Universiteit Leiden

ICT in Business and the Public Sector

Understanding the Impact of Large-Scale Agile Transformations

Name: Tim Poot
Student-no: s1514113

Date: 25/05/2022

1st supervisor: Dr. C.J. Stettina MSc
2nd supervisor: Prof.dr.ir. J.M.W. Visser

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

ABSTRACT

In the recent past more academic literature on the impact of large-scale agile transformations has appeared. Most of this literature agrees that large-scale agile transformations have a positive impact on organizational performance, however the findings differ in the exact benefits. Furthermore there is little to no research on the impact on the portfolio level of the organization.

As such in this thesis a framework for organizational performance is constructed. This framework is also extended with a set of portfolio metrics. This framework is then used to measure the impact of large-scale agile transformations on organizational performance. These results are then tested for correlation with transformation maturity, this is done to further the insight on how the experienced benefits change as a transformation progresses. Both the performance framework and the maturity model are also further analyzed to see if they can be improved further.

The framework is constructed by collecting previously reported benefits from academic literature, giving preference to literature on large-scale agile contexts wherever possible. This framework is then used in an online survey to measure the impact on organizational performance. The results on the impact are then tested for correlation with transformation maturity as measured by a maturity model from literature, which will also be included in the survey. To see how the collected impact metrics relate to each other, their correlation with each other will also be tested. To test the maturity model a graded response model will be used.

With a total of 61 completed responses, the results suggests that the majority of performance metrics do benefit from a large-scale agile transformation. There are several points of overlap on what the metrics are that benefit the most, like increased transparency, collaboration and earlier detection of defects. Less correlations between performance metrics and transformation maturity are found then in previous literature. Between the metrics themselves there are numerous high correlation coefficients which are also significant. The results of the graded response model agree for the most part with the maturity model but also show some discrepancies.

Overall the notion that large-scale agile transformations have a positive impact on organizational performance is confirmed again. Further study is required to make conclusive statements on the correlation between performance metrics and transformation maturity, of special interest would be a study that distinguishes between top-down and bottom-up transformation strategies. The created performance framework could also benefit from continued research, mostly by making it more concise through eliminating highly correlated metrics. Several suggestions for the maturity model are also made, including a suggestion for further testing through other methods such as a focus group.

Contents

1 Introduction

Due to the benefits experienced by agile times, large enterprises have tried to become more agile. This does not only apply to teams within the organization but throughout the entire organization [Digital.ai Software Inc., 2021]. Doing so proved to be challenging, as many early agile methods are often not well suited for large projects [Dyba and Dingsoyr, 2009]. To enable large enterprises to adapt agile ways of working various frameworks were created, such as the Scaled Agile Framework (SAFe) and Large Scale Scrum (LeSS). Over the years working agile at scale has become more and more popular Business Agility Institute [2020],

The large scale agile (LSA) framework SAFe makes various claims about performance benefits such as a $20-50\%$ increase in productivity and $30-75\%$ faster Time-to-Market[1]. Comparisons with these self reported figures in academic literature were not found outside of Stettina et al. [2021]. There has been an increase in academic literature on LSA in general, mostly focusing on the success-factors and challenges of such transformations [Dikert et al., 2016, Kalenda et al., 2018, Paasivaara et al., 2018, Sommer, 2019]. Literature on the benefits of LSA transformations also exists and certain performance benefits were found on multiple occasions [Petersen and Wohlin, 2010, Laanti et al., 2011, Putta et al., 2018, Laanti and Kettunen, 2019]. This has resulted in some academic understanding of the impact of LSA transformations on organizational performance. What seems to be missing from this body of academic literature is research on the impact on the portfolio level. Academic literature on agile portfolio management in general is also scarce [Sweetman and Conboy, 2018]. This leads to the research question of this thesis:

What performance benefits do organizations experience as they progress in LSA maturity across different organizational levels?

This research question consists of two components, the first component is measuring organizational performance in a context of LSA transformations. To be able to do this as many metrics as possible from already existing literature on the impact of LSA transformations will be reviewed and compiled. At the same time this thesis will aim to extend this collection of metrics by adding metrics that pertain to the portfolio level. With the hope of assisting in closing the research gap surrounding the impact of LSA transformations on the portfolio level. Doing so a framework of organizational performance in the context of LSA transformations will be created. This framework can then be tested iteratively through academic and practitioner feedback. Before arriving at the final framework, which can be used in a survey to collect data on organizational performance. This data can then be compared to results previously obtained in literature.

The second component is measuring transformation maturity. This will be done with the maturity model by Laanti [2017]. This model makes a distinction between three maturity levels for three organizational levels: portfolio maturity, program maturity and team maturity. This will allow a test to see if portfolio maturity impacts portfolio metrics more than other metrics.

Both of these will be useful tools to measure the constructs that make up the research questions. But since the validity of the results is dependent on the validity of these tools, some additional testing of these tools will also be done. For the performance framework the correlation between metrics will be analyzed. This will be done by calculating the Pearson coefficients. For the maturity model the internal structure of the model will be analyzed. This will be done by using a graded response model and comparing its results to the original model.

[1] https://scaledagile.com/what-is-safe/scaled-agile-benefits/

2 Background

In this section a description about the history of LSA will follow, both as a practice and as a subject of academic research. Within academic research, maturity models are especially relevant to this thesis and these will be discussed separately. Afterwards portfolio management will be discussed, both in the context of LSA transformations and in its more traditional form.

2.1 Large-scale Agile

Early agile frameworks such as Extreme Programming (XP) [Wells, 2013] and Scrum [Schwaber and Sutherland, 2020] often rely on small team sizes where team members would ideally all be in the same room, employing agile principles at a large scale proved difficult [Dyba and Dingsoyr, 2009]. Despite these difficulties various companies such as Nokia [Laanti et al., 2011] and Ericsson [Petersen and Wohlin, 2010] tried to integrate agile principles throughout their entire organization. Initially these transformations were done by trying to apply principles of established frameworks on a larger scale; while some benefits were identified, both Petersen and Wohlin [2010] and Laanti et al. [2011] also found some major issues and challenges such as switching from a waterfall type planning to an iterative type planning. At the 2010 XP conference there was a vote amongst practitioners, here they could vote for which topics they would like to see more research on. The result was that "Agile and large projects" was voted number one [Freudenberg and Sharp, 2010], later in the 2016 edition of the annual State of Agile report [Inc., 2016] 62% of the respondents had over a hundred employees. The need for more practitioner and academic literature on LSA was becoming more and more apparent.

2.1.1 Large-scale Agile Frameworks

Around 2013 the first LSA frameworks started going public, such as SAFe [Scaled Agile, 2021] and DAD [Ambler, 2012]. These frameworks dealt specifically with the problems of applying agile on a larger scale, providing guidance on issues such as having multiple agile teams working on the same product, how to apply agile on the product and portfolio levels of an organization, new roles etc. Since the creation of these frameworks more LSA frameworks have popped up, the following frameworks were used by at least 1% of the 2021 State of Agile respondents [Digital.ai Software Inc., 2021]: Scrum of scrums [Sutherland, 2001], Enterprise Scrum [Greening, 2010], Spotify Model [Kniberg and Ivarsson, 2012], Agile Portfolio Management [Krebs, 2008], Large Scale Scrum [Larman and Vodde, 2013], Nexus [Schwaber, 2018], Lean Management [Arnheiter and Maleyeff, 2005] and Solutions for Agile Governance in the Enterprise (SAGE) [Cprime, 2021]. These frameworks vary in their design and execution. For a more in depth analysis on the overlap and discrepancies between the frameworks see Alqudah and Razali [2016], who analyzed some of these frameworks extensively. In their analysis they noted differences such as variations in team sizes, available training in frameworks, methods and practices adopted, technical practices required and intended organization types.

On the team level these frameworks can adhere to the original agile principles [Fowler et al., 2001] and the frameworks that are used to integrate the principles with a team such as Scrum, which is a practice found in all the LSA frameworks researched by Alqudah and Razali [2016]. On the product and portfolio level there were no such available principles and practices [Laanti, 2014, Dingsøyr and Moe, 2014]. This is the gap that LSA frameworks try to fill, providing guidance on how to experience agile benefits in an organization that encompasses and manages multiple value streams.

2.1.2 Large-scale Agile in Academic Literature

The earliest available academic literature on LSA predates many of the existing LSA frameworks and widespread adoption of all such frameworks. As such they do not discuss specific LSA practices, rather they discuss how effective agile can be in a large scale setting and what pitfalls an organization may experience when applying agile at a large scale. Both Petersen and Wohlin [2010] and Laanti et al. [2011] found that employees report having issues with sprint planning, experiencing difficulty in prioritizing the backlog and having to wait on a new release cycle when a deadline is not met. Such troubles with implementing agile on a large scale were also voiced by practitioners, who voted "Agile and large projects" as their number one requested research topics at the 2010 XP conference [Freudenberg and Sharp, 2010]. However some benefits of agile were still present even at a larger scale. Both Petersen and Wohlin [2010] and Laanti et al. [2011] reported earlier fault detection, Laanti et al. reported more effective development, Petersen & Wohlin found it leads to less reworks, and employees reported on various other benefits. These benefits in combination with markets that were changing faster than ever made integrating agile practice an attractive proposition even for large enterprises [Laanti, 2014].

Since then there has been a substantial increase in LSA adaption, over half of the 433 respondents (across 359 different organizations) of the Business Agility Institute [2020] was from companies that have 200 FTEs or more. As the interest for LSA increased amongst practitioners, so did the interest in academia. Dikert et al. [2016] performed a systematic

literature review and went through 52 papers on LSA, with the aim of compiling all the reported challenges and succesfactors in literature. This compiled list can then provide practitioners with an extensive overview of succesfactors and challenges, which they can use whenever they are involved in such a transformation. Many papers around that time shared the same focus of identifying challenges and succesfactors in LSA transformations to be able to better guide organizations: Dingsøyr et al. [2018] developed a model to help practitioners with very large programs, Paasivaara et al. [2018] further analyzed Ericsson's transformation through interviews and created a table of challenges and how to mitigate these challenges, Kalenda et al. [2018] used their own focused literature review to conduct action research on the challenges and succesfactors, more papers researching how to implement agile on a large scale exist still [Paterek, 2017, Sommer, 2019, Uludağ et al., 2019].

Around the same time more literature researching the benefits of LSA transformations started appearing. In their systematic literature review Putta et al. [2018] were able to find six scientific papers reporting on performance benefits and combined this with gray literature (often reported by the SAFe framework itself) to compile a list of reported benefits. Expanding their research from a single company with various organizations within Finland, Laanti and Kettunen [2019] used an open question within their survey to ask about benefits of SAFe adoption, this resulted in a long list with transparency being by far the most reported benefit. In their closed question survey Putta et al. [2021] found similar results, with improved collaboration, dependency management and transparency being among the top reported results. Stettina et al. [2021] used Laanti et al. [2011] as a baseline for their own survey and compared the results against the reported benefits by SAFe[2], showing that on average organizations experience even larger benefits than SAFe is reporting. Deviating from Laanti et al. [2011]'s earlier established survey method by employing a slider scale rather than a Likert scale. Olszewska et al. [2016] took a Goal-Question-Metric (GQM) approach to define eight metrics that can be used in an LSA context to measure performance and also showed that most of these metrics behaved as expected when taking into practice. Even more research on the performance benefits of LSA exists, such as Gustavsson and Bergkvist [2019].

From this collection of research on the impact of LSA transformations some patterns can be observed. One such pattern is that overall LSA transformations do seem to have a positive impact on performance. Another pattern is that there are some metrics that consistently rank amongst the most impacted, such as collaboration, transparency. At the same time studies often use different sets of metrics, adding difficulty to making direct and complete comparisons. This issue of picking a set of metrics to measure an LSA transformation can also be seen in the SAFe customer stories[3], where SAFe shares the experienced impact of successful transformations. In these customer stories many different metrics can be found, often only used once for a specific case. This is not strange, as different transformations have different objectives. Nonetheless, this lack of a shared set of metrics to measure the impact of LSA transformations, does create some problems. The added difficulty to comparisons is an obvious problem, but a framework for measurement can also help practitioners measure their transformation.

2.1.3 Maturity Models

LSA transformations often take years to complete and it is not always obvious in which stage of the transformation an organization is. To help organizations understand how far along they are and to further steer their transformation, they can use a maturity model. Maturity models can help an organization by providing them a tool to benchmark their transformation. This benchmark can then be used to get an overview of what objectives have been obtained and what objectives still need to be obtained. In turn this insight on transformation maturity can provide an overview on what benefits organizations should already be experiencing based on where they are in the maturity model. If an organization is not experiencing certain benefits that are usually associated with their maturity, the organization might want to reasses the transformation process so far and take action to still be able to experience these benefits.

For research into performance benefits, these maturity models allow us to see when organizations start to experience benefits. The model by Laanti [2017] is especially interesting for the purposes of this thesis because of the distinction it makes between portfolio maturity, program maturity and team maturity. Each organizational level has its own independent stages of maturity, starting from the lowest to the highest level we have: beginner, novice, fluent, advanced and world-class. A level is defined by a combination of practices such as "Test-First Approach" and milestones such as "Production code practically error-free", if an organization wants to assess their maturity using this model they simply find the definition that best suits their current state for each of the three organizational levels. For the full model see figure 1. Assuming such a model measures maturity accurately, this distinction can help determine if the portfolio metrics used to measure portfolio performance correlate with portfolio maturity. If this is the case this would add to the construct validity of these metrics as a tool for measuring portfolio performance.

[2]https://scaledagile.com/what-is-safe/scaled-agile-benefits/

[3]https://scaledagile.com/insights-customer-stories/

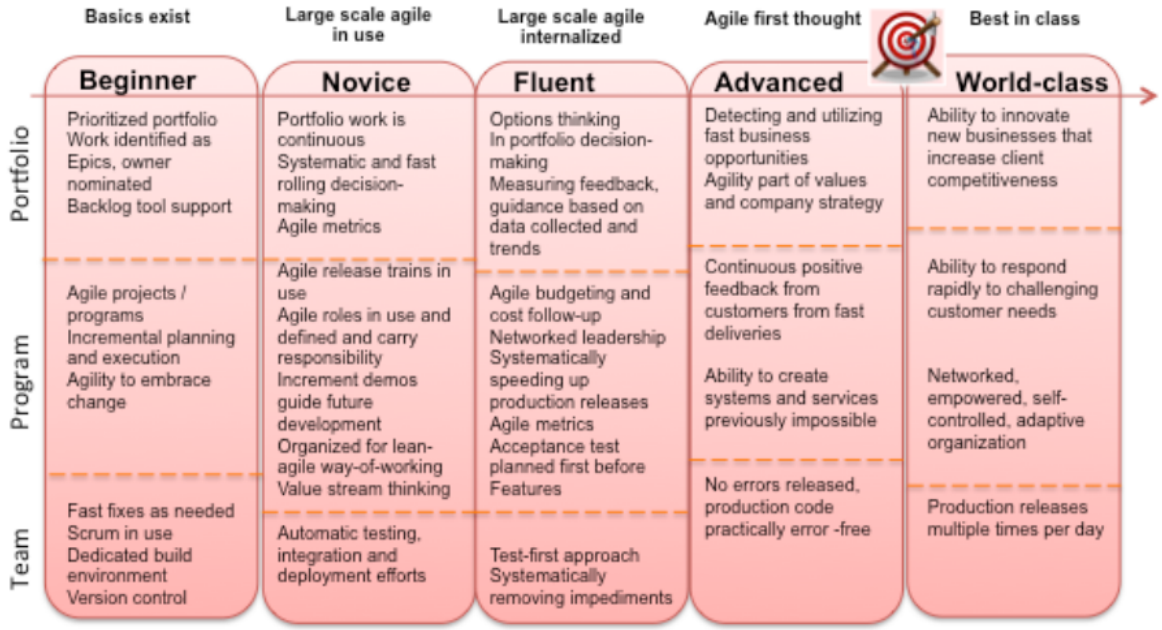


Figure 1: The maturity model by Laanti [2017] will be used to measure maturity

One other maturity model has been considered for the purposes of this thesis, the model by Turetken et al. [2017]. This model takes a different approach to defining LSA transformation maturity. They base their model on a previously made agile maturity model which looks at the principles and values of agile, these values are categorized and prioritized to form the dimensions of the model. The five levels of maturity are based on the most important aspects of agile, meaning that the first level concerns itself with practices surrounding the most important aspect of agile: collaboration. As you progress in levels you are incorporating practices that deal with other aspects of agile such as evolutionary development and efficiency. The second dimension consists of five categories such as human centricity, technical excellence and customer collaboration, each category representing a category of agile principles. The levels across the categories are defined by practices. To expand this model and make it work in a scaled context, Turetken et al. [2017] expanded the practices of certain levels and categories with practices that were specific to LSA. Ultimately this model could not be used as it focuses on only the team of specifically the SAFe framework, making it not as good a fit for studying the relationship between e.g. portfolio maturity and portfolio performance benefits.

2.2 Portfolio Management

The existing research on the performance benefits of LSA transformations shows the beginning of a performance benefits framework. Some benefits such as increased collaboration and transparency regularly find themselves amongst the most reported benefits [Laanti et al., 2011, Putta et al., 2018, Stettina et al., 2021]. But despite the recent increase in academic literature surrounding LSA, there still exists a literature gap around the impact on performance at the portfolio level of organizations. This while some LSA frameworks such as SAFe do include specific modules addressing agile portfolio management [Scaled Agile, 2021].

Portfolio management concerns itself with maximizing the value of the portfolio in terms of company objectives, achieving a balance of projects in terms of strategically important parameters or ensuring strategic direction of projects [Martinsuo and Lehtonen, 2007]. If an organization does not manage its portfolio effectively, it risks missing out on new value streams, a loss of quality on established value streams, an increased time-to-market and more [Cooper and Edgett, 2003]. It is in a company's best interest to have a good understanding of: their portfolio, how it aligns with the business strategy, if it reflects the available resources, if the prioritization is sound etc. Effective portfolio management is no trivial task and there are many pitfalls for management [Cooper et al., 2000].

2.2.1 Agile Portfolio Management

Some LSA frameworks contain principles and practices that can help guide management in effective portfolio management. Despite the apparent need for literature on agile portfolio management, there exists little academic literature on

this subject. Papers on agile portfolio management do exist but, with the aim of this thesis being measuring portfolio performance no relevant articles were found for various reasons. Cooper and Sommer [2020] do address the problems that arise when trying to apply traditional portfolio management in an agile organization, stating that the high uncertainty of agile products makes it harder to assess their economic value. They go on to discuss how traditional metrics of economic value can be adjusted to be more fitting of an agile context, allowing organizations to execute a stage-gate way of portfolio management. Outside of these economic measurements, Cooper and Sommer [2020] provides little insight on measuring portfolio performance.

Similar papers that discuss existing problems with agile portfolio papers were also found. While these papers are valuable, they do not provide the necessary knowledge required for this thesis. Sweetman and Conboy [2018] acknowledges the lack of literature on agile portfolio management and propose a framework for effective portfolio management. Lappi et al. [2018] seek to better understand how agile products are currently governed by comparing them to traditional governance dimensions and finds a gap in long term product governance. Niederman et al. [2018] propose a framework for surfacing and discussing issues surrounding the lack of theoretical basis amongst agile practices, although not specifically targeting the portfolio level the resulting framework does include it.

The paper by Drake et al. [2013] does contain a section which talks about measuring the leanness and agility of certain components. They talk about quality and cost being factors of leanness and flexibility and time being factors of agility. In turn each factor has its own three metrics addressing different aspects of these factors. Although the format fits very well with the subject matter of this thesis, the paper talks about physical components exclusively. This limit to physical components also becomes apparent in the individual metrics of the factors, such as "inventory cost" being an indicator of the cost factor. Such metrics often do not apply in an IT context, which is still one of the largest sectors when it comes to agile practices [Digital.ai Software Inc., 2021]. This means that these suggested measurements can not be used either.

2.2.2 Traditional Portfolio Performance metrics

To not be entirely reliant on gray literature and since there exists a literature gap when it comes to measuring performance of agile portfolio management, another source of knowledge surrounding this topic is necessary. Therefore, instead of using academic sources about measuring performance of the portfolio in an agile context, sources about measuring performance of the portfolio in general will be used.

As suggested by the papers addressing the challenges of implementing traditional portfolio management practices in an agile context (see section 2.2.1), portfolio management in general is a much discussed topic in academia. A lot of the papers on portfolio management can be traced back to a series of publications around the year 2000 done by Cooper et al. These publications cover a wide range of subjects, various methods for portfolio management are proposed such as the strategic bucket [Cooper et al., 1997] and the Stage-Gate [Cooper et al., 2000] methods. Companies were analyzed based on their portfolio management practices and stance of management on portfolio management. Showing that multiple portfolio decision making tools and engagement by management into the decision making is important for organizational performance [Cooper et al., 1999]. Research was done on the detriments of poor portfolio management. Showing that it does not only affect the portfolio level, but can also affect the lower levels of the organization [Cooper et al., 2000, Cooper and Edgett, 2003]. While these papers contributed greatly to the general base of knowledge of portfolio management around that time, the most relevant findings of these publications, given the context of this thesi,s are the objectives of portfolio management in Cooper et al. [1997] and the ways of measuring portfolio performance discussed in Cooper et al. [1999, 2004a] and various others.

In Cooper et al. [1997] they observe three main goals of portfolio management:

1. Maximizing the value of the portfolio against an objective such as profitability;
2. Balancing the portfolio in terms of risk, markets, attractiveness etc.;
3. Link the management to the strategy like in terms of resource allocation.

To answer the research question, portfolio performance needs to be measured. A definition of what constitutes portfolio performance is also required. These three goals given by Cooper et al. [1997] provide an initial division of the various facets of portfolio performance. Serving as a lens to look through when searching for metrics of portfolio performance in literature. Most notably in Cooper et al. [1999] and in Cooper et al. [2004a,b], they present various metrics for measuring portfolio performance such as: the alignment between spending and strategy, percentage of projects meeting objectives and portfolio reflecting available resources. These metrics fall nicely within the three main goals and the combination of the goals with the metrics serve as a base of knowledge. Although useful, it is important to remain critical of these goals and metrics and to make sure to adjust them when newly found information demands it. This

process will be further discussed in section 3.1, the full list of metrics used from Cooper et al. [1999, 2004a,b] (and others) as well as the decision making process will be covered in 3.1.2.

Cooper et al. are not the only authors in academia to write about portfolio management, various other publications have surfaced since Cooper et al.'s series of papers. Looking at these other publications, it becomes evident that they base their work at least partially on the work of Cooper et al. Especially so within the subset of sources that look into measuring portfolio performance. This can be seen in the citations and in the metrics that are used within these papers. Many of which can be categorized using the three main goals of portfolio management. Meskendahl [2010] proposes a framework to assist firms in implementing their strategies. To test this framework they use a definition of portfolio success based on four dimensions, with strategic fit falling in the third goal and portfolio balance in the second goal. As such half of these dimensions fit within the three goals framework.

