As a mitigation step, we checked multiple combinations of participant groups and considered the one with maximum balance. However, Kitchenham et. al. [9] stated that the Mann-Whitney U test is strongly affected by unequal variances, despite of their even sample size. Therefore, the variances observed in participant groups can still impose conclusion validity threat.

We assumed participants have experience with release decisions. But the risk of having no experience or different type of experience will impose serious construct validity threats. As a mitigation step, we identified participant's involvement in release decision along with her work experience.

We filtered 10 responses from participants who had less than a year experience and low involvement in release decisions. We assumed, these responses could introduce noise in our analysis. In contrast, inclusion of these responses might have allowed a broader perspective of release readiness practices. The underlying tradeoff is a validity concern for our study.

Four responses were filtered from participants, who don't perform measurement. These participants are a key part of the release decision process and should be thoroughly investigated. We had a survey instrument ready for identifying challenges behind skipping measurement. However, with only four responses we could not report any results.

9. CONCLUSIONS

It is inherently important to relate academic research to industry needs. To achieve this objective for release readiness research, we performed two industry surveys to understand industry practices and a comprehensive literature review to understand researcher's perspectives. We further compared results between literature and industry practices and the results are helpful for both researcher and practitioners in their release readiness decisions.

Practitioners heavily apply measurement in release decisions as 89% of participants said they perform measurement and 41% participants said they heavily (much, very much) rely on measurement for their release decisions. We observed, literature considered *poor quality* as the most frequent reason for release failure. In contrast, survey results suggest, not one but multiple reasons (e.g. *high cost, low functionality, immature service*) are equally frequent for failed releases.

Our findings show *new functionality* and *achieved quality* are the most influential factors on release shipping decisions. *Customer pressure, time pressure* and *experience* have a similar degree of influence on release decisions. While industry and literature both apply numerous metrics to measure *achieved quality* and *new functionality* rarely any measure is available for the others.

We also identified *feature completion rate, bug fix rate, defect find rate* and *build success rate* as the top four measures used for release shipping decisions. The majority of suggested approaches in the literature rely heavily on quality and test metrics for release readiness decisions. In practice, release readiness assessment requires multi-dimensional measurement instead of only quality or testing based metrics.

Survey-B results further support our conclusion to increase release readiness assessment scope towards multi-dimensional measurement. Exclusive focus on testing and quality of the release may narrow product managers view of success and end in wrong release decisions. Investing in both academia and industry perspectives will allow researchers to solve the most essential problems and facilitate achieving effective methods and approaches aligned with practitioner's perspective on release readiness.

In conclusion, we found that **release readiness decision-making** is a multi-dimensional concept that requires continuous and customized measurement. *Implemented new functionality* and *achieved quality* are considered the most important attributes in the trade-off process for release readiness.

ACKNOWLEDGEMENT

We thank all participants for their valuable responses and also thank ISPMA and organizers of “InnoTivum” newsletter for circulating our survey among the subscribers. This work was partially supported by the Natural Sciences and Engineering Research Council of Canada, NSERC Discovery Grant 250343-12, Alberta Innovates Technology Futures and institutional research grant IUT20-55 of the Estonian Research Council.

REFERENCES

- [1] Alam, S. et al. 2015. Monitoring and Controlling Release Readiness by Learning across Projects. *Managing Software Process Evolution*. Springer. 281–303.
- [2] Asthana, A. and Olivieri, J. 2009. Quantifying Software Reliability and Readiness. *Proceedings of the IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR)* (2009), 1–6.
- [3] Begel, A. and Zimmermann, T. 2014. Analyze This! 145 Questions for Data Scientists in Software Engineering. *ICSE* (2014).
- [4] Brettschneider, R. 1989. Is your software ready for release? *IEEE Software*. 6, 4 (1989), 100.
- [5] Chrissis, M.B. et al. 2011. *CMMI for development: guidelines for process integration and product improvement*. Pearson Education.
- [6] Garousi, V. et al. 2016. Challenges and best practices in industry-academia collaborations in software engineering: A systematic literature review. *Information and Software Technology*. 79, 106–127 (2016).
- [7] Hudson, W. 2013. Card sorting. *The Encyclopedia of Human-Computer Interaction, 2nd Ed.*
- [8] International Software Product Management Association Glossary: http://community.ispma.org/wp-content/uploads/2014/12/ISPMA_Glossary_2.0.pdf.
- [9] Kitchenham, B. et al. 2016. Robust Statistical Methods for Empirical Software Engineering. *Empirical Software Engineering*. (2016), 1–52.
- [10] Kitchenham, B. and Charters, S. 2007. Guidelines for performing Systematic Literature reviews in Software Engineering Version 2.3. *Engineering*. 45, 4ve (2007), 1051.
- [11] Likert, R. 1932. A Technique for the Measurement of Attitudes. *Archives of Psychology*. 22, (1932).
- [12] Mcconnell, S. 1997. Gauging software readiness with defect tracking. *IEEE Software*. 14, 3 (1997), 135–136.
- [13] Measurements to Reduce Risk in Product Ship Decisions: <https://www.universityalliance.com/info1/whitepaper>.
- [14] Motulsky, H. 2013. *Intuitive biostatistics: a nonmathematical guide to statistical thinking*. Oxford Univ. Press.
- [15] Murphy, B. et al. 2013. Have agile techniques been the silver bullet for software development at Microsoft? *Proc. ESEM* (2013), 75–84.
- [16] Nayebi, M. et al. 2015. Analytics for Software Project Management--Where are We and Where do We Go? *ACTION'15, co-located to ASE* (2015), 18–21.
- [17] Nayebi, M. and Ruhe, G. 2015. Analytical Product Release Planning. *The Art and Science of Analyzing Software Dat*. Morgan Kaufmann Publishers, Inc. 550–580.
- [18] Pearse, T. et al. 1999. Using Metrics to Manage the End-Game of a Software Project. *Proceedings of the Sixth International Software Metrics Symposium* (1999), 207–215.
- [19] Pfleeger, S.L. and Kitchenham, B.A. 2002. Principles of survey research part 2: designing a survey. *Software Engineering Notes*. 27, 1 (2002), 18–20.
- [20] Pfleeger, S.L. and Kitchenham, B.A. 2002. Principles of survey research part 4: questionnaire evaluation. *Software Engineering Notes*. 27, 3 (2002).
- [21] Port, D. and Wilf, J. 2013. The Value of Certifying Software Release Readiness: An Exploratory Study of Certification for a Critical System at JPL. *Proc. ESEM* (2013), 373–382.
- [22] Quah, T.-S. 2009. Estimating software readiness using predictive models. *Information Sciences*. 179, 4 (2009), 430–445.
- [23] Release Readiness Review: 2017. <https://technet.microsoft.com/en-us/library/cc5266>.
- [24] Satapathy, P. 2013. *Evaluation of Software Release Readiness Metric across the software development life cycle*. Department of Computer Science & Engineering, University of California, Riverside.
- [25] Software Release Practices – Survey Results: <http://www.clarrus.com/2013/05/software-release-practices-survey-results/> Last accessed November 2014.
- [26] Ware, M.P. et al. 2008. The Use of Intra-Release Product Measures in Predicting Release Readiness. *Proc. of the 1st International Conference on Software Testing, Verification, and Validation* (2008), 230–237.
- [27] Wild, R. and Brune, P. 2012. Determining Software Product Release Readiness by the Change-Error Correlation Function: On the Importance of the Change-Error Time Lag. *Proc. HICSS* (Jan. 2012), 5360–5367.