One of the hypotheses in the survey research by Killen et al. [2008] also requires the measurement of portfolio performance. Many survey questions fit the concept of the three main goals: "Our projects are done on time – in a timely and time efficient fashion" belongs the first goal, "We have the right number of new product projects for our resources – people, time and money – available" to the third, "Our portfolio of new product projects has an excellent balance in terms of long versus short term, high versus low risk, across markets and technologies" to the second and others fit the concept similarly. More such papers exist such as Martinsuo and Lehtonen [2007], Thornley [2012] and Müller et al. [2008] but for the sake of brevity we will not go over these in detail in this section.

Although many of the metrics used in publications adhere to the three main goals by Cooper et al. [1997], there were also some that did not. Furthermore, the three goals are broadly defined and encompass a lot of different aspects of portfolio management. If the framework used in a publication did adhere to the three main goals, they were often still able to address a new concept. Hence why papers both by, and not by Cooper et al. were still able to add to the understanding of measuring portfolio performance.

3 Methodology

To answer the main research question of this thesis, an appropriate research method has to be found. For related literature that also researches the impact of large scale agile transformations, it is not uncommon to see the use of a survey[Laanti et al., 2011, Gustavsson and Bergkvist, 2019, Putta et al., 2021, Stettina et al., 2021]. This can be done either through a self-administrated or an interviewer-administered questionnaire. Alternatives to survey based research are a quantitative Goal-Question-Metric approach [Olszewska et al., 2016] or a qualitative interview approach [Petersen and Wohlin, 2010, Pries-Heje and Krohn, 2017]. As this research further builds upon the results of Laanti et al. [2011] and Stettina et al. [2021], a similar research method will be used: a web-based closed-ended survey. By using the same method as other research, parts of previously held surveys can be reused. This increases reliability and external validity [Linåker et al., 2015].

To create a survey that answers the research question "What performance benefits do organizations experience as they progress in LSA maturity across different organizational levels?" internal questions will be formulated that align with the research objective. Internal questions are open-ended questions that represent the main research question and can help in forming the eventual survey [Linåker et al., 2015]. The research question can be split into two components that need to be measured to find an answer: the experienced impact on performance and the transformation maturity. These can be further translated into internal questions: "What impact on performance do organization's experience as a result of LSA transformations?" and "How can LSA transformation maturity be measured across different organizations?". For the former there exists a collection of literature that has previously researched the benefits of LSA transformation, from which a performance framework can be constructed. For the latter there exists various maturity frameworks of which the framework by Laanti [2017] will be used, as discussed in section 2.1.3.

Constructing this performance framework is undoubtedly paramount to answering the research question, as it will be the instrument by which performance will be measured. Therefore, if the framework is not sound, the construct validity of this research is at risk. The construction of this framework will be discussed in detail in section 3.1. This framework will have to be translated to survey questions, which will measure the performance impact as well as the LSA maturity. The translation of the performance framework and maturity model into survey questions will be discussed in section 3.2. How the final results will be discussed in section 3.3.

3.1 Creating the Performance Framework

To make sure that all aspects of organizational performance were covered in the final framework used in the survey, academic sources that discuss performance benefits of LSA and portfolio management need to be identified. After collecting these sources a decision had to be made as to which metrics had to be included and which had to be omitted, resulting in a list of metrics categorized in various dimensions. After compiling this list it was clear that it contained a lot of redundancy and many oddities, meaning that the list could easily be refined. After this initial refinement the remaining list of metrics was presented to various practitioners and academics for feedback, this feedback was then compiled and incorporated to arrive at the final version. Each of these four steps will be covered here in their own section: finding sources in section 3.1.1, initial selection in section 3.1.2, initial refinement in section 3.1.4 and finally the feedback sessions in section 3.1.5.

3.1.1 Finding Performance Benefits Sources

As discussed in section 2.1.2, literature surrounding LSA focuses on various subtopics. For the purposes of creating an organizational performance framework in the context of LSA, only sources that include performance benefits are of interest. The starting point for relevant literature was Stettina et al. [2021], Laanti et al. [2011] and Olszewska et al. [2016]. From this starting point a combination of manually looking at used references, connected papers and academic databases (Leiden University's E-Catalogue, and Google Scholar) were used to find additional papers. The following queries were used to find appropriate papers:

- "software development" OR "IT organization" AND "performance";
- "IT" AND "organization" IR "performance matrix";
- "measuring" AND "software" AND "performance" OR "success";
- "agile" AND "impact" OR "relation" AND "innovation";
- "agile" AND "innovation".

When searching for new sources, regardless of how the source was discovered, the same process was followed. First the title was inspected to see if it was related to agile working methods. If this was the case the abstract and conclusion

were read to see if it concerned itself with the performance benefits of agile working methods. Finally if this would also hold up, the entire source would be read. The search for new sources was stopped once the listed tools did not seem to result in new, relevant and accessible sources. This process resulted in ten sources that reported performance benefits from agile practices were found.

The process of searching for sources that researched performance benefits on the portfolio level used the same tools. Initially the goal was to find relevant literature through academic databases by combining search terms such as "agile" and "portfolio". This was unsuccessful due to the research gap that exists on agile portfolio management, as discussed in section 2.2.1. Afterwards the goal shifted to find literature within the broader context of portfolio performance benefits, such as the series of papers published by Cooper et al. From there a combination of Connected Papers and manually looking at references was used to find additional sources. This resulted in eight sources that reported performance benefits from portfolio management practices, for a total of nineteen sources.

3.1.2 Selecting metrics

Through this search there now is a collection of literature on the impact of LSA transformations. From his collection a selection of metrics has to be made that should be included in the performance framework. Previous papers that exist and discuss organizational performance both at the team, product and the portfolio level have been briefly mentioned (section 2.1.2). In this section a description will follow on the construction of the initial version of the performance framework. What metrics were used and which were not, will be discussed in this section per paper. As well as the underlying reasoning behind these decisions.

For readability's sake, the complete lists of metrics obtained at this stage are shown separately. This is done in table 1 for the team and product level metrics, and table 2 for portfolio level metrics. It is important to note that at this stage the only concern is to collect impact metrics from literature, with no further consideration. Metrics may seem very similar, not fit for survey research, ambiguous in their wording etc. Further refinement of the framework will be discussed in section 3.1.4.

Stettina et al. [2021] provides a baseline performance framework. All the academic sources from the comparison table from Stettina et al. [2021] will be used, that is to say, all metrics except those from SAFe itself. The metrics included in their survey research are based on earlier reported benefits from Laanti et al. [2011]. They took these metrics and grouped them by the dimensions of performance, which they do to be able to compare the results to results reported by SAFe. In this table they make one other comparison to Olszewska et al. [2016]. The metrics used in this study differ from those in Stettina et al. [2021] and Laanti et al. [2011], but do still fit within the same dimensions. As a baseline, all metrics and dimensions from the table by Stettina et al. [2021] will be the context in which other metrics from other papers will be looked at.

Putta et al. [2018] did a systematic literature review on the benefits and challenges of adopting SAFe. Their literature review used both whitepapers from SAFe itself and academic sources. Although a case is made for also including gray literature in the paper, only benefits reported in academia were considered to avoid any bias that might come from using benefits that are reported by SAFe about SAFe.

In a follow-up study Putta et al. [2021] did another survey research on the benefits and challenges of adopting SAFe across many companies in different areas of the world. Using a five-point Likert scale, they looked at a total of eleven benefits. The resulting means of all these benefits were above the median of three, with the lowest mean being 3.21. Because of the positive results all the benefits are selected for further use.

In their survey to better understand the current state of SAFe adoption in Finland, Laanti and Kettunen [2019] looked at various things including impact. Using an open text questionnaire they got 47 replies about SAFe adoption benefits, resulting in over 20 reported impact metrics. Since many of these metrics only have very few participants mentioning them, a threshold of 10% (or five or more mentions) will be used to determine which benefits to continue with.

Gustavsson and Bergkvist [2019] research SAFe adoption at three different companies in Sweden from three different sectors. With an open text question they asked about the benefits and drawbacks and coded the resulting answers with the help of the key areas found by Laanti et al. [2011]. They distinguish the results between a group that has indicated to feel positively about the adoption and a group that has indicated to feel negatively about the SAFe adoption. Participants could also answer thay they felt neutral about the adoption but these responses were left out of the analysis by the original authors. Interestingly enough the top three reported benefits between these groups are identical. They present their results, of how often benefits were mentioned by participants, in a table of percentages. Since the method used is also open ended, a threshold of 10% is applied again. Meaning that any benefit that is mentioned by at least 10% of either the positive or the negative group is selected.

Table 1: **Team and product** metrics as they were presented in their original papers, per paper

Paper	**metrics taken from paper**
Lee and Xia [2010]	Software team response extensiveness, software team response efficiency, on time completion, on budget completion, software functionality
Petersen and Wohlin [2010]	More stable requirements led to less rework and changes, everything that is started is implemented, estimations are more precise, early fault detection and feedback from test, lead time for testing is reduced, moving people together reduced the amount of documentation
Laanti et al. [2011]	Increases effectiveness of development, increases quality of product, incnreases the transparency of development, increases collaboration, makes work more fun, makes work more planned, increases the autonomy of development teams, enables earlier detection of bugs/errors/defects, makes work less hectic
Olszewska et al. [2016]	Features per money spent, customer service request turnaround time, lead time per feature, number of external trouble reports, average number of days open of external trouble reports, number of realeses per time unit, number of days between commits
Recker et al. [2017]	Software team response extensiveness, software team response efficiency. customer satisfaction, process performance, software functionality
Putta et al. [2018]	Focus on continuous improvement, increased alignment between teams, enhanced collaboration, improved dependency management, cross team dependencies are transparent, more self organizing teams, improved visibility, improved morale, improved employee satisfaction
Gustavsson and Bergkvist [2019]	Requirements / goals / planning, visbility / overview / transperency, people / communication / collaboration, productivity / focus / efficiency, dependencies / co-operation
Laanti and Kettunen [2019]	Transparency, co-operation, cadency and rhythm, better speed, continuous improvement
Business Agility Institute [2020]	Collaboration & communication, better ways of working, speed to market, customer satisfaction
Putta et al. [2021]	Improved collaboration between teams, improved dependency management between teams, improve transparency, enable faster feedback, have more frequent deliveries, improve customer satisfaction, have shorter time to market, increase delivery predictability, increase responsiveness, improve team autonomy, improve software quality
Stettina et al. [2021]	Increases effectiveness of development, improves time-to-market, increases quality of product, enables earlier detection of defects, makes work more planned, makes work more organized, makes work more fun, makes work less hectic, increases the autonomy of development teams, increases collaboration, increases the transparency of development

Petersen and Wohlin [2010] take an early look at LSA by looking at agile practices adoption within Ericsson, a Swedish networking and telecommunications company. As part of their case study the authors used interviews to ask about improvements experienced by employees and categorized coded answers based on how common they were. For example, any issue that was mentioned by more than $1/10$ of the interviewees is deemed "common improvements". For the purposes of this thesis, this definition will be adhered to. All improvements mentioned by at least $1/10$ of the interviewees are included in the selection, staying consistent with the selection process used on other open ended research methods (10% of the respondents). Ericsson also kept track of a requirements waste and maintenance cost indicator, both of which improved with the adoption of incremental methods. These metrics will also be used.

Lee and Xia [2010] looked at the relation between software team characteristics and software development agility. In turn they studied if this impacted software development performance. They sent out a survey to North American project managers who had recently managed a software development project. With 399 respondents, they were able to confirm many of their hypotheses. Most importantly to the context of this thesis: both agility metrics had a positive relation to at least one performance metric. As such all used agility and all performance metrics will be used.

Using a very similar model Recker et al. [2017] research the impact of agile practices on the customer of one large anonymous organization. They take individual agile practices such as stand-up meetings, collective code ownership and

pair programming, and see how they impact agility and in turn how this agility impacts development success. They found that all practices impact either aspects of agility or success, albeit not always directly. Therefore the paper's aspects of agility and aspects of success wil be used further.

The Business Agility Report [Business Agility Institute, 2020] is not an academic source but rather an annual report on the state of agile in businesses across the world. They look at the different industries in which agile is being employed, which regions are adopting agile practices at which rate, company sizes etc. As part of their report they also ask businesses what they think the most significant organizational benefit of business agility is. They publish the top ten but unfortunately without exact figures, only a horizontal bar-chart illustrating roughly how frequently some benefits were reported over others. In their 2020 edition they also presented this top ten and wrote an additional paragraph on every benefit in the top four. No further explanation is given as to why only the top four were discussed further, but to adhere to the process of the original authors the same will be done in this thesis.

On the portfolio level the biggest contribution comes from the series of papers by Cooper et al. Although they do not establish a framework like the one in Stettina et al. [2021]. In Cooper et al. [1999] they look at 205 US based companies and group them by management's view on the portfolio into one of four categories. They then find that companies where management takes the portfolio more serious also score better in six performance goals for portfolio management. All six performance goals that were mentioned were taken as metrics of portfolio management performance, as they all performed in accordance with the hypothesis.

Here Cooper and Edgett [2003] take a different approach and discuss six common ailments of poor portfolio management based on previous research. This also implies that proper portfolio management results in improvement in these six areas, which is why these inverse metrics were included for further analysis. These ailments are identified mostly from data collected in previous study by Cooper et al. such as Cooper et al. [1999].

Cooper et al. [2004a] presents the most extensive list of portfolio performance measures in the series. By using both qualitative on-site research and quantitative survey research they look into the performance of various companies across different industries. In doing so they look at what portfolio performance metrics are used by practitioners as well as gauge how companies perform according to metrics from literature. This difference in intent and approach results in different sets of portfolio performance metrics, some of which are better suited for the purposes of this research than others. For example, the authors point out a flaw with one of the most popular metrics amongst practitioners: percentage of sales of new products is a popular metric because it will lead to more short-term projects and little long-term ones. As such metrics that could only be found in the results of popular practitioner metrics were omitted, metrics that were used to measure portfolio performance by the authors themselves were included.

These are the only papers from the series by Cooper et al. that were used for the purpose of collecting metrics in this thesis. More papers do exist but these were excluded either because they contained the same exact metrics as other papers [Cooper et al., 2001, 2004b] or because they do not contain any portfolio performance metrics at all [Cooper et al., 1997, 2000].

Killen et al. [2008] test three hypotheses in their survey research, two of which concern themselves with portfolio performance. They use a five-point Likert scale to measure the impact of portfolio performance on product success (hypothesis two), and the impact of various portfolio management methods on different dimensions of portfolio performance (hypothesis three). Their first hypothesis does not concern itself with portfolio performance. The survey was distributed amongst sixty Australian organizations from varying industries. The metrics used to test hypothesis two align with those from Cooper et al., as do the results. Those belonging to hypothesis three seem to deviate further from metrics from Cooper et al. Their results and in depth expalantions are not included in the paper either. Due to the positive results they will be included in further steps.

Müller et al. [2008] is another paper that also looks at the effect of various portfolio management methods on portfolio performance. However they also research if this relationship is moderated by various contextual factors such as geography, industry, governance type and others. They used data collected from Blomquist and Müller [2006], where a five-point Likert scale worldwide survey was distributed to collect 136 responses. They do split portfolio performance into three dimensions but these categories do not match those of Cooper et al. [1997], they are: achieving results, achieving purpose and balancing priorities. In turn each dimension contains its own metrics, and all metrics were considered for further use.

Martinsuo and Lehtonen [2007] research the role of single-project management on portfolio performance. They also use a survey with a five-point Likert scale to ask participants from 279 firms about the dimensions of perceived portfolio performance. The authors state that this survey is based on Cooper et al. [1999] but direct parallels between the two studies are not immediately apparent. All metrics that were used for portfolio management were included due to their apparent correlation to portfolio success.

Table 2: **Portfolio** metrics as they were presented in their original paper, per paper

Paper	**metrics taken from paper**
Cooper et al. [1999]	Having the right number of projects in the portfolio for the resources available, avoiding pipeline gridlock in the portfolio-undertaking projects on time and in a time-efficient manner, having a portfolio of high-value projects (or maximizing the value of the portfolio)-profitable/high return projects with solid commercial prospects, having a balanced portfolio (long term versus short term / high risk versus low risk / across markets and technologies), having a portfolio of projects that are aligned with the business's strategy, having a portfolio whose spending breakdown mirrors the business's strategy and strategic priorities
Cooper and Edgett [2003]	Quality of execution suffers, vital activities don't get done, time-to-market lengthens, game-changers are missed, active projects are dumbed down, the project team's morale suffer
Cooper et al. [2004a]	Success / fail / kill rate of projects, percentage of revenue and profits coming from new projects, percentage of projects that are on time and on budget, percentage of projects meeting financial objectives, time-to-market, speed & efficiency, profitability versus spending, met profit objectives, overall profitability versus competitors, technical success rating, reduction of cycle time, opened up new markets, entered new product categories, integrated new scientific knowledge, entered new technologies
Martinsuo and Lehtonen [2007]	The objectives of projects are aligned with strategy, company strategy is realized well by the project entity, resource allocation to projects is aligned with strategy, portfolio management supports the strategy process excellently, priorities across projects are known, the project entity yields an optimal return, portfolio management is efficient, portfolio management focuses on the right issues
Killen et al. [2008]	The projects in our portfolio are aligned with our business objectives and our business' strategy, our portfolio of new product projects contains only high value ones to our business - profitable high return projects with solid commercial prospects, the breakdown of spending in our portfolio of projects truly reflects our business strategy, our projects are done on time / in a timely and time efficient fashion, our portfolio of new product projects has an excellent balance in terms of long versus short term / high versus low risk / across markets and technologies, our new product program develops our existing technologies and technological competencies, Our new product program brings new technologies to our organization, Our new product program leads our organization into new product arenas, Our new product program enables our organization to enter new markets. All metrics for hypothesis two and three were included
Müller et al. [2008]	Customer satisfaction, time, cost and quality results, financial results, user requirements, projects purpose, program purpose, resource turnover, timely accomplishments of programs, stakeholder satisfaction
Meskendahl [2010]	average single project success, use of synergies, strategic fit, portfolio balance.
Thornley [2012]	alignment of programs to business-unit strategic goals, projected future income from program road map, program portfolio distribution, external customer satisfaction, percentage of the program milestones accomplished, alignment between spending and portfolio priority

Thornley [2012] take a similar approach to Olszewska et al. [2016], in that they take a more quantitative approach and try to define how portfolio performance can be measured outside of a self-reported context. However they do not do their own research and only apply some of their proposed metrics on projects previously covered in academia. Nonetheless the elaborate description on the calculation of various metrics does add another perspective. As such all metrics from this paper were included.

Meskendahl [2010] perform a systematic literature review to look at the relation between business strategy and project portfolio as well as the success thereof. Herein they also refer to the three main goals of portfolio management by Cooper et al. [1997], as well as the influence it has had on other papers. However they split the first goal into two separate goals resulting in four metrics, all of which are used in further steps of this research. This is done because they all seem rooted in literature, both by Cooper et al. (e.g. [Cooper and Edgett, 2003]) and by other authors (e.g. [Martinsuo and Lehtonen, 2007]).

3.1.3 Eliminating Similar Metrics

Although the set of collected metrics from literature as presented in table 1 and in table 2 is extensive, it is problematic in other ways. One such way, is the many metrics that measure the same aspect of a dimension, such as the seeming similarity between "Increase collaboration" from Stettina et al. [2021] and "Co-operation" from Laanti and Kettunen [2019]. To get a more accurate view of what benefits have been reported in literature, and also to see how often they have been reported, similar metrics will be treated as if they are the same metrics. This was done if two or more metrics measured the same aspect of a dimension. For some of the collected metrics, identifying this similarity is trivial. For others a closer inspection of the source material is required. In this section all metrics that have been merged into another metric will be discussed per paper, to illustrate the underlying thought process. The resulting tables of this process can be seen in table 3 and table 4.

One source has been omitted entirely from further analysis, that source being Müller et al. [2008]. Initially this source seemed to present an interesting contrast to the often reported benefits of Cooper et al. [1999] and related papers. It did this by using mostly different metrics and categorizing them by components (see table 2). However the metrics used are often vague such as the "achieving purpose" component consisting of "project purpose" and "program purpose". It is not clear what "purpose" is referring to in this context and no further explanation is given. Neither in Blomquist and Müller [2006] where the used questionnaire is discussed in detail. Despite referring to Cooper et al. [1999] and Martinsuo and Lehtonen [2007] in the context of portfolio management performance, the parallels between those studies and Müller et al. [2008] study are not immediately clear. Metrics that do not coincide with mentioned sources are also ambiguous. As such for the purposes of this thesis, Müller et al. [2008] has been deemed unfit as a source and the metrics used in this paper have not been included in the framework.

The papers Laanti et al. [2011] and Stettina et al. [2021] share many metrics with each other. This is due to the fact that Stettina et al. [2021] takes Laanti et al. [2011] as a point of comparison and largely based the metrics in their survey on those used in Laanti et al. [2011], as stated in the paper itself. Therefore if a metric is indicated as being identical in the original paper, they are also considered identical for the purposes of this thesis. Also in Stettina et al. [2021], Olszewska et al. [2016] is taken as a point of comparison but it does not share any metrics with the other two. These three are used to form an initial framework to which metrics from other papers will be compared.

Putta et al. [2021] contains various duplicates, of which two are presented in a manner that is not similar to the metric they duplicate. The "To have shorter time to market" metric is a duplicate to the "Improves time-to-market" metric used in Stettina et al. [2021]. The metric "increases responsiveness" is considered the same, as an improved time-to-market is also an increase in responsiveness [Stettina et al., 2021]. The metric "More frequent deliveries" equals the "number of releases per time unit" metric from Olszewska et al. [2016]. This paper contains many duplicates from Putta et al. [2018] as it was a follow up research with a lot of the same authors. Other metrics in either Putta et al. [2018] or Putta et al. [2021] are either presented in a similar way to other metrics or measure new aspects.

Laanti and Kettunen [2019] have metrics that for the most part resemble other metrics. The "better speed" metric is somewhat ambiguous and can most likely be seen as a duplicate of increased effectiveness or increased time-to-market. Since other metrics that refer to some type of velocity (such as faster feedback and lead time per feature) are considered to belong to the responsiveness dimensions, "better speed" is treated likewise and seen as a duplicate of the generic time-to-market metric. The paper also contains a "co-operation" metric, without further explanation this is considered to be analogous to "increases collaboration" which can be found in many other papers.

Gustavsson and Bergkvist [2019] present a challenge in the way they coded their responses into metrics. Instead of having metrics that are written down in short sentences or single nouns, they often combine various nouns into one overarching metric like "people/communication/collaboration". The metric "productivity/focus/efficiency" is considered to be the same as the "Features per money spent" metric as this belongs to the productivity dimensions and is considered an efficiency metric in its original paper [Olszewska et al., 2016]. "Requirements/Goals/Planning" is considered to be the same as "Makes work more planned". The "dependencies/co-operation" is of special interest because it contains co-operation, which in the case of Laanti and Kettunen [2019] was considered to be analogous to collaboration. Since in this case there is also a metric called "people/communication/collaboration" it seems that the authors do not consider this the same thing. Therefore "dependencies/co-operation" is considered identical to "better dependency management" from Putta et al. [2018, 2021], and "people/communication/collaboration" identical to the collaboration metric. The last

remaining metric "visibility/overview/transparency" is considered equal to the increased transparency metric reported in many papers.

Petersen and Wohlin [2010] contains three metrics that are very similar: "Everything that is started is implemented", "More stable requirements led to less rework and changes" and "Estimations are more precise". These three metrics all seem to cover various aspects of the "makes work more planned" metric originally used in Laanti et al. [2011], as a more planned work environment would impact all of these. Since no other source makes this distinction they are all considered to be identical to the planned work metric. Most other metrics cover novel aspects, with the exception of the "Early fault detection and feedback from test" metric which is similar to the early fault detection metric also from Laanti et al. [2011].

Recker et al. [2017] is based largely on the research model of Lee and Xia [2010] and therefore these two contain many similar metrics. "Response efficiency" can be found in both papers and although it specifically pertains to responses, it also pertains to the amount of resources spent per response. Therefore it is considered a duplicate of the "Features per money spent" metric. The "Software functionality" metric can seem ambiguous at first but the papers elaborate that the metric concerns itself with the alignment between the end product and the initial requirement. Although this is not a duplicate in itself the name used does not fit the measurement well which is why it is considered as "alignment between product and requirements" in this thesis. "Software team response extensiveness" can also seem ambiguous. When looking at both surveys it becomes clear that this metric refers to the percentage of requirement changes that were actually implemented, across various categories (system scope, data structure, user interface etc.). This metric is similar to the "Everything that is started is implemented" from Petersen and Wohlin [2010], and this metric is considered to measure a subset of the "work planned" metric from Laanti et al. [2011]. Therefore this metric is also considered to be a duplicate of the "work planned" metric. The "on-time completion" and the "on budget completion" metrics can only be found in Lee and Xia [2010] and these are both considered to be subsets of the "Increases effectiveness" metric.

Business Agility Institute [2020] contains only duplicate metrics but since they are presented in a conventional manner no further explanation is required to establish which metrics they duplicate.

The role of baseline framework that is fulfilled by Laanti et al. [2011], Olszewska et al. [2016], Stettina et al. [2021] for the team and product level, is fulfilled by Cooper et al. [1999, 2004a] for the portfolio level. Together these two papers contain a collection of various metrics that span all three main goals of portfolio management, as discussed in Cooper et al. [1997]. Furthermore Cooper et al. [2004a] also contains additional metrics that do not seem to adhere to the three main goals, four of which are of special interest: "Opened up new markets", "Entered new product categories", "Integrated new scientific knowledge" and "Entered new technologies". The first four of these seem to largely coincide with three other metrics used in Killen et al. [2008], where they are called "opportunity metrics". As such a new dimension for the portfolio is considered outside of the three main goals of portfolio management, the opportunity dimension. Besides these, Cooper et al. [2004a] also contains various metrics that coincide with metrics from the team and product levels. "Profitability vs spending" is very similar to the "Features per money spent" from Olszewska et al. [2016], and: "Time to market", "Reduction of cycle time" and "Speed & Efficiency" are all very similar to the "Time-to-market" metric from Stettina et al. [2021]. It is important to note that despite their similarities these metrics measure something different than their team and product counterparts. For example, measuring the time-to-market for a single value-stream mostly looks at mechanics within that value-stream. Whereas looking at the time-to-market of multiple value-streams also requires you to look at the interactions between value-streams. Finally Cooper et al. [2004a] contains one metric that does not fit any previously established dimension. The "Technical success rating" metric does not fall within the three main goals of portfolio management and does not seem to coincide with metrics from other sources on the portfolio level nor the team and product levels. Therefore this metric is for now put in a separate "Other" dimension.

Killen et al. [2008] is very verbose about the metrics they use but the metrics are in large based on the more concise metrics of Cooper et al. [1999]. The metrics "Projects align with business objectives", "Spending reflects business strategy", "Portfolio contains high value products", "Percentage of projects that are on time and on budget", "Portfolio has good balance of products" and "Portfolio reflects resources available" can all be found in this paper in a more verbose form. Additionally the paper also contains some metrics that coincide with Cooper et al. [2004a] and a novel metric about developing existing technologies and competencies within the organization.

Martinsuo and Lehtonen [2007] for the most part follows in the steps of Cooper et al. [1999], although their metrics do sometimes seem more detailed. The metrics: "The objectives of projects are aligned with strategy", "Company strategy is realized well by the project entity" and "Portfolio management supports the strategy process excellently" all seem and score very similar to each other. As such they are treated as duplicates of "Products align with business objectives" from Cooper et al. [1999]. The "Resource allocation to projects is aligned with strategy" also scores very close to these metrics but since its subject is closer to spending, it is seen as a duplicate from "Spending reflects business strategy". The metric "The project entity yields an optimal return" seems very similar to "Portfolio contains high value products"

from Cooper et al. [1999] and it is treated as a duplicate. The paper also contains some metrics that are very similar to metrics that have been reported as benefits on the team and product level: "Priorities across projects are known" and "Portfolio management is efficient". The former concerns itself with transparency, which is a metric that can also be found in Putta et al. [2018], Laanti et al. [2011], Stettina et al. [2021] and others. The latter concerns itself with efficiency like Olszewska et al. [2016], Gustavsson and Bergkvist [2019] also do.

Thornley [2012] contains four metrics that coincide with Cooper et al. [1999], but are described in such a way that they can easily be made quantifiable in a business context: "Alignment of programs to business-unit strategic goals", "Alignment between spending and portfolio priority", "percentage of the program milestones accomplished" and "program portfolio distribution". Furthermore the metric "Projected future income from program road map" which talks about the net present value of the road-map, is novel. Finally it contains a metric on customer satisfaction, which coincides with the team and product levels metric, by asking customers about their opinion on a scale from one to five.

As discussed in section 3.1.2, the approach of Cooper and Edgett [2003] is different from the other papers. Instead of looking at benefits it looks at ailments of poor portfolio management. For the purposes of this thesis these ailments have been inverted and are seen as benefits of proper portfolio management. The paper contains four metrics that coincide with metrics on the team and product levels. Unlike other papers about the portfolio level, the description of these metrics suggest that they pertain also on the team and product level. For example the morale metric specifically states "The project team's morale suffers". The same is true for the "Quality of execution suffers", "Time-to-market lengthens" and the "active projects are dumbed down" down metrics. The "Vital activities don't get done" metric is explained to lead to a lower success rate of projects, therefore it is considered a duplicate of the "Success/fail/kill rate of projects" from Cooper et al. [2004a]. Finally the "Game-changers are missed" metric is stated to lead to a lost opportunity cost. It is seen as a duplicate to "Entered new product categories", but one could also argue that it belongs under the "Opened up new markets" metric. Both of these metrics are from Cooper et al. [2004a].

Meskendahl [2010] has three metrics that are other instances of metrics from Cooper et al. [1999] or Cooper et al. [2004a]. The metrics "Average single project success", "strategic fit" and "Portfolio balance" are all mentioned in these papers under similar names. The fourth metric "Use of Synergies" is explained to pertain to "Technical and market synergies between projects within the portfolio". This suggests that it is another instance of the "Our new product program develops our existing technologies and technological competencies" form Killen et al. [2008], which also measures the mechanics between value-streams that allow technologies between value-streams to further develop.

Compiling all metrics into one list and removing any duplicates results in 52 unique metrics. Of these the majority belongs to the team and product level, a total of 34 (table 3). The portfolio level only has 19 metrics (table 4). This large range between the amount of collected metrics can be explained by looking at the disparity in literature when it comes to agile methodologies on the team/product level and on the portfolio level, as discussed in section 2.2.1. There are 6 metrics that are mentioned in both a team and product context and a portfolio context.

Table 3: The 34 unique team and portfolio metrics mentioned in literature, categorized in six dimensions. The leftmost column contains the metrics, each dimension is in boldface. An "X" in a cell means that the metric in that row was mentioned in the paper of that column. The final column ("#") shows how often a metric was mentioned in total.

[a] Indicates a metric was omitted for being too similar to another metric

[b] Indicates a metric was omitted for only being mentioned by one source and not having a detailed explanation

[c] Indicates a metric was omitted for not fitting an agile context

	Lee and Xia [2010]	Petersen and Wohlin [2010]	Laanti et al. [2011]	Olszewska et al. [2016]	Recker et al. [2017]	Putta et al. [2018]	Gustavsson and Bergkvist [2019]	Laanti and Kettunen [2019]	Business Agility Institute [2020]	Putta et al. [2021]	Stettina et al. [2021]	#
Productivity												
Increases effectiveness of development	X		X								X	3
Features per money spent	X			X	X		X					4
Responsiveness												
Improves time-to-market[a]								X	X	X	X	4
Customer service request turnaround time				X								1
Lead time per feature				X								1
Enable faster feedback										X		1
Decreases lead time for testing[a]		X										1
Quality												
Increases quality of product			X							X	X	3
Enables earlier detection of defects		X	X								X	3
Average days open external trouble reports				X								1
Number of external trouble reports				X								1
Software Functionality	X				X							2
Workflow Health												
Makes work more organized											X	1
Makes work more planned	X	X	X		X		X				X	6
Number of days between commits				X								1
Number of releases per time unit				X						X		2
Focus on continuous Improvement						X		X				2
Improved dependency management						X	X			X		3
Increases delivery predictability										X		1
Cadence and rhythm[b]								X				1
Better ways of working[b]									X			1
Process performance[c]					X							1
Employee Satisfaction & Engagement												
Makes work more fun			X								X	2
Makes work less hectic			X								X	2
Increases autonomy development teams			X							X	X	3
Increases collaboration			X			X	X	X	X	X	X	7
Increases transparency of development			X			X	X	X		X	X	6
Moving people together reduced documentation[a]		X										1
Better team alignment[b]						X						1
Improved morale[a]						X						1
Improved visibility[a]						X						1
Improved employee satisfaction[b]						X						1
More self organizing teams[a]						X						1
Customer Satisfaction												
Overall satisfaction					X				X	X		3

Table 4: The 19 unique portfolio metrics mentioned in literature, categorized in six dimensions. The leftmost column contains the metrics, each dimension is in boldface. An "X" in a cell means that the metric in that row was mentioned in the paper of that column. The final column ("#") shows how often a metric was mentioned in total. For the sake of space many metrics are presented in a shorter form.

[a] Indicates a metric was omitted for being too similar to another metric

[b] Indicates a metric was omitted for only being mentioned by one source and not having a detailed explanation

[c] Indicates a metric was omitted for not fitting an agile context

	Cooper et al. [1999]	Cooper and Edgett [2003]	Cooper et al. [2004a]	Martinsuo and Lehtonen [2007]	Killen et al. [2008]	Meskendahl [2010]	Thornley [2012]	#
Productivity								
Features per money spent			X	X				2
Responsiveness								
Improves time-to-market[a]		X	X					2
Quality								
Increases quality of product		X						1
Workflow Health								
Employee Satisfaction & Engagement								
Increases transparency of development				X				1
Improved morale[a]		X						1
Customer Satisfaction								
Overall satisfaction							X	1
Business Alignment								
Products align with business objectives	X			X	X	X	X	5
Spending reflects business strategy	X			X	X		X	4
Portfolio management focuses on the right issues[b]				X				1
Financial Performance								
Portfolio contains high value products	X			X	X			3
Success/fail/kill rate of products		X	X			X		3
Percentage of revenue and profits coming from new products[a]			X					1
Percentage of projects that are on time and on budget[c]	X		X		X		X	4
Percentage of projects meeting objectives			X					1
Profitability versus competitors[b]			X					1
Projected future income from portfolio roadmap							X	1
Opportunity								
Integrated new scientific knowledge[a]			X					1
Entered new technologies[a]			X		X			2
Entered new product categories[a]		X	X		X			3
Opened up new markets[a]			X		X			2
Portfolio Balance								
Portfolio has good balance of products	X				X	X	X	4
Portfolio reflects resources available	X				X			2
Portfolio develops our technologies and competencies					X	X		2
Other								
Technical success rating[b]			X					1

3.1.4 Initial Refinement

This set of 52 metrics covers many aspects of organizational performance but it is still problematic in other ways. Duplicates have been grouped together but still some metrics are very similar. Other metrics are mentioned only once and without explanation as to what they are measuring. Therefore an initial refinement is done on this set of metrics. With the goal of getting a more concise framework, which can then be used presented to practitioners for further feedback. In both table 3 and in table 4 the metrics that were omitted in this process are marked, alongside with the reason for their omission. In short, there are three reasons for a metric to be eliminated in this refinement step: it is too similar to another metric, it is only mentioned by one source and without explanation or it does not fit an agile context. In this section the reasoning behind all omitted metrics will be explained in further detail. This will be done per performance dimension, if a performance dimension includes such a metric.

The responsiveness dimension includes one metric that will not be included for feedback. The "improves time-to-market", "lead time per feature" and the "decreases lead time for testing" metrics are all similar. The "decreases lead time for testing metric" is considered a subset of the other two, as lead time for testing is one of several elements that makes up the time-to-market metric [Stettina et al., 2021] and the lead time per feature metric [Olszewska et al., 2016]. This leaves the time-to-market and the lead time per feature metrics who differ foremost in how they are used in their original papers. The time-to-market metric can be found in various papers as can be seen in both table 3 and 4 but its description is always much less detailed than the one source of lead time per feature. Due to this similarity these metrics will both be included in feedback rounds with the intention to omit the less recognizable metric.

The workflow health dimension contains three metrics that were omitted for feedback and further use. In their open-ended questionnaire Laanti and Kettunen [2019] listed "cadence, rhythm" as their third most reported SAFe benefit. This is not further elaborated upon, as the authors choose to focus more on the state of respondents' transformation, leaving the exact meaning unclear.

The "process performance" metric is used in a questionnaire by Recker et al. [2017]. They have taken this metric from Wallace et al. [2004] where they used it to assess the risks of software projects. According to the original source, process performance refers to the success of development in terms of the extent in which it was delivered on schedule and within budget. Such a metric contradicts the principles of agile, the second principle of the agile manifesto reads: "Welcome changing requirements, even late in development. Agile processes harness change for the customer's competitive advantage." [Fowler et al., 2001]. Due to this contradiction of the metric with agile contexts, it seems irrelevant for the purposes of this thesis.

The "better ways of working" is one of the four metrics that are discussed in the survey by Business Agility Institute [2020]. This metric seems very broad and the description further confirms this: transparency, engagement, collaboration, value stream focus, a reduction in hand-offs and a decreasing delivery time and cost are all mentioned as being part of a "better way of working". This ambiguity makes it less useful as a research metric, as a positive effect on this metric can mean many things.

Employee satisfaction & engagement contains six metrics that were not included for feedback. The "less required documentation" or as it is presented in the original paper "Moving people together reduced the amount of documentation that was not reused due to direct communication", was the last commonly perceived improvement reported by Petersen and Wohlin [2010]. Commonly perceived in this context means that at least 10% of the interviewees mentioned this as a benefit in some way, as stated by the authors. The provided examples of responses that were coded with this benefit, give a clear understanding of what this metric means. On page 29 of the paper two statements from different interviewees are given, in these statements the following is mentioned: "it is easier to communicate", "knowing what others are doing for that day as a result of stand-ups" and "walls being broken between testing and design". Given the context of metrics used in other sources, such as "increases collaboration" and "increases transparency of development" [Stettina et al., 2021], it seems that less required documentation is not the resulting benefit but rather a collection of other benefits. Like those mentioned by the interviewees. Due to this similarity between other metrics "less required documentation" in itself will not be included further.

The metrics: "increased team alignment", "improved visibility", "improved employee satisfaction" and "more self organizing teams" are mentioned only in Putta et al. [2018] which in turn obtained these from other academic sources. Unfortunately the "increased team alignment" and the "improved employee satisfaction" are both from sources that are not available to the author of this thesis at the time of writing. The original authors were approached for a copy of their papers but no response was ever received. Therefore the only available information about these metrics is their names as presented in Putta et al. [2018], making it impossible to know what these metrics are measuring and as a result they are unfit for further use. The source of the metric regarding self organizing teams is Razzak et al. [2017], which was not used for in depth analysis in this thesis due to its small sample size. In its original paper the metric is measured by five survey questions that are categorized under the "Team health" metric. The questions ask about

collaboration, planning events, cyclical development and communication. Compared to other metrics that were found this seems to be a very broadly defined metric. For example one of the five questions reads: "Team members are self-organized, respect each other, help each other complete sprint goals, manage inter-dependencies and stay in-sync with each other". This question in itself contains various aspects covered by multiple metrics, such as collaboration and autonomy of development teams [Stettina et al., 2021]. This is also the case for other questions which is why the less ambiguous individual metrics will be used in favour of this metric. The metric about visibility is taken from Pries-Heje and Krohn [2017] which is a case study using interviews to look at the lessons learned and challenges faced in a SAFe transformation. The paper is very short and does not talk about what visibility means in this context resulting in an unclear measurement. As there already is a transparency metric which is explained in further detail [Stettina et al., 2021], the visibility metric will be superseded by the transparency metric in further use. Pries-Heje and Krohn [2017] also contains "improved morale" as a benefit which can be found in one other paper besides Putta et al. [2018]: Cooper and Edgett [2003] where "the project team's morale suffers" is the last mentioned ailment of poor portfolio management. Much like the visibility metric Pries-Heje and Krohn [2017] provide no further clarification however Cooper and Edgett [2003] do provide further context. It states that nobody wants to be in teams anymore and see more work as a punishment, which makes it seem very similar to the fun metric found in both Laanti et al. [2011] and Stettina et al. [2021]. Since these two papers directly talk about the impact of LSA transformations, the fun metric will be used over the morale metric.

The business alignment dimension contains only one metric that is omitted: the "Portfolio management focuses on the right issues" metric from Martinsuo and Lehtonen [2007]. The authors never expand on what is meant by "the right issues", i.e. what are the criteria for an issue being right. For the section of their research pertaining to portfolio management they cite various papers by Cooper et al. as their source, such as Cooper et al. [1997] and Cooper et al. [2000]. However, this particular use of "right issues" can not be found in these papers. Due to the ambiguous nature of this term the metric will not be used any further.

The financial performance dimension has several omitted metrics which were omitted for various reasons. The "overall profitability versus competitors" is one of many metrics in Cooper et al. [2004a] that is only discussed briefly in the original paper and which cannot be found in other academic sources. What overall profitability means in this context is not further explained, allowing multiple interpretations. The comparison against competitors is also not elaborated upon, which is troublesome due to how different companies use different portfolio management tools and measurements as mentioned in the same paper. Because of these considerations this metric will not be used further.

The "percentage of revenue and profits from new products" metric covers a very direct aspect of financial performance: revenue and profits. However, this particular metric only focuses on the revenue and profits of new products instead of the entire portfolio. This ignores the effect that new products might have on already existing products, these might compete for the same resources which can hurt already existing products [Cooper et al., 1999]. To get a more complete measurement of portfolio revenue and profit, a metric is used that does not only look at new products but the entire portfolio road-map. That being the "Projected future income from portfolio road-map" metric form Thornley [2012].

The "projects that are on time and on budget" metric can be found in various forms in many papers [Cooper et al., 1999, 2004a, Killen et al., 2008, Thornley, 2012]. Despite its frequent appearance, the metric does not seem to fit this thesis due to its clashing nature with an agile context. Much like the "process performance" metric from Recker et al. [2017] this metric contradicts the second principle of the agile manifesto which is about welcoming change. In an agile context the budget and deadline of a product should be subject to constant change, and not be used as measurements of success. Which is why this metric will not be further used in this thesis.

The entire opportunity dimension is interesting in the sense that all its metrics are very similar. These metrics originate from Cooper et al. [2004a] and Killen et al. [2008] and in both sources they are not discussed as thoroughly as other subjects, leaving a lot open for interpretation. In Killen et al. [2008] these metrics are called portfolio opportunity measures, presumably due to them all having something to do with entering a new market or integrating a new technology. The difficulty within this dimension is not necessarily a single metric, but rather the difficulty in making a distinction between all of its metrics. As an example take the metrics "Integrated new scientific knowledge" and "Entered new technologies", what does it mean for a new technology to be integrated instead of entered. Furthermore, no sources could be found that suggest that this particular breakdown of the ability to seize opportunities, is used by practitioners. This ability can be beneficial for an organization and it is also mentioned in Cooper et al. [2004a], but as a single ailment of poor portfolio management rather than this breakdown into four separate metrics. This lack of distinction between the four metrics, while at the same time having an apparent benefit of some form of opportunity metric, is the reason the four separate metrics well be replaced by one general opportunity metric. Since this metric deals with bringing new opportunities to the organization, it is a nice counterpart to the "portfolio develops our technologies and competencies" metric from Killen et al. [2008]. This metric deals with existing elements within the organization. As such this newly merged opportunity metric will be put under the portfolio balance dimension.

The "technical success rating" was not categorized under a dimension as it did not seem to fit any. This is another relatively unexplained metric from Cooper et al. [2004a] where nothing but the survey result is mentioned. No other source seems to use this metric, or a similar metric either. This further limits the available information. For these reasons this metric is omitted, as well as the entire "other" dimension as it only contained this single metric and had no further basis in literature.

3.1.5 Academic and Practitioner Feedback

Now there exists a more concise framework of metrics obtained from academic literature. To broaden the range of inputs for the framework, several feedback sessions are held with experts. In total there are eight separate feedback sessions with eight different experts. Six of these experts are practitioners (consultants) with experience in the area of large-scale agile transformations, one expert is an academic and one expert is active both as an academic as well as a practitioner. Each session is roughly thirty minutes, in which participants are presented the framework of metrics and asked if they would change anything. Be it adding a metric, deleting a metric, moving a metric to another dimension or changing up the dimensions all together. Participants are presented the framework ahead of time but at the beginning of the sessions a short amount of time is spent going over the framework as well. Notes are kept on anything a participant might have said. After all the sessions are done a compiled list of all notes is created to get an overview of all the collected feedback. These sessions are not conducted, and therefore should not be regarded as, scientific interviews but can still provide a useful way to incorporate practitioner and academic views outside of literature.

It is common for an individual metric to be deemed ambiguous in one or more feedback sessions, however no metric seems to stand out in this regard. The "customer service request turnaround time" metric from Olszewska et al. [2016] does receive the same critique from three participants. That critique being that a decreased customer service request turnaround time does not necessarily lead to an increase in responsiveness within development teams. The argument being that the responsibility of product development and dealing with customer service requests might be split between separate teams. This might lead to a more responsive experience on the customer's end, but it could be misleading to say that a team has become more responsive. The metric is left in the survey despite these critiques because results from the survey might confirm these suspicions. The survey aims to measure the benefits of large scale agile transformations as perceived by employees within that transformation. If the customer service request turnaround time metric does not experience an impact it could be that this is indeed a bad metric to measure responsiveness with.

On the issue of making a decision between the "improves time-to-market" metric and the "lead time per feature" metric, two experts remark that the term "time-to-market" is somewhat ambiguous. Of these two only one explicitly states their preference for the "lead time per feature" metric. A third expert remarks that the "improves time-to-market" metric implies that there is an end-product that can be shipped, which is not always the case. Because of this slight preference amongst experts, the decision is made to continue with the "lead time per feature" metric over the "improves time-to-market" metric.

Various suggestions for additional metrics are also made. For one participant it is unclear what the "makes work more planned" and "makes work more organized" metrics were measuring exactly. After some discourse they suggest adding an "amount of unexpected work has decreased" metric to see if more people recognize this metric and how it compares to the other two. Their underlying thought process being that wording it in this way could cover similar aspects of workflow health, while at the same time being more direct.

Two participants suggest both a "decrease in employee turnover" metric and an "allows to attract more employees" metric as part of the employee satisfaction & engagement dimension. They argue that these metrics were often already measured and could be a result of a more fun and autonomous working environment. Therefore they can provide a less ambiguous metric in this particular dimension.

For portfolio level metrics, one expert mentioned that the lack of a transparency metric is missing. Although transparency is mentioned in Martinsuo and Lehtonen [2007] as an element of portfolio management efficiency, it was unclear what this meant on portfolio level therefore it is not included. One participant elaborates that a portfolio backlog and portfolio dependencies should be transparent within the organization. This can give an organization wide understanding of the decision making process and allow teams to align their vision with that of the company. Given this explanation the metric is included for the survey.

Taking into account all of these points gathered from the feedback sessions results in a final comprehensive framework that can be used to measure performance benefits (figure 2). This final framework consists of nine dimensions, six for the team and product level and three for the portfolio level. In turn these dimensions contain 26 metrics for the team and product level and 11 for the portfolio level, for a total of 37 metrics.

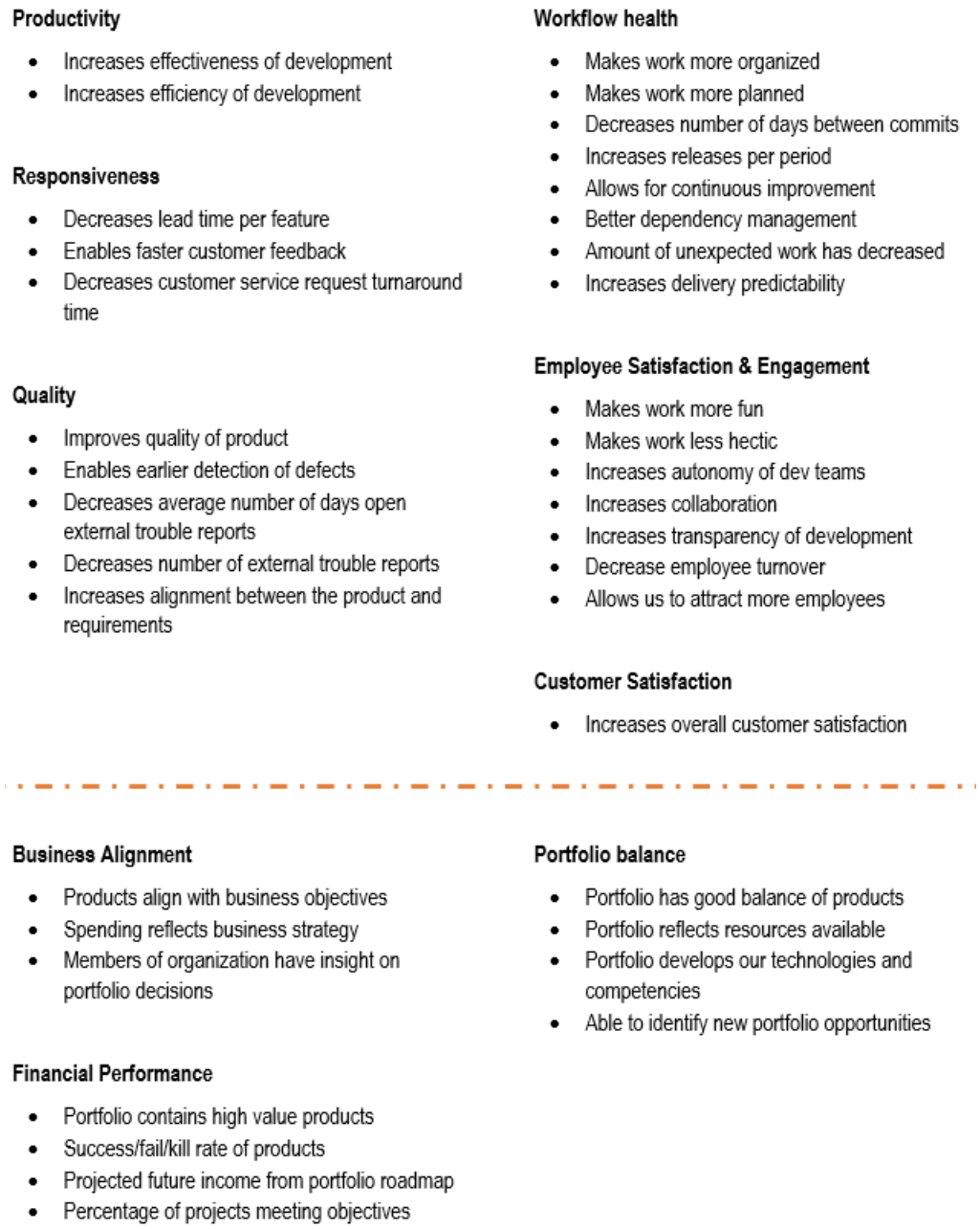


Figure 2: Final framework that will be used to measure performance benefits. The orange dotted line separates the team and product benefits (above) and portfolio benefits (below).

3.2 Survey Design

This survey was designed from the top down, from the research question to internal questions to eventual survey questions as described by Linåker et al. [2015]. The aim of this survey is to measure the relation between LSA transformation maturity and organizational performance. Transformation maturity will be measured with the agile practices mentioned in Laanti [2017], which is discussed in section 2.1.3. Organizational performance will be measured through a framework consisting of metrics from literature, as discussed in section 3.1.

Data needs to be collected to be able to answer the research question. In order to make sure data is collected from the right individuals, units of observation and the unit of analysis will be defined [Linåker et al., 2015]. The unit of analysis can be obtained from the research question. It concerns organizations undergoing an LSA transformation that need

to be measured, making it the unit of analysis. The individuals from which data-points will be obtained need to be a part of an organization undergoing such a transformation. The used maturity model puts further limitations on the units of observation due to the practices it uses, which are picked specifically for an IT context. This makes the units of observation employees who work within an IT context in an organization that is undergoing an LSA transformation.

For answering the research question data is required on both the transformation maturity and on the perceived performance benefits, for both of which instruments have been chosen. These two instruments will get their own section in the survey. Besides these two sections two more sections will also be included. The first of which can be categorized as demographic questions [Linåker et al., 2015] which in this case serves two purposes: it will provide insight on who completed the survey which helps with exposing biases in the data and it will also provide a comfortable starting point to the survey as to encourage respondents to go on. The second section that will be added is not related to the research question directly but rather contains a collection of exploratory questions that might provide additional insights. These questions are optional (as opposed to all other questions in the survey) and put at the very end of the survey. These questions will be further referred to as cluster questions, as they might expose interesting clusters of respondents.

In this section of the thesis each of these individual sections will be discussed in further detail. This will be done in the same order that they are shown to respondents in the survey. Finally the distribution and sampling strategies for the survey will also be discussed. For the whole survey see appendix A.

3.2.1 Demographic Questions

The first section of the survey after the introduction, will consist of demographic questions. These are simple questions that ask about the respondents position within the organization, the organization itself and about the basics of the transformation such as the used framework. These results of this section provide insight on the demographics of the respondents, which can help explain certain patterns. For instance if the relative number of developer respondents largely differs from those in Stettina et al. [2021], the end results might also differ. The simplicity of these questions also provide an easy start to the survey, encouraging participants to continue with the survey [Linåker et al., 2015].

At the end of this section respondents are presented with three questions that ask about their familiarity with their organization's transformation. Each question asks about their familiarity on each of the three organizational levels: team, product and portfolio. Since a single unit of observation can be anything from a junior developer to an executive it is possible that their understanding of the transformation on one level might differ from their understanding at another level. Which is why for each of these levels the respondent is asked about their familiarity. The answers to these questions have no impact on what parts of the survey a respondent is shown, each respondent gets shown the entire survey, however including these questions does provide the possibility to filter based on familiarity.

3.2.2 Maturity Model

To measure the maturity of a respondent the maturity model by Laanti [2017] will be used, which can be seen in full in figure 1. This model measures three levels of maturity: portfolio level maturity, program level maturity and team level maturity. On each of these levels there are five stages of maturity, going from the lowest stage of maturity to the highest these are: "beginner", "novice", "fluent", "advanced" and "world-class". The way the model measures maturity is by asking the user of the model to rate themselves for each level. In each cell of the maturity model there is a set of practices and/or milestones associated with LSA transformations. In order to be at a certain stage of maturity on an organizational level, you need to have implemented and/or achieved all of the practices and milestones that are in the stages before that particular stage. For example, if an organization is at the fluent stage on the team level, they must have implemented and achieved all of the practices and milestones in both the beginner and novice stage of the team level. The resulting measurement will be a maturity rating in the form of one of the stages, for each of the three organizational levels.

First respondents will be presented the individual practices and milestones that constitute the three levels of the model. Each level will have its own section in the survey. Within these three sections the practices and milestones will be presented in a random order to prevent any sort of bias. The organizational level is explicitly stated at the beginning of each section and the scale is repeated every five questions, as to have the options always be on screen. For each practice or milestone respondents will be asked how often they do that practice or achieve that milestone on a five point Likert scale: "never", "seldomly", "sometimes", "frequently" and "always". This data can then be used to do a test of the validity of the maturity model as it is presented by Laanti [2017], see section 3.3.1 for the details of how this will be done.

After the final section of individual practices respondents will also be shown the entire model and asked to rate themselves according to the instructions of Laanti [2017]. Although there is value in testing the validity of the model, getting data directly from the model makes it possible to test the correlation between maturity and the performance

metrics in a similar way to Stettina et al. [2021]. Therefore, for the sake of comparability, this measurement is also taken. The model is purposefully presented after showing the respondents the individual practices and milestones to prevent any bias that might result from having previous knowledge on the model.

3.2.3 Framework Metrics

After the maturity section respondents will be asked about which performance benefits they perceived. This will be done on the basis of the performance benefits framework as it is portrayed in figure 2, and whose development is described in section 3.1. To use this framework the individual metrics will have to be translated into survey questions, however in this process various considerations can be made.

In literature it is shown that a Likert scale can be used to measure organizational performance in the context of LSA transformations [Laanti et al., 2011, Putta et al., 2021], but a slider is also an option [Stettina et al., 2021]. An argument that has been made for sliders is that they might be more engaging for respondents and that they might lead to superior data, but researchers were unable to find convincing evidence for either argument [Roster et al., 2015]. Furthermore, the average percentage of a slide type scale can also be compared to the mean of a Likert type scale[Stettina et al., 2021]. As this research is based on the framework used in Stettina et al. [2021] and because no reason to choose a Likert scale over a slider scale can be found, a slider scale will be used in this survey to benefit the comparability with Stettina et al. [2021]. One change to this scale will be made. In the original paper the scale was an improvement scale, starting at a 0% improvement (i.e. no improvement) and going up to a 100% improvement. In this survey the scale will range from -100% to 100%, where any negative answer will mean a decrease in a metric. This gives the respondents the ability to indicate a perceived decrease in a metric. Respondents are notified of this through the means of an example.

The extension of the range into the negative in combination with the metrics as they are presented in the framework, can lead to double negatives. For example, the metric addressing the lead time per feature is shown as "Decreases lead time per feature". An answer within the negative range would mean a double negative which can be confusing for a respondent [Roster et al., 2015]. To avoid such confusion any use of words implying improvement in performance or decrease in performance will be omitted. Using the previous example of lead time per feature, "Decreases lead time per feature" will become "Lead time per feature". A negative answer in this case will mean a decrease in lead time per feature, which in turn is a positive impact on the performance. The same will be done for any instances of words like "increases" or "improves". For cases where a respondent might not be familiar with a metric at all, each metric will also have a "Don't know" option. The amount of "don't knows" in itself can also provide an insight on which metrics are recognized by practitioners.

The metrics will be divided into two blocks, one for team and product metrics and one for portfolio metrics. Unlike with the maturity section of the survey, respondents will not be explicitly told what level of the organization the metrics pertain to. Within these two blocks metrics are further divided into smaller sections. This makes sure that there are not too many metrics on the screen at any time, therefore ensuring that the scale is visible as often as possible. Within these individual sections the metrics are presented in a random order to a respondent.

3.2.4 Cluster Questions

The final block consists of a collection of questions that do not directly relate to the research question. Rather these questions ask about how the respondent's organization is run. These questions ask about how often the impact of the LSA transformation is measured, if a line structure is present, if Scrum Master is a permanent role etc. (for a full list see appendix A for the entire survey). These questions have no basis in literature but are exploratory questions to see if any of these factors might also influence the impact of an LSA transformation. If for any of the questions this is the case, this could then be used in possible future research.

3.2.5 Survey Distribution

To distribute the survey and find respondents, a sampling strategy was formed. As discussed in section 3.2, the units of observations are employees that work within an IT context, in organizations that are undergoing an LSA transformation. To reach these units of observation, a snowball sampling strategy [Linåker et al., 2015] will be used. Employees with manager-like positions will be approached such as: Scrum Masters, Release Train Engineers (RTE), Product owners etc. These managers will be informed about the survey and the research and asked to participate, as well as to spread the survey within their team or teams to reach developers and similar positions. By requesting managers to further spread the survey, potential respondents can be reached while still maintaining a concise target demographic for promotional material. This is also done in part due to the fact that managers are more easily reachable for given the position of the author, and therefore seems like the more effective strategy. This gives the sampling strategy some elements of accidental sampling [Linåker et al., 2015] as well.

Various channels have been identified to reach these managers. As this thesis is written in combination with an internship at KPMG The Netherlands, the network of employees of KPMG The Netherlands can be used to reach out to people that fit the demographic. Another channel is an RTE Summit, where various RTEs will gather to discuss their experiences with SAFe. The author of this thesis is asked to co-host a workshop at this summit. At the end of this workshop participants will be given a chance to leave their contact information such that they can be contacted later about participating in the survey. Other people at the summit can also be approached. Finally, there exist various LinkedIn groups focused on (large scale) agile and related topics. These groups will also be joined to spread promotion material.

3.3 Analysis

To answer the main research question of this thesis, two components will be used. One to measure LSA transformation maturity and one to measure organizational performance. These are the maturity model by Laanti [2017] and the performance framework from section 3.1 respectively. Since the correlation between transformation maturity and organizational performance depends on these two measuring instruments, the individual instruments will also undergo some analysis. In this section the methods for performing these tests are discussed, as well as the underlying reasons for picking these tests. This will be done in section 3.3.1 for the maturity model, and in section 3.3.2 for the performance framework. In section 3.3.3 the method with which the relation between LSA transformation maturity and organizational performance will be discussed.

3.3.1 Maturity Model

The model by Laanti [2017] was constructed through the collaboration of practitioners and academics, where colleagues of the author were interviewed about their own LSA experiences. Based on this information they developed their categorization. The author then validated the model in a 12,000 FTE banking and insuring organization. Within this organization 42 interviews were conducted with 117 people and the reactions were generally positive. In their paper the author goes on to say that the model has been taken into use by several other organizations [Laanti, 2017] but no further validation of the model could be found in academic literature.

Because the model was never validated outside of its original context one could question its external validity. Furthermore, the authors of this thesis have used the model with LSA practitioners from various organizations and countries to let them reflect on their own transformation. This was done twice in a workshop setting, both times the model sparked some discussion. These workshops are by no means evidence that put the model in doubt, but it did bring further attention to the issue of external validity. Therefore with this survey the validity of the model will be further tested through the graded response model (GRM) [Samejima, 2011]. GRM is a family of statistical models that allow somebody to measure some latent trait based on a collection of responses, if the response format is supported. These models allow for two analyses: assessing a latent trait of an individual respondent and assessing the difficulty in attaining a certain score for each individual question. Using such a model provides the ability to both measure the transformation maturity of a respondent (the respondent's latent trait), as well as further test the model by looking at individual practices and milestones and seeing how strong the latent trait needs to be to implement or achieve it.

A response format that can be used as input by a GRM model is a Likert scale. The example used by Samejima [2011] is a scale ranging from one through four: strongly disagree, disagree, agree and strongly agree, but any Likert scale can be used. By taking the individual practices used in Laanti [2017]'s maturity model and allowing respondents to answer how often they do a practice or accomplish a milestone using a Likert scale, data can be collected. This data can then be used to calculate the required maturity level to implement a practice or achieve a milestone, allowing for a comparison between these elements. For the range of the Likert scale, 5 different stages will be used indicating how often a practice is done or how often a milestone is accomplished: "never", "seldomly", "sometimes", "frequently" and "always". Using a GRM with this type of scale will result in an output of four coefficients (C_1, C_2, C_3, C_4), each one representing the required ability in the latent trait to have a 50% chance of answering that or a lower category. That is to say, if $C_3 = 1.23$ then an ability score of 1.23 is required to have a probability of 50% for a respondent to pick either the never, seldomly or sometimes option. Using this data the practices belonging to an organizational level can be compared between themselves.

The original model requires users to have fully implemented a practice or fully achieved a milestone to move to the next maturity level. The analysis of the GRM results will therefore focus on the C_4, which is the required maturity to have a 50% chance of giving an answer of "frequently" or lower. This same coefficient also means that it is the maturity required to fully implement that practice or achieve that milestone (the always option). Sorting the elements by this value results in a new ranking of the elements of the model, which can then be compared to the original model.

3.3.2 Performance Framework

Through the means of literature research, a large framework of performance metrics has been constructed. Although the majority of these metrics have been tested before in literature individually, no such tests exist for using them together in the context of an unsupervised survey. This can be a problem due to the ambiguous nature of some of the metrics, for example the term "effectiveness" can be interpreted in various ways. To see if respondents can distinguish between metrics properly, the Pearson correlation coefficients between metrics in the same dimension will be calculated. A high and significant coefficient could suggest that respondents have difficulty distinguishing between the two correlated metrics, suggesting that the combined use of the metrics is unfit for the context of an unsupervised survey. Respondents are also able to answer "Don't know" for each individual metric, as a result the amount of responses per question can also differ. Both these pieces of data can assist in making the framework more concise by exposing hard to distinguish metric pairs, as well as less recognized metrics.

3.3.3 Impact on Performance

To see which metrics are impacted the most by an LSA transformation, the mean response for each metric will be used. These will be presented alongside the standard deviation, minimum and maximum answers and each $25th$ percentile answer. This data can then be used for comparisons between other survey based literature like Laanti et al. [2011], Putta et al. [2021] and Stettina et al. [2021]. These papers share at least some of the metrics as the performance framework was created (in part) from these papers.

To test the relation between transformation maturity as defined by Laanti [2017] and organizational performance metrics, Spearman correlation coefficients will be used. This differs from Stettina et al. [2021] where Pearson coefficients are used. The decision to use Spearman coefficients rather than Pearson is due to the way that the maturity model measures transformation maturity. Since Pearson tests for linear relationships, it would work best if each step from maturity stage to the next one requires the same increase in maturity. However the structure of the maturity model does not suggest that this is the case, some steps between stages have much more requirements than others. Furthermore it might be the case that some requirements are harder to fulfill than others, something that will also be tested in this thesis. For these reasons it is assumed that it is more likely for the relation between maturity and performance metrics to be monotonic rather than linear, making Spearman coefficients the better fit.

4 Results

In total 139 responses to the survey were gathered over the course of 5 months. Of the total 139 there are 61 responses that completed the survey, resulting in a completion rate of 43.88%. From here on out the term "responses" will refer to the set of completed responses as opposed to the total amount of responses.

The responses were collected from twelve unique countries across five continents, although a majority of the responses came from The Netherlands (34). France had the second most respondents at five, followed by Germany (four), Belgium, The United Kingdom and The Czech Republic (all with three respondents) constituting the top five. Outside of Europe responses were collected from the US (two), Canada (one), China (one), Australia (one) and Israel (two).

In the rest of this section the results from the survey will be presented. Starting with descriptive statistics describing both respondents and the transformations they are experiencing in section 4.1. Next the results of the familiarity question for the team, product and portfolio level will be shown in section 4.2. The maturity as measured by the model from Laanti [2017] will follow in section 4.3. Afterwards the results of the impact metrics will be shown in section 4.4. Section 4.5 will be about the graded response model results about the individual practices in the model by Laanti [2017]. The final subsection, section 4.6, will show the results of the cluster questions in the survey.

4.1 Descriptive statistics

The survey contains two categories of descriptive questions, those asking about the respondent and their organization and those asking about the transformations. Questions about the respondent and organization were about the industry in which the organization of the respondent operates, the size of the organization and the role of the respondent within that organization (figure 3). Questions about the transformation were about the departments included in the organization, the total scope of the transformation as a percentage and frameworks used in the transformation (figure 4).

4.1.1 Respondents

Industry: The largest group of respondents works in the IT industry (32.79%), followed by financial services (26.23%) and various others (18.0%). All other industries only constitute 10% or less of the total respondents. This is similar to Stettina et al. [2021] where the number one and two represented industries were software and and financial services, the Business Agility Institute [2020] also has a similar top three.

Organization size: Roughly a quarter (26.23%) of respondents work in an organization of $5,001 - 20,000$ full time employees (FTEs), another quarter (22.95%) work in organizations with $20,001 - 50,000$ FTEs. Followed by organizations of $1,001 - 5,000$ FTEs (19.67%), $201 - 1,000$ FTEs (18.03%) and $> 50,000$ FTEs (13.11%). Compared to Stettina et al. [2021] this study has less respondents from organizations larger than $20,000$ FTEs (47.8% versus 36.2%) but a bigger amount of organizations larger than $5,000$ FTEs (47.8% versus 63.8%)

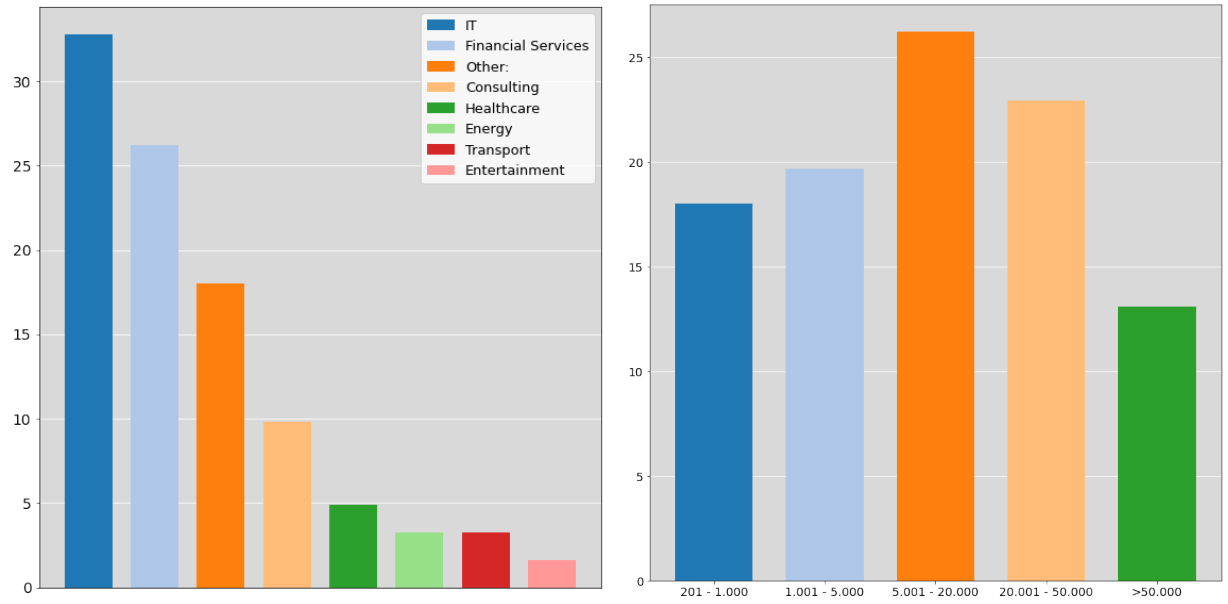
Roles: Agile coaches were the most represented roles (18.03%), followed by both Managers, Release Train Engineers and Product Manager/Owner (all at 11.48%). Other roles fill only 10% or less of the data-set. These sections of respondents are overall much higher than in Stettina et al. [2021], likely because the latter contained many options whereas the survey used in this study only contains eleven. In both studies the best represented roles are Agile Coaches.

4.1.2 Transformation

Transformed departments: 33.11% of respondents were part of a transformation that included the IT department, 12.84% included the production and finance departments, 12.16% included the R&D and marketing departments. Less than 10% included HR or other departments. This is similar to Stettina et al. [2021] where the most included department was also IT, with all other departments following in a range of $[9.4, 16.8]$. It should be noted that respondents were able to give multiple answers on this question.

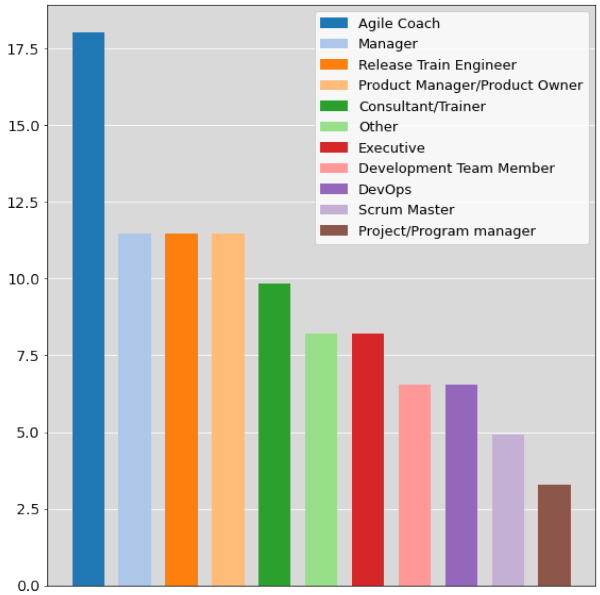
Transformation scope: The distribution of the scope of transformations ascends perfectly with the available options, 44.26% of respondents belong to organizations whose transformation scope is no more than 25%. Followed by a scope of $26 - 50\%$ with 21.31%, and even fewer with over 50% or 100%. This is in contrast with Stettina et al. [2021] where over 50% of respondents were part of a transformation that had a total scope of $26 - 75\%$ compared to only 37.9% in this study.

Used frameworks: As was the case in both Stettina et al. [2021] and in Digital.ai Software Inc. [2021] the most used framework by respondents is SAFe with 39.56%, the Spotify Model being the second most used at 16.48%. This is a much larger difference between the number one and two spots than in Stettina et al. [2021]. LeSS (12.09%), own



(a) Industry

(b) Organization size



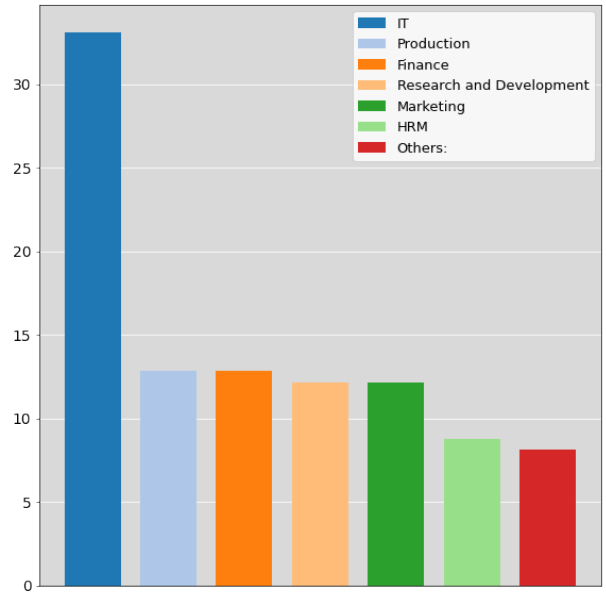
(c) Roles

Figure 3: Distribution of various characteristics of the respondents to the survey, shown as percentages

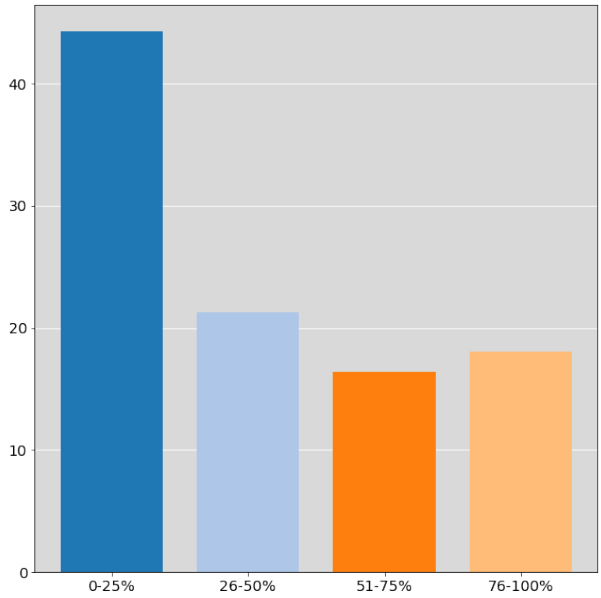
framework (9.89%) and Lean Management/Scrum of Scrums (both at 5.49%) constitute the rest of the top five. Please note that on this question respondents were able to give multiple answers.

4.2 Familiarity

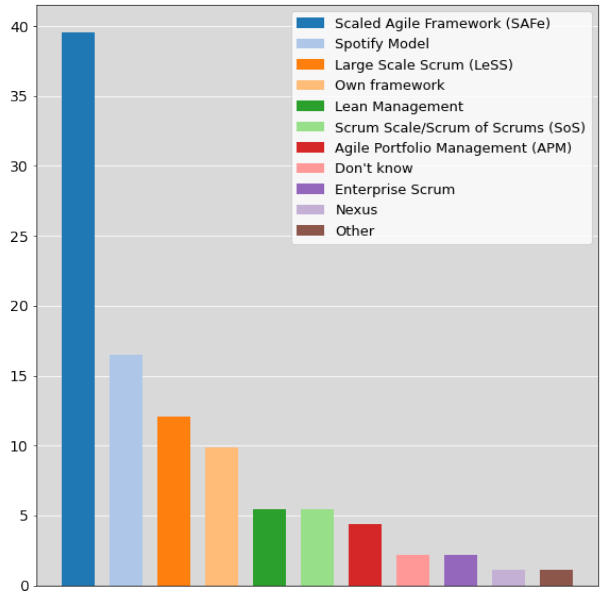
After the descriptive questions respondents were asked to rate their own familiarity for each of the three organizational levels specified in the model by Laanti [2017]. This was done with the following question: "How familiar are you with your organization's transformation on team/product/portfolio level?". The results of this question can be seen in figure 5. The numbers inside the bars indicate the total amount of answers for that specific option and level combination. The



(a) Transformed Departments



(b) Percentage of organization part of transformation



(c) Frameworks used in organization

Figure 4: Distribution of various characteristics of the transformations of respondents to the survey, shown as percentages. Note that for 4a and for 4c respondents could give multiple answers

means were calculated by mapping each answer to a numerical value where "Not familiar at all" is one and 'Completely familiar' is a five.

Respondents feel most familiar with the team level of the transformation of their organization with 88.52% stating they feel either fairly or completely familiar. This familiarity drops at the product level where 63.93% of respondents answer that they are either fairly or completely familiar. The familiarity at the portfolio level is even smaller with 54.10% giving an answer of fairly or completely familiar and only eleven respondents feeling completely familiar.

The amount of respondents that state they are not familiar at all with their organization transformation is relatively small for the team, product and portfolio level. The portfolio level has the most of these answers with three total, which is less then 5% of the total data-set.

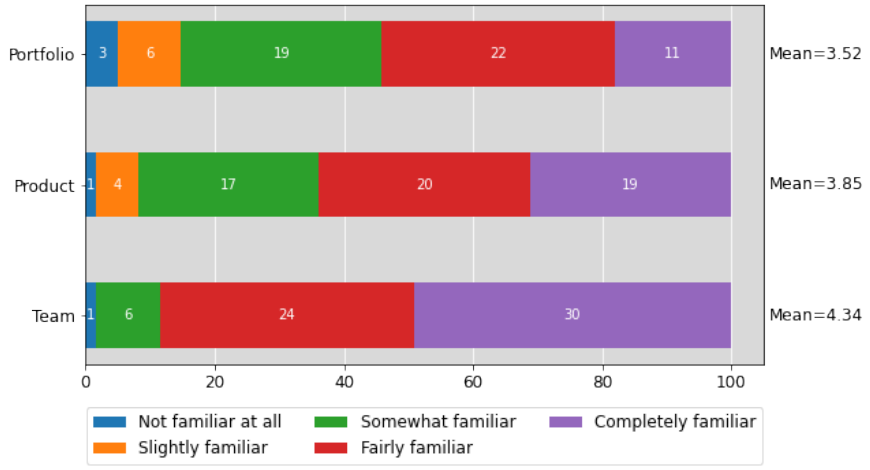


Figure 5: Distribution of confidence level of respondents per each level specified in Laanti [2017], shown as percentages.

4.3 Maturity Model Results

Figure 6 shows the distribution of answers for the questions which ask the respondents to rate their transformation maturity based on the model by Laanti [2017]. Numbers inside the bars show the total amount of responses for that option and at that level. Means were calculated by mapping each possible answer to a numerical value where "Beginner" is a one and "World-Class" is a five.

This figure shows that for all three levels most respondents indicate that they are at the "Novice" stage. Only one respondent indicated to be at the "World-class" stage for the portfolio level whereas five respondents did for the team and program level. When looking at the means of the three individual levels it becomes apparent that maturity decreases as you move from the team level (2.92) to the program level (2.52) and even further to the portfolio level (2.18). A similar decrease in maturity across these levels can be seen by looking at the medians, where over 50% of respondents answered that they were Fluent or higher at team level, this was only Novice at the program and portfolio level. When

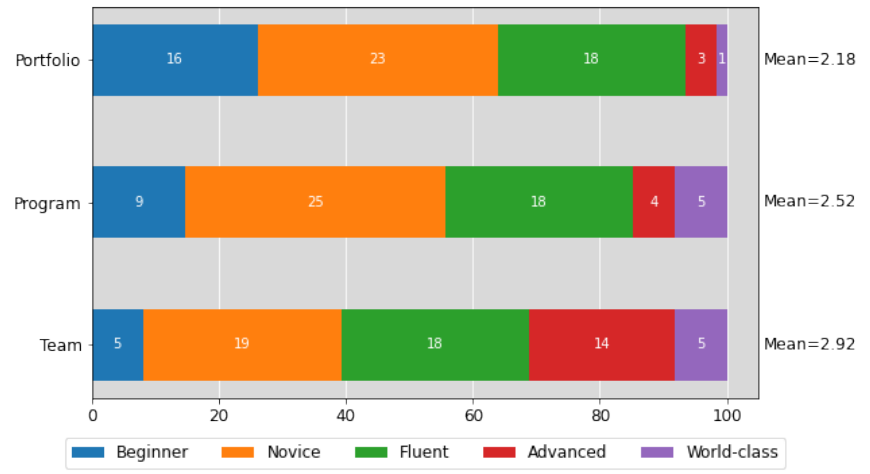


Figure 6: Distribution of maturity level of respondents per each level specified in Laanti [2017], shown as percentages.

comparing these means to Stettina et al. [2021] a slight increase can be seen on the team and program level and a

decrease on the portfolio level, as shown in table 5. It should be noted that although Stettina et al. [2021] was published in 2021 the data was collected in 2018.

Table 5: Comparison of the maturity means between the survey results and the study by Stettina et al. [2021]

Maturity Level	Survey data	Stettina et al. [2021]
Portfolio	2.18	2.28
Program	2.52	2.49
Team	2.92	2.87

4.4 Impact Metrics

In this section the results surrounding the impact metrics will be presented. Starting off with the correlation between metrics that make up the various dimensions in 4.4.1. As discussed in section 3.1 there is a distinction to be made between two categories of metrics: the first are the team and product level metrics which are supported by a body of academic literature on LSA contexts. The second category are the portfolio level metrics which are based on academic literature on more traditional ways of working, due to a lack of research in the portfolio level in LSA transformations. This same distinction will be maintained throughout this section. As a result the section 4.4.2 regarding the impact on team and product metrics will contain comparisons with other studies, as there is comparison material available. Section 4.4.3 regarding the portfolio metrics will only present the results obtained in this study.

4.4.1 Metric to Metric Correlation

For the sake of readability there are two figure groups that display the correlation between team and product metrics rather than one big group: figure 7 and figure 8. The "Customer Satisfaction" is missing from these figures because it only consists of one metric, therefore no pairs could be tested for correlation. Also due to readability the metrics in these figure are presented in a shorter manner compared to their presentation in the survey. Each dimension uses the same coloring and significance indicators. If a metric itself (rather than a coefficient) is marked with an "*" it means that if a respondent answered with a negative number, they indicate a positive impact. For example, if a respondent answered -25% for "Lead time per feature" this would mean that the lead time per feature decreased, which is a positive impact. As a result metrics with a "*" should have a negative correlation coefficient with metrics which do not have a "*". All obtained results show that whether or not respondents picked up on this intention is questionable, which is further discussed in section 5.1.2. Due to the "Don't Know" option which was present in the survey for each metric, the N-value of correlation pairs may differ.

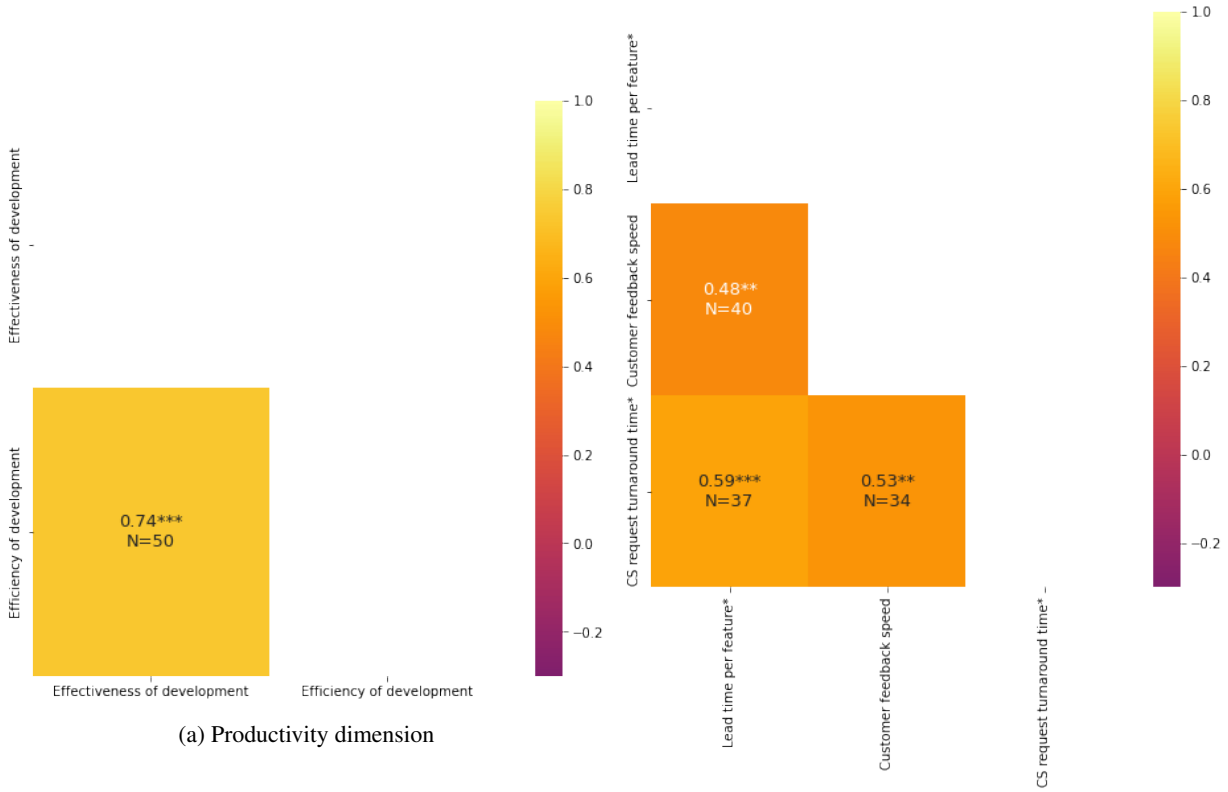
Of interest are any metric pairs that have a high Pearson correlation coefficient and whose correlation is also significant. Four such pairs stand out such, having coefficients of 0.7 or above and with significance at the 0.001 level: "efficiency of development & effectiveness of development", "work is planned & work is organized", "autonomy of dev teams & work is fun" and "transparency of development & collaboration".

These figures also show that between metrics in any particular dimension there is a lot of correlation. Ideally each metrics should cover a separate aspect of a performance dimension but this does not appear to be the case.

The metric pairs that consist of a metric with a "*" and a metric without a "*" are expected to have a negative correlation. This is not always the case, as with the "amount of trouble reports & quality of the product" pair. In general these pairs do seem to have coefficients closer to zero as well as having less significant correlations.

The metrics for the portfolio level can be seen in figure 9. Many of the things that apply for the team and product metrics also apply here. There is only one figure group but the metrics are still presented in shorter manners than in their presentation in the survey for readability's sake. Each dimension uses the same coloring and significance indicators. Due to the "Don't Know" option which was present in the survey for each metric, the N-value of correlation pairs differs. There are no metrics marked by a "*" because there are no such metrics for the portfolio level.

The portfolio metrics seem even more correlated than the team and product metrics. Unlike with the team and product metrics, there are no metric pairs which are expected to have a negative correlation, which is reflected in the results. There are also more portfolio metrics that have a coefficient above 0.7 and are significant at the 0.001 level than team and product metrics that meet the same requirement, despite the lower total amount of metrics. These pairings are: "insights on portfolio & spending reflects business strategy", "future income from road-map & high value product", "% of projects meeting objectives & high value products", "portfolio develops technologies & balance products in portfolio" and "identify portfolio opportunities & portfolio reflects resources available".



(a) Productivity dimension

(b) Responsiveness dimension

Figure 7: Pearson correlation coefficients for the productivity and responsiveness metrics. A single "*" means significance at the 0.05, a "**" means at the 0.01 and "***" means at the 0.001 level.

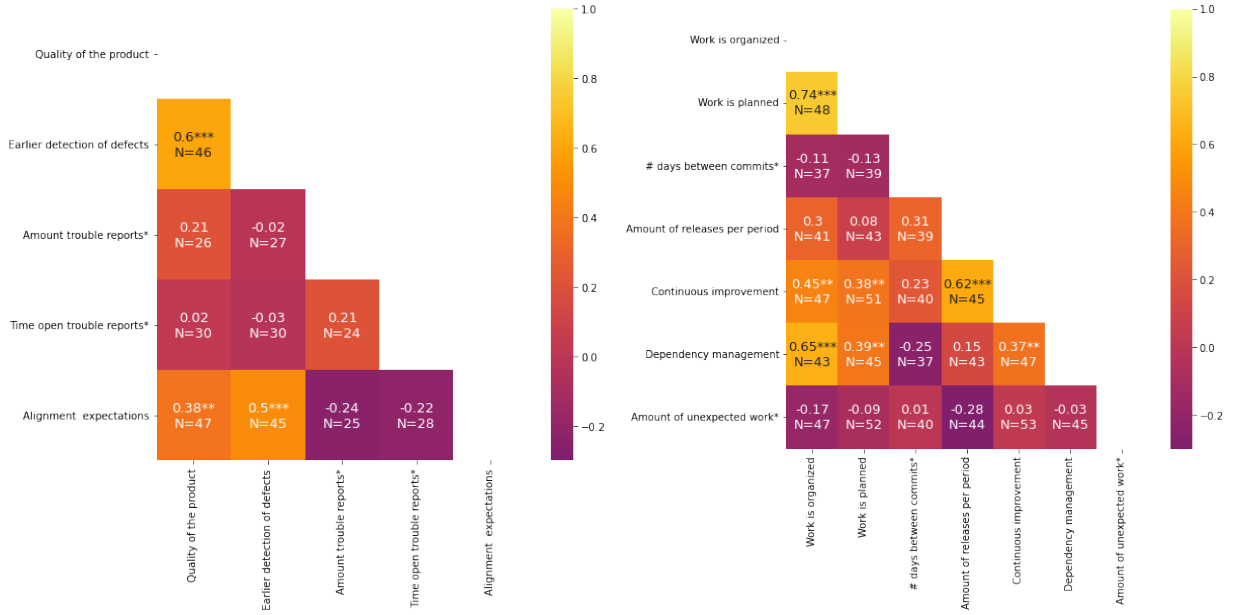
4.4.2 Team and Product Level Impact

Table 6 shows a wide range of statistics for each individual team and product metric, categorized by their impact dimension. Although the amount of responses is 61, the amount of individual responses per metric can differ. This is because the "Don't Know" option does not count towards the total for a metric. This can give some additional insight on what metrics are recognized by practitioners and which are not. The "Total amount of external trouble reports" metric is recognized by less than half of the total respondents. The "Time external trouble reports remain unsolved" metric is recognized by just over half of the respondents.

Other statistics presented in this table are the mean answers, the standard deviation, the answer with the lowest number, the answers were respectively 25%, 50% and 75% of the responses are represented and the answer with the highest number.

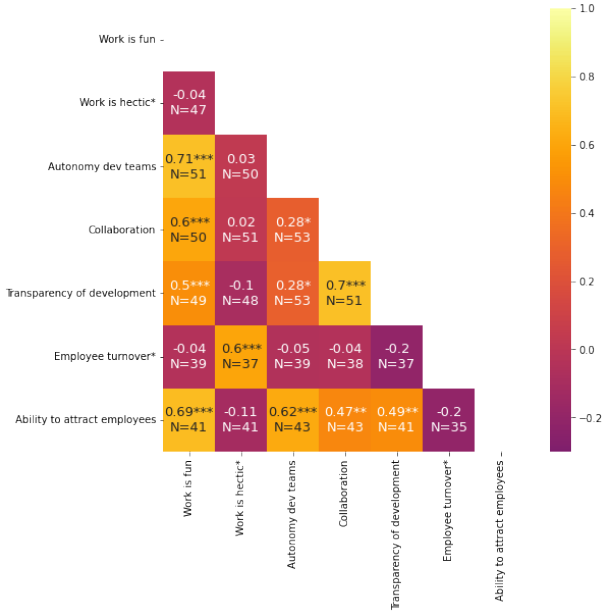
Metrics that have an "*" at the end of them are metrics where a negative answer indicates a positive impact. That is to say, a negative answer to the "Lead time per feature" metric would mean a decrease in the amount of lead time per feature which is a positive impact. Respondents to the survey were made aware of this fact through the means of an example but these explicit markings present in the table were not present in the survey. Interestingly all the means of all these marked metrics are still positive although these metrics have some of the lowest means such as "Employee turnover" and "Amount of unexpected work". On the opposite end there is not a small group of metrics that stand out in how high their mean is, with 9 of 26 (34.61%) metrics having a mean between 30 and 35.

The Spearman correlation between maturity of each organizational level and the metrics is shown in figure 10. Although are are quite some significant Spearman correlation coefficients (marked with a "*" for $p < 0.05$ or with a "**" for $p < 0.01$), no metrics seem to benefit from an increase in maturity for all three organizational levels. Some other things of note are metrics that have positive correlation on some levels but negative correlation on others, such as the "the degree to which work is hectic" with a relatively high N and the "customer service request turnaround time" with a relatively low N.



(a) Quality dimension

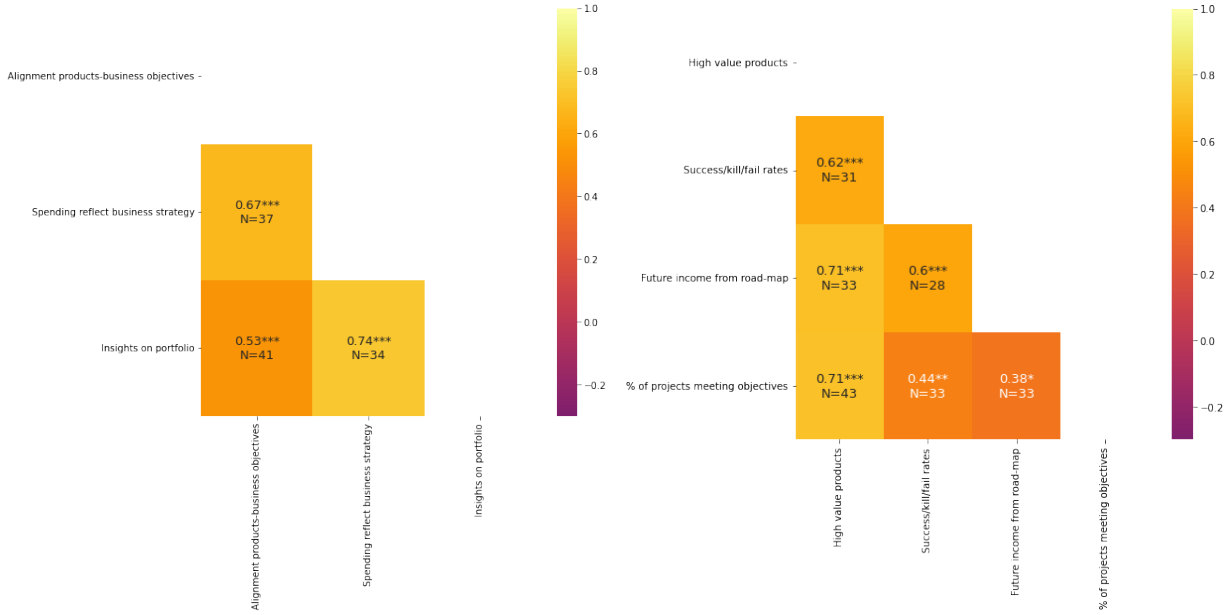
(b) Workflow Health dimension



(c) Employee Satisfaction & Engagement dimension

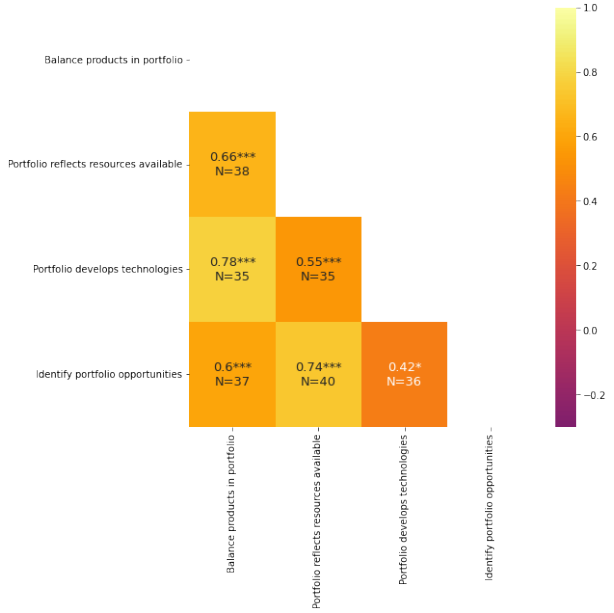
Figure 8: Pearson correlation coefficients for the quality, workflow health and employee satisfaction & engagement metrics. A single "*" means significance at the 0.05, a "**" means at the 0.01 and "***" means at the 0.001 level.

As explained in section 3.3.3, Spearman correlation was chosen over Pearson due to the nature of the maturity model. As such a direct comparison to Stettina et al. [2021] is not possible, as they use Pearson correlation. Both the Spearman and the Pearson correlations of the results obtained in this thesis, were compared against each other. There are few notable differences: "Amount of releases per period" and program maturity is significant at the 0.01 level with Spearman correlation, as opposed to 0.05 for Pearson. The same is true for both the "Autonomy of development teams" and the "Ability to attract new employees" metrics, whose Spearman correlation with program maturity is also 0.01. For all of these correlation pairs, the coefficient is also slightly higher with Spearman correlation. There is only one correlation pair that decreases in both significance and coefficient, which is the "Overall customer satisfaction" which now only



(a) Business alignment dimension

(b) Financial performance dimension



(c) Portfolio balance dimension

Figure 9: Pearson correlation coefficients for the portfolio metrics. A single "*" means significance at the 0.05, a "**" means at the 0.01 and "***" means at the 0.001 level.

correlates with portfolio maturity at the 0.05 level. Other than that no differences in 0.01 level significance correlations were found between the two methods.

A comparison between the survey results and the results obtained by Stettina et al. [2021] for the means of metrics that were included in both studies can be seen in table 7. Across all dimensions and metrics a decrease in the range of $[40, 60]$ percent can be seen in the survey results, with the exception of the "The degree to which work is hectic" metric which decreased by 83.75%. The results in Stettina et al. [2021] were also obtained through an online survey but there were differences between the two methodologies. Respondents to their survey could not indicate a negative impact, as the range of the scale began at 0%. The results of the survey can be adjusted for this by mapping each mean to a value

Table 6: Team and product level metrics general statistics. If a metric is marked with an "*" it means that a negative score means a positive impact.

Dimensions/Metrics	N	mean	std	min	25%	50%	75%	max
Productivity								
Effectiveness of development	52	35.37	28.45	-20	20	31	45	100
Efficiency of development	52	31.75	27.39	-40	18	27	49	97
Responsiveness								
Lead time per feature*	47	28.85	32.60	-28	0	25	46	100
Customer feedback speed	45	33.76	26.24	0	20	30	43	100
Customer service request turnaround time*	38	22.21	36.00	-31	0	16	29	100
Quality								
Quality of the product	51	32.82	29.35	-20	15	25	53	90
Earlier detection of defects	48	35.54	32.70	-20	17	25	53	100
Total amount of external trouble reports*	27	9.07	36.38	-44	-16	0	28	100
Time external trouble reports remain unsolved*	31	9.45	23.04	-34	0	0	25	51
Alignment between the product and stakeholder expectations	54	34.24	26.89	-20	18	30	53	100
Workflow Health								
The degree to which work is organized	48	28.83	34.19	-30	0	25	56	100
The degree to which work is planned	53	30.43	34.01	-40	0	30	50	95
Number of days between commits*	41	18.00	43.23	-100	0	15	46	100
Amount of releases per period	46	35.13	40.10	-100	13	36	63	100
Allows for continuous improvement	55	32.45	26.54	-23	18	32	50	100
Dependency management	47	20.49	40.02	-100	0	21	40	100
Amount of unexpected work*	56	4.07	38.03	-100	-19	0	26	80
Delivery predicatability	56	30.59	32.52	-50	10	33	46	100
Employee Satisfaction & Engagement								
The degree to which work is fun	51	27.47	34.98	-43	0	23	50	100
The degree to which work is hectic*	51	6.78	28.61	-60	-10	0	22	81
Autonomy of development teams	55	31.16	37.03	-100	15	26	50	100
Collaboration	54	35.54	31.57	-41	19	38	52	100
Transparency of development	54	33.56	34.01	-34	10	27	61	100
Employee turnover*	39	7.72	25.59	-43	0	0	21	88
Ability to attract new employees	44	24.52	29.77	-38	0	21	41	100
Customer Satisfaction								
Overall customer satisfaction	49	25.41	24.28	-20	10	20	40	89

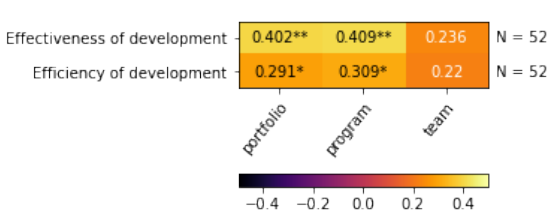
within the $[0, 100]$ range. This can be done by adding 100 to the means and then dividing them by two, as this maps all values in $[-100, 100]$ to a value in $[0, 100]$. Transposing the means in this manner results in more similar results.

The wording of the questions was also different due to the changes made in this study to avoid the possibility of double negatives in the answering mechanism, as per Roster et al. [2015]. The effect of having a slider with a range that includes negative values can also affect the obtained results [Schwarz et al., 1991, Tourangeau et al., 2007]. Therefore any comparison made between the results obtained in this study and those obtained in Stettina et al. [2021] must be made taking the context of both studies in mind. This will be discussed in further depth in section 5.1.2.

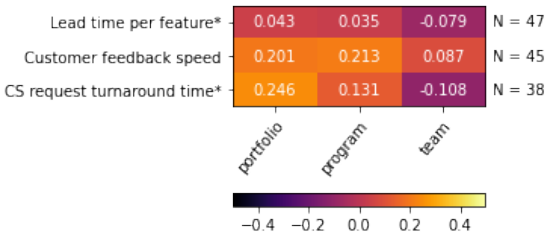
4.4.3 Portfolio Level Impact

Table 8 shows the same results for the portfolio metrics as table 6 did for team and product metrics. There is an overall increase of "Don't Know" answers for portfolio metrics, which can be seen by the decrease in N values. This suggests that respondents were less familiar with portfolio metrics than with team and product metrics. Here there are no metrics marked by an "*" because for all of these metrics a positive answer meant a positive impact.

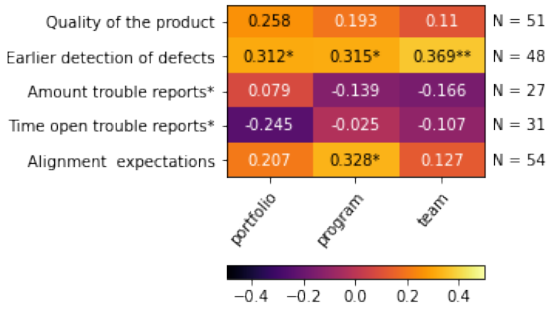
The portfolio metric with the lowest mean is "Success/kill/fail rates" with 12.53, the highest scoring portfolio metric is "Ability to provide insights on portfolio decisions across the organization" with 25.04. Also of note is that for every metric except "Ability to provide insights on portfolio decisions across the organization" the 25^{th} percentile of the respondents is zero, which also was the starting position of the slider.



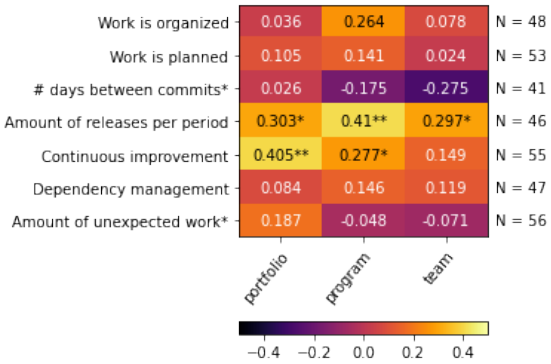
(a) Productivity dimension



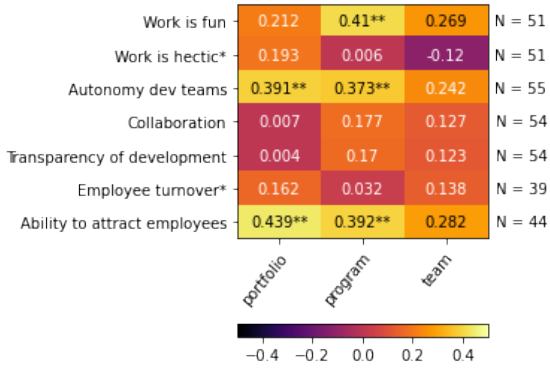
(b) Responsiveness dimension



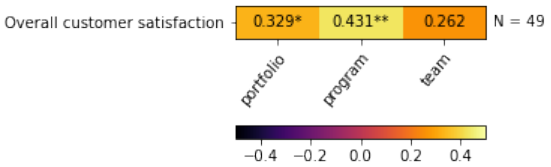
(c) Quality dimension



(d) Workflow Health dimension



(e) Employee Satisfaction & Engagement dimension



(f) Customer Satisfaction dimension

Figure 10: Spearman correlation coefficients between team and product metrics and each of the three organizational levels. A single "*" means significance at the 0.05 level and a "**" means at the $p < 0.01$.

As with the team and product metrics, the Spearman correlation between the portfolio metrics and maturity levels was calculated. The reason that these are not shown in this section is that none of these correlations were significant, not at the 0.01 level nor at the 0.05 level. This means that there were no significant correlations found between portfolio maturity and portfolio metrics, in contrast with the team and product metrics which had several significant correlations with portfolio maturity. Due to this lack of significant correlations and to safe space, the correlation matrix for portfolio metrics is omitted from this thesis.

4.5 Maturity Graded Response Model Results

Tables 9, 10 and 11 show the results of the GRM model on the various elements that constitute the maturity model by Laanti [2017] (shown in figure 1). The tables show the portfolio, program and team levels respectively and these should be interpreted independent of each other, as they were used in their original model.

To understand these results it is important to know how respondents were asked about these practices. For each practice a respondent could answer with six options: "Never", "Seldomly", "Sometimes", "Frequently", "Always" and "Don't Know". The first five options should be seen as a rise in implementation for each step, "Don't Know" answers are

Table 7: Comparison between the survey results and the study by Stettina et al. [2021] who collected data in 2018. Translated means x_t are defined as $x_t = (x + 100)/2$ where x is the regular mean.

Impact Metrics	Survey	Stettina et al. [2021]	Change %	Survey Translated	Change % Translated
Effectiveness of development	35.37	60.58	-41.63%	67.68	11.72%
Quality of the product	32.82	61.32	-46.47%	66.41	8.31%
Lead time per feature	28.85	66.70	-56.75%	64.43	-3.41%
Collaboration	35.54	74.32	-52.18%	67.77	-8.81%
The degree to which work is fun	27.47	63.48	-56.73%	63.74	0.40%
The degree to which work is planned	30.43	55.56	-45.23%	65.22	17.38%
The degree to which work is organized	28.83	57.02	-49.44%	64.42	12.96%
The degree to which work is hectic*	6.78	49.50	-86.29%	53.39	7.86%
Earlier detection of defects	35.54	66.94	-46.90%	67.77	1.24%
Transparency of development	33.56	70.13	-52.15%	66.78	-4.78%
Autonomy of development teams	31.16	63.48	-50.91%	65.58	3.30%

Table 8: Portfolio level metrics general statistics.

Dimensions/Metrics	N	mean	std	min	25%	50%	75%	max
Business Alignment								
Alignment between products and business objectives	47	24.28	27.08	-29	0	21	41	100
Ability to have spending reflect business strategy	38	19.26	23.06	-34	0	15	29	77
Ability to provide insights on portfolio decisions across the organization	46	25.04	25.96	-15	7	20	37.75	100
Financial Performance								
Portfolio contains high value products	49	23.78	26.10	-26	0	20	41	100
Success/kill/fail rates	34	12.53	17.68	-18	0	12	21	74
Projected future income from road-map	34	16.26	23.38	-20	0	7	32	80
Percentage of projects meeting objectives	49	18.76	28.23	-49	0	15	33	100
Portfolio Balance								
Good balance of products in the portfolio	42	15.69	29.10	-40	0	10	30	100
Portfolio reflects resources available	46	13.48	26.89	-54	0	10	29.25	79
Portfolio develops technologies and competencies	40	19.75	25.66	-30	0	16	27.5	100
Ability to identify new portfolio opportunities	43	14.56	25.90	-53	0	12	25.5	81

ignored. "Extrmt1" is the level of maturity that a respondent must have for a probability of 50% of a "Never" answer, "Extrmt2" is the level of maturity a respondent must have for a probability of 50% of a "Never" or a "Seldomly" answer, etc. These maturity levels are shown as a percentage of the highest required level in an individual table, that is to say a value of 0.73 means that the required maturity level is 73% of the highest maturity level in that table. This is done because the purpose of this analysis is to test if the order used in the original model is valid. Therefore it is only of interest how the practices compare against each other. Since the original model requires total implementation of a practice for the user to move on to a new maturity stage, the practices in the table are ranked by Extrmt4. This column shows the maturity required to have a 50% probability to have a "Frequently" or lower answer, which also means that it is the level at which a respondent has a 50% chance of answering "Always". This means they have fully implemented the practice. As a result the table is sorted from most difficult to implement to easiest to implement, reading the table top to bottom. So for the portfolio layer the most difficult practice to fully implement is "Ability to innovate new businesses" and the easiest practice to fully implement is "Backlog tool support in use".

For the GRM model to produce any output it requires that every possible answer occurs at least once for every question. That is to say, every question needs to have at least one response each of "Never", "Seldomly", "Sometimes", "Frequently", "Always". This was not the case for the data obtained through the survey. Since the model cannot produce an output without satisfying this precondition, "dummy entries" were added to the data-set. Five dummy entries were added, one for each of the possible valid answers. Each dummy entry contained only answers of its respective answer type for each question. To address concerns of how this might affect the output of the model, it was tested

what happened if progressively more of these dummy entries were added in sets of five (again one for each answer). Although this did affect the absolute ability values and the "Dscrmn" values it did not affect the relative ability values of the practices. This is the data of interest since, to test the validity of the maturity model, the relative ability values are of interest.

The "Dscrmn" column indicates how well a practice is able to distinguish between respondents with low maturity and respondents with high maturity. That is to say, if a practice has a high "Dscrmn" value then there is a relatively high chance for a low maturity respondent to answer differently than a high level respondent. A practice with a low "Dscrmn" value is less likely to be able to make a distinction between low and high level respondents. As an example, on the program layer "Organization is networked" makes a very clear distinction between high maturity and low maturity respondents, whereas "Agile release trains in use" has high maturity and low maturity respondents answering similarly.

Behind every practice between square brackets is a number representing the maturity stage of that practice in the original model. The mapping used for this is the same one used in section 3.2.2. Since the tables are sorted by the required maturity to fully implement a practice in descending order, the more the numbers between brackets follow the same descending order the more the GRM results reflect the order used in the original model.

Table 9: Normalized graded response model results for all portfolio practices. Numbers between square brackets indicate the stage of the practice in the original model. Practices are sorted on "Extrmt4" value, as a result if the results fully agree with the original model the numbers between brackets should descend in order.

Practices	**Extrmt1**	**Extrmt2**	**Extrmt3**	**Extrmt4**	**Dscrmn**
Ability to innovate new businesses [5]	0.0	0.4	0.78	1.0	2.03
Detecting utilizing fast business opportunities [4]	0.14	0.42	0.68	0.99	2.23
Systematic fast rolling decision making [2]	0.22	0.48	0.71	0.98	3.69
Measuring feedback guidance based on data [3]	0.21	0.44	0.68	0.95	2.26
Options thinking in decision making [3]	0.17	0.41	0.64	0.89	1.79
Agile metrics in use [2]	0.05	0.47	0.59	0.8	2.71
Agility is part of values [4]	0.21	0.42	0.52	0.8	2.43
Backlog prioritized [1]	0.21	0.4	0.5	0.76	3.77
Portfolio work is continuous [2]	0.18	0.37	0.55	0.75	2.99
Work identified as Epics features [1]	0.01	0.27	0.48	0.64	2.24
Backlog tool support in use [1]	0.03	0.33	0.5	0.64	2.56

Table 10: Normalized graded response model results for all program practices. Numbers between square brackets indicate the stage of the practice in the original model. Practices are sorted on "Extrmt4" value, as a result if the results fully agree with the original model the numbers between brackets should descend in order.

Practices	**Extrmt1**	**Extrmt2**	**Extrmt3**	**Extrmt4**	**Dscrmn**
Ability to create systems services previously impossible [4]	0.23	0.45	0.72	1.0	1.76
Organization is networked [5]	0.3	0.5	0.7	0.94	3.05
Ability to respond rapidly to changing needs [5]	0.0	0.37	0.65	0.93	1.65
Acceptance test before feature [3]	0.23	0.5	0.7	0.93	1.29
Organized for lean agile WoW [2]	0.22	0.49	0.65	0.92	2.58
Agile release trains in use [2]	0.21	0.5	0.64	0.9	1.08
Ability to embrace change [1]	0.04	0.38	0.55	0.9	1.94
Continuous positive feedback from customers [4]	0.24	0.43	0.58	0.89	2.82
Agile budgeting [3]	0.42	0.56	0.7	0.89	1.52
Agile metrics in use [3]	0.05	0.47	0.62	0.89	1.95
Systematically speeding up production releases [3]	0.35	0.47	0.67	0.88	2.62
Value stream thinking [2]	0.22	0.41	0.64	0.88	1.87
Networked leadership [3]	0.43	0.54	0.69	0.87	2.74
Product programs are agile [1]	0.26	0.43	0.61	0.86	2.77
Incremental demos guide development [2]	0.04	0.43	0.63	0.81	1.9
Incremental planning execution [1]	0.19	0.26	0.53	0.76	2.24
Agile roles in use [2]	0.11	0.32	0.52	0.75	1.67

There are some discrepancies on the portfolio and team layer between the results and the original model [Laanti, 2017]. On the portfolio layer the "Systematic fast rolling decision making" practice is higher in difficulty compared to the

Table 11: Normalized graded response model results for all team practices. Numbers between square brackets indicate the stage of the practice in the original model. Practices are sorted on "Extrmt4" value, as a result if the results fully agree with the original model the numbers between brackets should descend in order.

Practices	Extrmt1	Extrmt2	Extrmt3	Extrmt4	Dscrmn
Multiple releases per day [5]	0.36	0.51	0.71	1.0	1.15
No errors released [4]	0.15	0.36	0.58	0.98	1.44
Automatic testing integration deployment [2]	0.07	0.29	0.47	0.73	1.61
Systematically removing impediments [3]	0.0	0.31	0.46	0.73	1.83
Test first approach [3]	0.06	0.35	0.55	0.7	2.53
Fast fixes done as needed [1]	0.07	0.14	0.38	0.61	1.57
Scrum in use [1]	0.06	0.2	0.31	0.52	1.5
Dedicated build environment [1]	0.0	0.22	0.34	0.52	1.91
Version control in use [1]	0.08	0.19	0.3	0.49	1.67

original model, the "Agility is part of values" is lower. On the team level the "Automatic testing integration deployment" is higher in difficulty compared to the original model, although the difference in Extrmt4 value is not as stark compared to the portfolio discrepancies. For the most part the results seem to align with the original model.

There are more discrepancies on the program layer, especially practices with an Extrmt4 value in the range [0.86, 0.94]. This subset of program practices include thirteen out of the total of seventeen practices. Practices in this subset are close in difficulty and there are many practices that are out of order compared to the original mode. Furthermore, the practice that scores the highest in difficulty in the results, the "Ability to create systems services previously im- possible", is only in the second highest stage in the original model.

4.6 Cluster Statistics

The cluster questions at the very end of the survey can be divided into three categories: questions that ask about measuring an organization's transformation, questions about certain aspects of the organization and questions about how an organization implements certain agile aspects. The results of these categories will be shown in section 4.6.1, 4.6.2, 4.6.3 respectively. Unlike questions before this section, respondents did not have to answer these questions to be able to complete the survey. As a result the amount of answers to each question can differ, this amount is mentioned in the caption of each figure. The intent of these cluster questions was to use it to slice the data-set, with the aim of comparing groups of respondents for which scrum master was a part-time role against those for which it is a full-time role. Unfortunately with the amount of respondents being lower than expected, this would result in very small slices. Because of these reasons these comparisons were not done as part of this thesis.

4.6.1 Measuring the Transformation

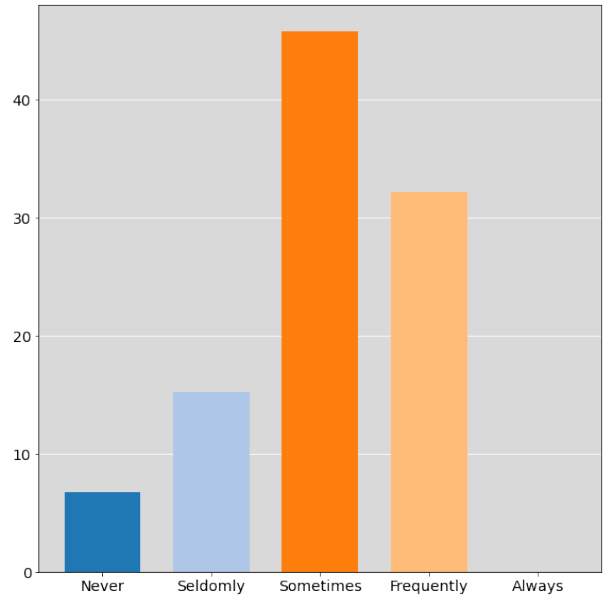
Respondents were asked how often they measure their transformations (figure 11a) and how often they steer their transformations based on these measurements (figure 11b). The distribution of answers to these questions have the same descending order: Sometimes, Frequently, Seldomly, Never, Always. Where in the first question no respondent answered that they always measure their transformation. 6.78 of respondents never measure their transformation, 8.47 never steer their transformation based on measurements.

4.6.2 Organizational

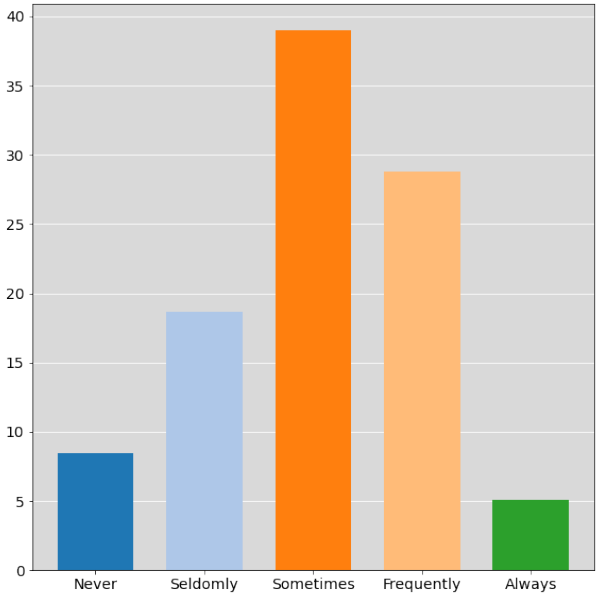
This category consists of two questions: "Does your organization maintain/have a line organization next to its team structure?" (figure 12a) and "What percentage of teams in your organization is outsourced?" (figure 12b). The first question shows that 79.31% of the respondents maintain a line structure next to a team one. The second question shows that, with 55.0% of the respondents, more than half the respondents' organizations only outsource $0 - 25\%$ of their work in terms of employees.

4.6.3 Agile Implementation

This third category consists of three questions: "What kind of role is the Scrum Master in your organization?" (figure 13a), "What is the work distribution like for part-time Scrum Masters?" (figure 13b) and "What is your average story size?" (figure 13a). The second question was only shown to respondents who answered that the Scrum Master role is a part-time role alongside other tasks on the first question.

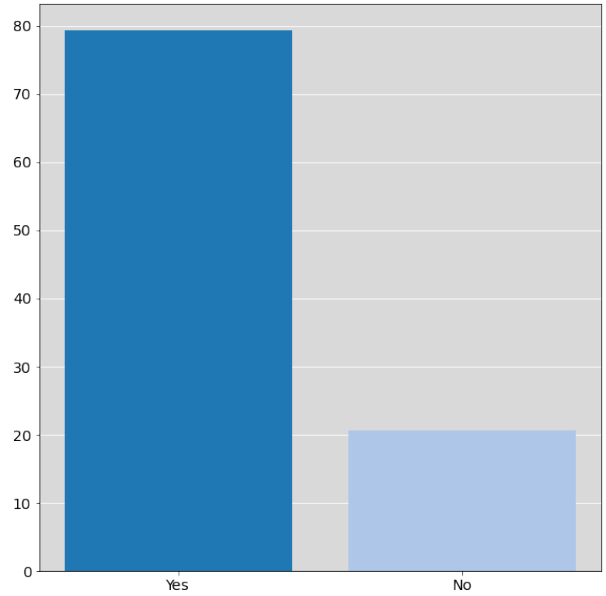


(a) How often do you measure the transformation? No "always" answers were given.
$N = 59$"

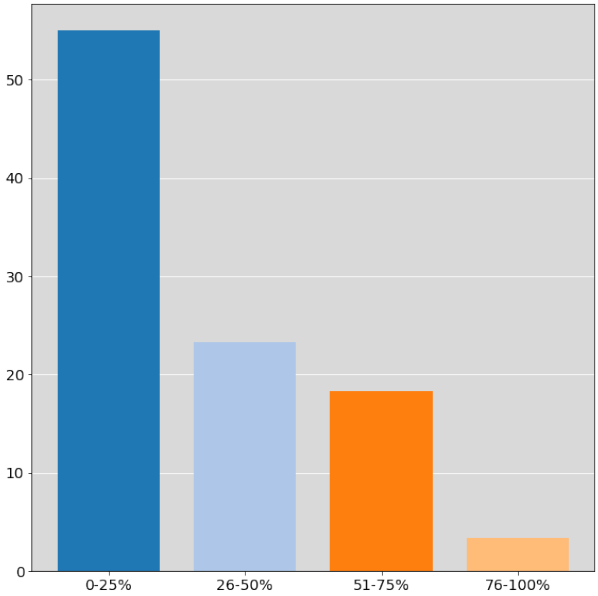


(b) How often do you steer based on measurement.
$N = 59$

Figure 11: Distribution in percentages of cluster questions about measuring the transformation



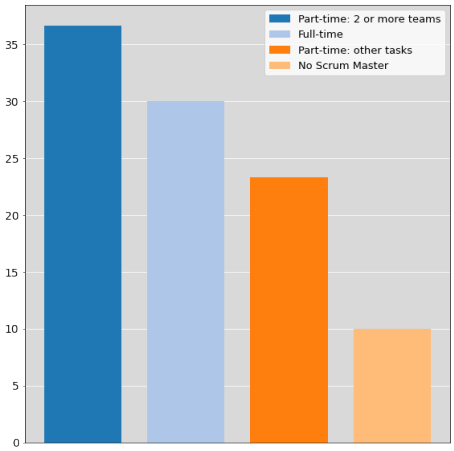
(a) Line structure next a team one.
$N = 58$



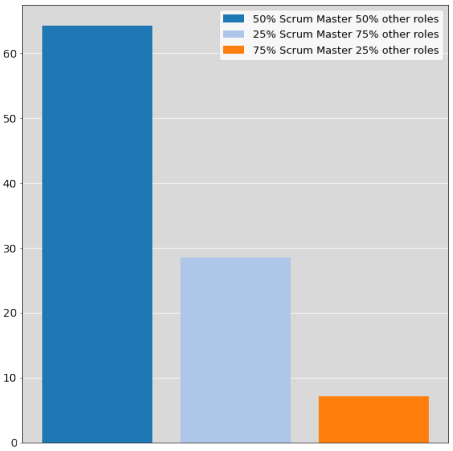
(b) Percentage of outsourced employees.
$N = 60$

Figure 12: Distribution in percentages of cluster questions about the organization.

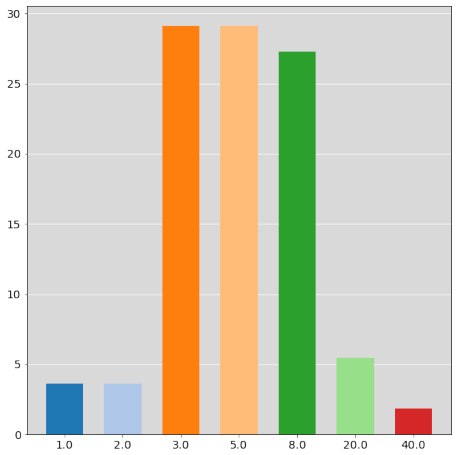
A possible solution to this issue would be replacing the slider scale with a radiobutton based Likert scale, as was done in Laanti et al. [2011] and Putta et al. [2021]. Rather than asking respondents directly about the impact on various metrics where depending on the metric a positive or negative answer can mean different things, respondents would be asked if they agree or disagree that



(a) Scrum Master role.
$N = 60$



(b) Scrum Masters work distribution.
$N = 14$



(c) Average story size.
$N = 55$

Figure 13: Distribution in percentages of cluster questions about the agile implementation of the organizations. Figure 13b is only shown to respondents who answered that Scrum Master is a part-time role with other tasks.

Figure 13a shows that 36.67% of respondents are in an organization where Scrum Master is a part time role, where Scrum Masters are responsible for two or more teams. For another 30.00% Scrum Master is a full-time role.

Figure 13b shows that if respondents were in an organization where Scrum Master was a role shared with other tasks, for only 7.14% the Scrum Master is the majority of their tasks.

Figure 13c shows that the most common average story sizes are three and five, both with 29.09%. Third is an average size of eight which is the average story size for 27.27% of the respondents. Only 7.27% of respondents have an average story size of 20 or more.

5 Discussion

5.1 Guidance for a Data Driven Performance Management Framework

The impact of LSA transformations on organizational performance has been measured before in academic literature, as discussed in section 2.1.2. There often is some overlap in the used metrics but the complete set of metrics across all literature is large. Many of the used metrics are also used in only one academic source, as can be seen in table 3. In this thesis all of these metrics have been collected and put into one large performance framework. With the results collected with this framework, several things can be learned and improvements can be made. The performance framework from section 3.1 is further refined based on these results. Metrics that are strongly correlated with each other are trimmed in such a way that only one metric of each pair remains. Furthermore, metrics that appear to be less recognized are also removed. The resulting framework is shown in figure 14. For practitioners and academics to be able to use this framework, the individual metrics have to be translated into survey questions. Based on obtained results, several guidelines are also given in this section. Academic literature and the results show that radiobutton, sliders and visual analog scales are all valid choices, but they also come with their own caveats [Cook et al., 2001, Couper et al., 2006, Roster et al., 2015, Simms et al., 2019]. Regardless of the type of scale used, any form of double answering mechanism should be avoided.

5.1.1 Framework Metrics

For the team and product metrics the results show that there are various strongly correlated metric pairs which are also significant at the 0.001 level, as shown in section 4.4.1. One such pair is the "efficiency of development & effectiveness of development" pair, which has a coefficient of 0.74. Definitions of the term "effectiveness" do exist such as the one from Fagerholm et al. [2015]: "*Effectiveness refers to accomplishing the right objectives e.g. those that have the greatest value.*". From practitioner feedback it was indicated several times that this term can be interpreted many ways. Perhaps respondents were not able to properly differentiate between effectiveness and efficiency. The same principle might be at play with the "work is planned" & "work is organized" pair with a coefficient of 0.74, who are also similar to each other in how they are worded. This was also a critique from practitioners when gathering feedback on the framework. Due to their high correlation and seeming ambiguity to practitioners, the use of these metric pairings in the same survey is questionable.

Another piece of valuable data when it comes to further improving the framework is the amount of "Don't know" answers from table 6 and table 8. From these tables it becomes apparent that the term "external trouble report" is not widely recognized. The metric "Total amount of external trouble reports" has answers by less than half of the total respondents. The metric "Time external trouble reports remain unsolved" has answers by just over half of the total respondents. In initial feedback on the framework practitioners also questioned the use of the term. They understood its meaning but doubted if the term had widespread use and was therefore recognizable. These two metrics are also likely candidates for removal.

Overall the portfolio metrics seem to be more correlated than their team and product counterparts, as shown in figure 9. Even though there are in total less portfolio metrics, there are more portfolio metrics with a coefficient above 0.7. Amongst respondents the familiarity of portfolio level of the transformation is lower than their familiarity with the team and product level, as can be seen in figure 5. This could mean that for the portfolio metrics, respondents were less able to distinguish between the various metrics. If this is not the case and respondents were able to distinguish properly between the metrics, it would imply that there is a relatively large amount of overlap in the underlying principles that are being measured by these metrics. In both cases a different composition of portfolio metrics might be beneficial when measuring the impact of a LSA transformation on the portfolio level. What further puts the used set of portfolio metrics in doubt, is the lack of correlations between portfolio maturity and the portfolio metrics. There is no data amongst the results that suggests concise changes amongst the portfolio metrics. There are metrics that have a lot of "Don't Know" options, but overall the portfolio metrics are less recognized when compared to their team and product counterparts. When further improving the framework, making major changes to the portfolio metrics should be considered.

Taking all of these possible improvements in mind a new, more concise, framework can be created. This framework contains only team and product metrics, this is due to the ambiguous nature of the portfolio metrics results. This ambiguity of the portfolio metrics results is further discussed in section 5.2.2. The final framework is shown in figure 14. In this improved version the effectiveness metric from the productivity dimension has been removed, as well as the organized metric from the workflow dimension. They are now represented by the metric with which they have a high correlation. Efficiency was chosen over effectiveness due to its less ambiguous meaning and due to its more frequent appearance in literature (see table 3). Planned was chosen over organized as the organized metric was only used once in literature. On top of this both metrics containing the term "external trouble report" have been removed. This framework

can be used by academics and practitioners alike, when trying to measure the impact of an LSA transformation on organizational performance through the means of a survey.

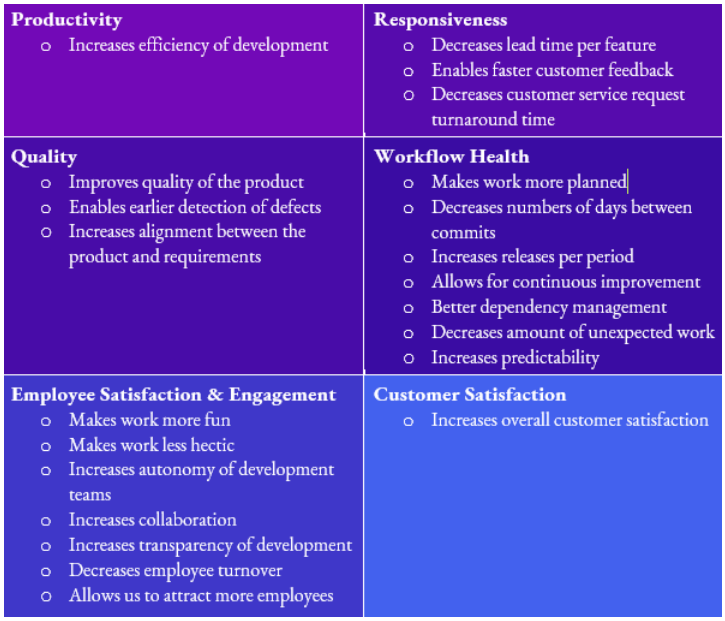


Figure 14: Organizational performance framework after adjustments made based on results.

5.1.2 Framework Survey Design

When translating this framework of metrics into a survey various design choices can be made. The results of this thesis can also offer guidance in this process. One design choice is which type of scale should be used and what its accompanying labels should be. A criticism that Stettina et al. [2021] make about their own survey is that it does not allow respondents to indicate a decrease in performance. Therefore the slider in this thesis can go into the negative, effectively doubling the amount of options for respondents.

Besides doubling the amount of options there is another consequence of this decision, there is now a possibility of a double negative in the answering mechanism. Double negatives in answering mechanisms can lead to confusion amongst respondents [Roster et al., 2015], something that was also noted during testing of the survey. To avoid such confusion each metric was proposed in a neutral manner. For example, when asking about the impact on collaboration Stettina et al. [2021] presented the following question to respondents: *"In your opinion, what is the impact of a scaled agile transformation. Agile development has increased the following topics with what percentage: Increases collaboration"* whereas in this study this became: *"In your opinion, from -100% to 100%, how has your organization's large scale agile transformation impacted the following aspects? Collaboration"*. As a result this meant that for some metrics respondents had to give a negative number in order to indicate a positive impact. As an example: a decrease in the amount of unexpected work is an increase in performance.

Given the obtained results (see table 6) it is unclear whether this decision had the intended effect. Overall these metrics score lower than their counterparts: the average impact of all metrics where a negative number indicates a positive impact is 13.27, where for all other team and product metrics this was 30.59. At the same time none of the results for these metrics indicate a positive impact whereas all other metrics do. This can be due to a variety of underlying reasons, it could be that these metrics experience less or no positive impact as a result of LSA transformations. This seems unlikely as this contradicts results previously obtained by Laanti et al. [2011], Olszewska et al. [2016] and Stettina et al. [2021]. However if respondents were more confused with these questions a higher standard deviation is also to be expected which is not the case. The correlation matrix (figure 10) between the maturity levels and the metrics tells a similar story. In total there are 20 correlations at the 0.5 level, but none of these is a correlation with a metric where a negative number indicates a positive impact. The same can be seen when looking at the correlation between metrics. If respondents answered as expected, metrics with a "*" should correlate negatively with metrics without a "*". Often such pairs do have lower coefficients but many of these are not significant. All of this means that it is inconclusive whether or not the decision to present all metrics in a neutral manner was effective.

A possible solution to this issue would be replacing the slider scale with a radiobutton based Likert scale, as was done in Laanti et al. [2011] and Putta et al. [2021]. Rather than asking respondents directly about the impact on various metrics where depending on the metric a positive or negative answer can mean different things, respondents would be asked if they agree or disagree that the metric has been affected positively. In this manner the issue of double negatives in the answering mechanism would be avoided along with any ambiguity it might cause. Both a radiobutton scales and slider scales can yield reliable results that are similar in distribution [Cook et al., 2001, Couper et al., 2006, Roster et al., 2015, Simms et al., 2019]. The results of this survey suggest the same, when adjusting for the scale of the answering mechanism both Laanti et al. [2011], Stettina et al. [2021] and the results of this thesis all fit within similar ranges. Such a comparison is shown in table 7. As such both academic literature and results over multiple studies suggest that respondents do not answer based on the value labels of the scale, but rather based on the relative position of the scale.

The results of this thesis also provide insight on another survey design choice, that of the slider starting position. When looking at the obtained results on the impact metrics in table 6 and in table 8 it can be seen how often 0% appears as the border of the 25% or 50% of answers. Meaning that for a lot of metrics 25% or 50% of the answers are equal to or lower than 0%. This is especially apparent for the portfolio metrics shown in table 8 where all metrics except for the "Ability to provide insights on portfolio decisions across the organization" metric have 0% as their 25% limit. This indicates that the answer of 0% is unusually well represented in a data-set where the possible range of answers is $[-100, 100]$. When looking at how often 0% is given as an answer this also becomes apparent, on average 0% makes up over 30% of the total amount of answers for portfolio metrics. All of this is of interest since this is also the starting position of the slider in the survey.

Respondents were not required to move the slider for it to be a valid answer, as the starting position of 0% was a valid answer. As a result it is not possible to distinguish between intentional 0% responses and non-responses [Buskirk, 2015]. The survey did include an explicit "Don't Know" option so that respondents could indicate intentional non-response, but there is no consensus on the effect this has on missing data and reliability [DeCastellarnau, 2018]. Therefore it would be worth considering methods that can avoid such ambiguity. The earlier proposed alternative of a radiobutton based Likert scale would be a solution, as these always start without a valid answer. Another solution would be the use of a visual analog scale instead of a slider, where a respondent starts with an empty slider and uses a single click or pen stroke to indicate their response. The most analogous solution to the slider scale used in this survey would be to require respondents to move the slider at least once for it to count as a valid answer.

All of the above provides further insight for translating a performance framework into a survey, and can provide guidance in this process. There is the option of a radiobutton scale, slider scale and a visual analog scale. All of these are valid, reliable and comparable in distribution [Roster et al., 2015]. The results of this thesis show that a numerical scale with a range into the negative should be avoided. When using a radiobutton scale it is important to have at least six options [Simms et al., 2019]. When using a slider scale some action is required to prevent default answers from being valid answers, as can be seen in both the results and in Buskirk [2015].

5.2 Comparing Apples and Pears: Impact Patterns From Previous Studies

In this section the obtained results transformation impact will be discussed. This includes the impact means per performance metric and also the relation between transformation maturity and organizational performance. The results show three main findings. First of all a trend becomes apparent when comparing the results of this thesis against results from previous studies. Collaboration and transparency consistently rank amongst the most impacted metrics in survey studies, work is planned/organized consistently ranks amongst the least impacted metrics in survey studies [Laanti et al., 2011, Putta et al., 2021, Stettina et al., 2021]. The collaboration and transparency metrics are also often mentioned in studies that are not survey based [Gustavsson and Bergkvist, 2019, Laanti and Kettunen, 2019]. Secondly the correlations between transformation maturity and impact metrics suggest that there are some metrics that experience near immediate impact. Looking at the scatterplots (figure 15) of the "effectiveness of development" metric and the "collaboration" metric shows that at low transformation maturity there is already an impact of around 20%. Finally the results show that when adjusted for the scale used in different studies, the range of impact means across various studies is comparable.

5.2.1 Impact on the Team and Product Levels

When looking at the results of the impact metrics almost all results seem to indicate a positive impact. There is one set of metrics for which this is not the case, the results for all metrics where a negative number response indicated a positive impact (marked with an "*" in table 6). The results indicate that all of these metrics do not benefit from a large-scale agile transformation. The underlying cause of this might have something to do with how respondents were asked about these metrics, this is discussed in further depth in section 5.1.2. For now it should be said that the construct validity of these metrics is questionable. As such from here on out when any general statement about the results of the

impact metrics is made, the metrics marked with a "*" are not included in that statement. This statement is made once now to avoid redundancy later on.

When comparing the values of the means with results obtained in other survey based research, the results of this thesis seem less optimistic regarding the impact of LSA transformations [Laanti et al., 2011, Putta et al., 2021, Stettina et al., 2021]. This is also shown in table 7 by making a comparison with Stettina et al. [2021], which also uses a slider scale with percentage labels. In this same table a possible explanation is also given, showing that if you translate the values in a way that maps every value from a $[-100, 100]$ range to a $[0, 100]$ range you get more comparable results. This suggests that the underlying reason might be the difference in scale used, and also gets the resulting means closer to those in Laanti et al. [2011] and in Putta et al. [2021]. These papers use a Likert scale with agree/disagree labels instead of a slider scale with percentage labels, but translating the Likert scale results to a percentage yields similar results. Performing such translations does raise questions about impact the various answering mechanisms and scales have on respondents, does it change their answering patterns? Both Roster et al. [2015] and Simms et al. [2019] found no significant difference between the use of a radiobutton scale and a slider scale, although it should be noted that neither studies were done in an organizational performance context. As such there is a collection of results obtained over various studies and academic literature that both suggest that there is no difference in using a Likert or slider scale. This suggests that generally respondents pay less attention to the labels attached to the scales, rather they answer based on the position of their answer relative to the maximum possible answer.

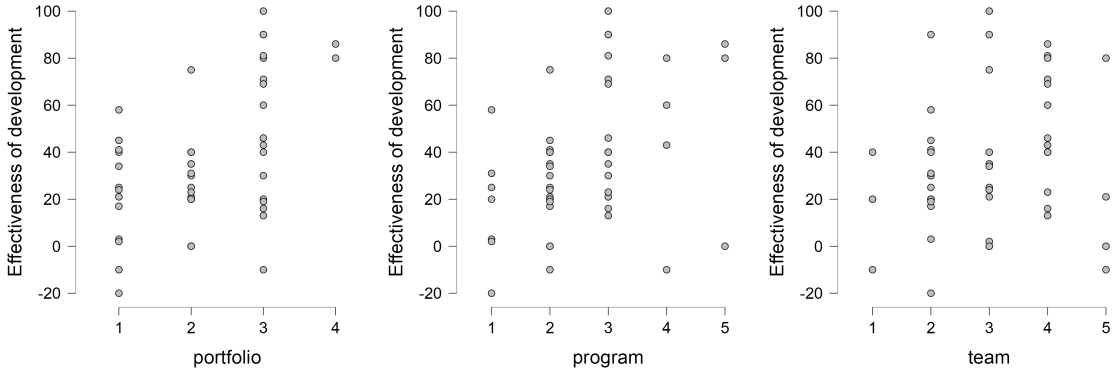
There is literature that found that having a slider with a scale that goes into the negatives does yield different results than a slider with a scale that only contains positive labels [Schwarz et al., 1991, Tourangeau et al., 2007]. In their research Schwarz et al. [1991] compared two ranges: $[0, 10]$ and $[-5, 5]$. Despite having the same number of options to choose from, the average response was higher with the $[-5, 5]$ range. Respondents were less likely to give an answer with a negative value. The comparison of this study and Stettina et al. [2021] is not a direct parallel because the amount of options are not the same, rather this study has double the amount of options. The effect observed in Schwarz et al. [1991] and Tourangeau et al. [2007] cannot be observed across all variables, as shown in table 7. Most metrics do have higher means but there are also some metrics with lower means, even though the average respondent's transformation maturities are comparable between the two studies (table 5). When comparing the average impact across all metrics that are shared between Stettina et al. [2021] and this thesis, a slight increase can be seen. The average of all of the impact means from Stettina et al. [2021] is 62.64, for this study the same average comes out to 64.83. This includes the "Lead time per feature" and the "The degree to which work is hectic" metrics, which are both metrics whose results suffer from the double negative answering mechanism. Without these two metrics the average of this thesis comes out at 66.15. The effect observed in Schwarz et al. [1991] and Tourangeau et al. [2007] might therefore be at play here as well.

Keeping the impact of survey design choices in mind, comparisons can be made to other academic literature on the impact of LSA transformations. When comparing the results of this thesis to previously obtained results, some patterns can be seen. Amongst Laanti et al. [2011], Stettina et al. [2021] and this thesis, the overall range of the means is comparable after translating each scale to the $[0, 100]$ range. For Laanti et al. [2011] this range is $[64, 73]$[4], for Stettina et al. [2021] this range is $[50, 74]$ and for this thesis this range is $[60, 68]$. The results of Putta et al. [2021] seem to indicate an overall higher impact, although this might be explained by their use of a five point Likert scale as opposed to six or more which is suggested by Simms et al. [2019]. When it comes to survey based surveys, both collaboration and transparency seem to be metrics that experience a big impact [Laanti et al., 2011, Putta et al., 2021, Stettina et al., 2021]. These two metrics are also often mentioned as benefits in literature that makes use of interviews or open ended questions [Gustavsson and Bergkvist, 2019, Laanti and Kettunen, 2019]. On the other end metrics about how organized or planned work is, seem to consistently rank amongst the metrics that experience the least impact [Laanti et al., 2011, Putta et al., 2021, Stettina et al., 2021]. As such this thesis has succeeded in improving the reliability of the findings surrounding LSA transformation impact that suggest: that there is an overall positive impact, that collaboration and transparency seem to be impacted the most, that how planned and/or organized work is seems to be impacted the least.

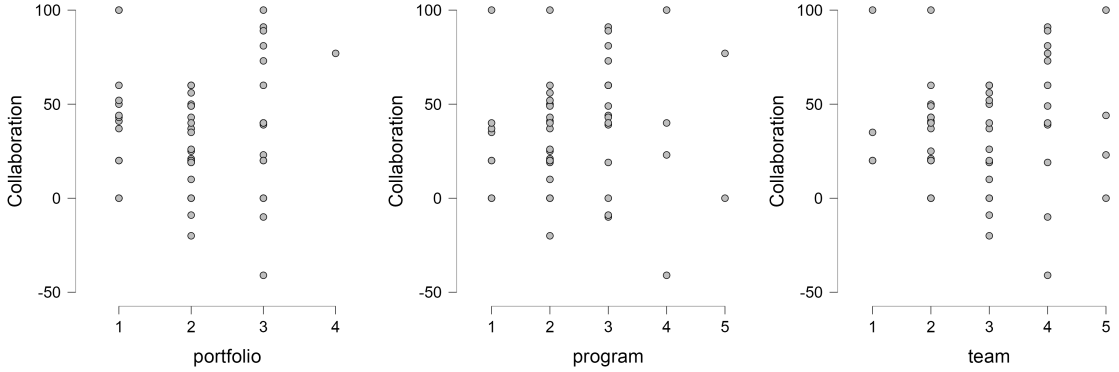
The relation between transformation maturity as specified by Laanti et al. [2011] and the metrics of the organizational performance, can be seen in the correlation matrix shown in figure 10. Due to the large number of relations tested the term significant will refer to significance on the 0.01 level in this context. The only correlation between a metric and team maturity with that level of significance is "Earlier detection of defects". This could be because the team level in the maturity model contains many practices dealing with testing such as "Automatic testing integration deployment", "Test first approach" and "No errors released". It is interesting that no other metrics seem to correlate with this level, one could argue that the same testing related practices should lead to more effectiveness or efficiency of development. Especially due to the high mean impact of effectiveness of development. A reason could be that this impact is experienced immediately at an early stage of the transformation. This coincides with the first stage of the maturity model, which

[4]Ignoring the "hectic" metric due to it also having the double negative answering mechanism and therefore questionable construct validity.

contains many practices which are essential to working agile. The collaboration metric shares this pattern: a high mean impact but no correlation, in this case not on any level. Possibly also sharing the same underlying reason, it was also prioritized as a basic aspect of agile in Turetken et al. [2017]. To see if it is indeed the case that the effectiveness of development and collaboration metrics immediately experience this benefit, their scatter plots against the team maturity can be seen in figure 15. In these figures it is shown that both metrics do seem to benefit immediately from stage two team maturity. The metrics do increase somewhat alongside team maturity but there is also a large spread of answers, this likely being the reason that the correlations are not significant. The same immediate benefit can be found on the portfolio and program levels, for both metrics. This indicates that the practices that belong to the second and above stages of team maturity, do not have that much of an impact on effectiveness of development. As opposed to their program and portfolio counterparts. The second portfolio stage contains practices such as "Agile metrics" and the second program stage contains practices such as "Organized for lean-agile way-of-working", which both can have an impact on development.



(a) Effectiveness - portfolio maturity. (b) Effectiveness - program maturity. (c) Effectiveness - team maturity.



(d) Collaboration - portfolio maturity. (e) Collaboration - program maturity. (f) Collaboration - team maturity.

Figure 15: Scatter plots for further analysis of the correlation between the various levels of maturity and the "effectiveness of development" and "collaboration" metrics.

The program and portfolio levels contain a larger amount of significant correlations, some of which shared between the two levels. Effectiveness of development is correlated with both levels. The practices and milestones in the program level offer more of an explanation than those in the portfolio level, it is not hard to see how the following could have an impact on development: "Incremental planning execution", "Incremental demos guide development" and "Acceptance test before feature". It could also be the case that if the program and portfolio level are behind in transformation maturity the full potential of development teams is not realized. These levels might need to progress alongside the team level in order for them not to become a bottleneck. Something similar might be happening with "Amount of releases per period", "the degree to which work is fun" and "Autonomy of development teams". The program level contains some practices and milestones regarding the organization of processes between teams such as the use of agile release trains. It could be that these enable teams to release more, have more fun and increase their autonomy but it could also be that the program level needs to keep up with the team level in a transformation.

The portfolio level contains only one significant correlation with an impact metric that is not shared with the program level: "Allows for continuous improvement". There are several practices and milestones in this level that might enable

this metric: "backlog tool support in use", "agile metrics in use" and "systematic fast rolling decision making" can all contribute to the ability of an organization to continuously improve itself. The program level also contains such elements but has no correlation, this level does correlate with the amount of releases per period metric. Perhaps the way this metric is formulated in combination with the fact that these two metrics were always presented in the same block (in random order within that block) lead to respondents associating the continuous improvement metric with the portfolio level. It does share the term continuous with the portfolio milestone: "Portfolio work is continuous".

5.2.2 Impact on the Portfolio Level

Compared to their team and product counterparts the portfolio metrics seem to experience less of an impact. Where the means of the team and product metrics lie within the range of $[20, 35]$, the portfolio metric means lie within the range of $[12, 25]$. This could be due to the lower portfolio maturity of respondents compared to team and program maturity, as shown in table 5. The relatively lower familiarity with the portfolio level of the transformation could also be a factor, which can be seen in figure 5. This would also fit with the role distribution of the respondents, executives made up less then 10% of the data-set. This unfamiliarity with the portfolio level of the transformation could result in respondents giving more reserved answers. Although there exists no significant correlation between portfolio maturity and portfolio metrics, nor between portfolio familiarity and portfolio metrics. Both of these were calculated for this thesis, but omitted due to the lack of relations between the data. It could also be that the selected metrics are not the correct metrics for measuring the impact of an LSA transformation. As discussed in section 2.2, very little academic literature about the impact of LSA transformations on the portfolio level of organizations. To still be able to make some measurements of the portfolio level, metrics were used that came from non-agile ways of working. It could be that these metrics are not impacted by LSA transformations. To be able to measure the impact of LSA transformations on the portfolio level, instead of using metrics from more traditional ways of working perhaps entirely new metrics need to be constructed.

5.3 Reflections on the Maturity Model: Graded Response Model Based Improvement

One of the aims of this study is to assess the validity of the agile maturity model created by Laanti [2017] (shown in figure 1). This is done by asking respondents about their implementation of the individual practices and milestones used in this model to rank maturity. The data gathered from this process can be used in combination with a graded response model (GRM) to see how the ranking used in the model compares to data obtained in this study. Overall the data from the GRM suggests that there are various improvements that could be made. Some of these issues might be solved by having somebody with prior knowledge of the original model help with the assessment by explaining certain elements. Although in its original paper [Laanti, 2017] the model is represented as a stand alone tool. One could also question the usefulness of a general purpose agile maturity model, as is done by various experts in Turetken et al. [2017]. These experts argue that dividing maturity into stages and categories shouldn't be attempted as agile transformations are a fluent process and any general purpose model would not be able to reflect that. However there can still be merit to such tools, not only for the purpose of making comparisons between academic papers but also as a moment of reflection, from which a more in depth discussion can start. If one does decide to use the Laanti [2017] model, various improvements can be made based on the GRM results. For the team and portfolio level these improvements are simple, using the GRM results several practices can be moved to different maturity stages. For the program level this process is more complicated. One result based solution would be the removal of certain practices. Doing so makes the distinction between groups of practices more defined, therefore making it easier to divide them back into stages. When removing practices likely candidates are practices that have similar "Extrmt4" values to other practices, but also a low "Dscrmn" score. For further academic research it would be interesting to see if a focus group study would produce similar results to those of the GRM. Participants of this focus group could be asked to rank or categorize these practices based on maturity.

The results of the portfolio level contain two main points of contention when comparing it to the original model. Most notably the practice of "Systematic fast rolling decision making" is ranked as a second stage maturity in the original model but requires almost the same maturity as level four and five stage maturity practices according to the GRM data. On the opposite end the "Agility is part of values" milestone is a fourth stage maturity milestone according to the model but ranks similar to first and second maturity stage practices. The latter might be explained by the fact that it is easy to have agility as part of organizational values on paper but much harder to determine whether these values are actually reflected in the daily way of working. Meaning that respondents might disagree with Laanti [2017] on what this exactly means. Another practice worth mentioning "Backlog prioritized" which is in stage one in the original model but ranks higher than other stage one practices.

On the program level there are many differences between the results obtained through the survey and the original model. First of all the most difficult milestone or practice according to the data is the "Ability to create systems services

previously impossible" milestone which in the original model is placed in stage four but here outranks both stage five milestones. Besides this top scoring milestone the following practices and milestones all score very closely to each other. These are the thirteen (out of seventeen total) practices that all score within the range $[0.86, 0.94]$. They also all differ 0.01 or less in relative maturity compared to practices and milestones directly above or below them. If one were to try to divide the practices and milestones into five categories again based on these results, dividing this middle section would be difficult. The practices and milestones in this section might therefore not be the best suited to have respondents rank their own individual program maturity. The bottom three practices are all stage one and two practices. Even though for the program level they are relatively distant from the other program practices and milestones in "Extrmt4" value, the overall range is still small when compared to the portfolio and team levels. This implies that for the respondents to this survey the used practices might not be a good fit to rank their maturity. A solution that is also supported by the data, is to use less practices and milestones for the program level. Currently the model uses seventeen total practices and milestones, more than the portfolio and team levels. By eliminating some of these the distinction between the maturity required to implement certain practices and milestones becomes more distinguished. As a result it would make it easier to categorize these sets which in turn might make the model easier to use for respondents. The most likely candidates for removal would be practices with low "Dscrmn" scores, as these are not able to distinguish between high and low maturity respondents. An alternative solution would be to look at entirely different program practices and milestones to assess maturity but with the data collected it is not possible to make any such suggestions.

The results of the final organizational level, the team level, has the most resemblance to the original model. Only one practice is out of order, the stage two practice "Automatic testing integration deployment. This practice is above the stage three practice of "Test first approach" and tied with the stage three practice "Systematically removing impediments". This is curious because one would assume automatic testing is required for a proper test first approach. Compared to the test first approach, the automatic testing practice does score lower for both "Extrmt2" and "Extrmt3" and is very similar in "Extrmt1". This implies that partial implementation of automatic testing is easier than partial implementation of a test first approach. Perhaps respondents were more hesitant to state that they always do automatic testing because there is at least one part of the pipeline that is not fully automated, whereas a test first approach is a less tangible matter. It should also be noted that the difference in "Extrmt4" values of the two practices is small, only 0.03. All other practices in milestones on the team level follow the order of the original model.

5.4 Threads to Validity

There are several possible threads to the validity of this research. For external validity there are two possible threads. First of all the data-set consists of only 61 entries, a majority of which were collected from The Netherlands. It is hard to make an estimate of how much of the population this covers, nonetheless this is a relatively low number of entries compared to Laanti et al. [2011] and Stettina et al. [2021]. This raises the question how generalizable these results are, especially outside of The Netherlands. Secondly, the distribution of the roles of respondents could also be a thread. There is a relatively small number of development team members amongst the respondents even though they should make up a larger part of the population. The most represented role in the data-set are agile coaches, which have an incentive to be positive about LSA transformations. It should be noted that Laanti et al. [2011] obtained similar results (when taking the translation in 7 into account) while their data-set consists of 41% of scrum team members.

Matters surrounding construct validity have previously been discussed in this thesis, as was done in section 5.1.2. Various analyses suggest that the survey design decision of including questions where a negative response indicated a positive impact, has led to questionable construct validity for these questions. The selected group of metrics to measure portfolio performance also seem to have questionable validity, at least as a tool to measure the construct of portfolio performance. This was also discussed in 5.2.2.

One interesting thread to validity is the maturity model by Laanti [2017], used to test correlation between transformation maturity and organizational performance. As shown in section 4.5 the results of this thesis are in conflict with some parts of the model. At the same time this model was used to measure the transformation maturity of a respondent which was then used in the correlation test. Therefore these correlations might not accurately represent the relation between transformation maturity and organizational performance.

5.5 Future Work

The results of this research can serve as a basis for several new studies. This is the first time such a comprehensive collection of metrics was used to measure organizational performance. The figures 7 and 8 show that there is still room for improvement. The team and product dimensions could be made more concise. For the portfolio level it would be interesting to see more drastic changes, as the used metrics did not seem the right fit to measure portfolio performance. Perhaps this should also be contained in its own study, as it proved difficult to get executives to participate in this study.

With the results of the GRM model it would be interesting to see if this was reproducible in a focus group context. This focus group would consist of various participants from various roles and organizations. They could be presented the same elements as in Laanti [2017] and asked to categorize them again. The results from this focus group can then be compared to the results obtained in this thesis through the GRM.

An interesting finding within the correlation results is that there appear to be very little significant correlations between maturity on the team level and performance metrics. This even though there are some metrics such as effectiveness of development which one would benefit from the practices inside the team level of the maturity model. Furthermore this metric does correlate significantly with the program and portfolio levels. One possible explanation is that the team level of an organization might be held back if the program and portfolio levels do not progress at the same rate. This would also fit with the higher team level maturity amongst respondents compared to their program and portfolio level maturity. To test this hypothesis, it would be interesting to see a study that also tests this correlation but makes a distinction between bottom-up and top-down transformation strategies. As this problem should occur less with top-down transformations, where the portfolio and program levels should be ahead of the team level in maturity.

There might also be merit in running (parts of) the same survey again but with the suggested adjustments of section 5.1.2. Amongst various analyses there are parts with questionable validity due to survey design decisions. A prime example would be the questions where a negative answer indicates a positive impact, like "lead time per feature" and "amount of unexpected work". As a result of their questionable validity it is unwise to draw any conclusion from these metrics. Even though the issue might not be with the metrics but rather with the way they were presented to respondents.

Because of the time constraints there are some analyses that are possible with the current data-set that were left outside of the (already large) scope of this thesis. The GRM can also produce its own output on the maturity of each respondent. It would be interesting to see how these compare to the maturity as measured by the maturity model. It would also be possible to do a type of regression between a selection of practices and milestones from the maturity model, and some performance metrics.

6 Conclusion

The aims of this thesis were to construct a framework for measuring organizational performance, which can then be used to measure the impact of LSA transformations on organizational performance. These results can then add to the reliability of previously obtained results and also extend the understanding of the impact, especially on the portfolio level. Together with the maturity model by Laanti [2017] this performance framework could be used to test the relation between transformation maturity and organizational performance. The combination of the maturity model and the performance framework could then be used to test the correlation between these variables through Spearman coefficients. To increase insight on the relation between transformation maturity and organizational performance, the maturity model itself is also tested through the means of a graded response model (GRM). With the data obtained through the GRM, data driven feedback can be given on the maturity model itself.

Through literature research and practitioner feedback, an initial performance framework is constructed. This framework is then used to measure impact on organizational performance, based on these results further improvements are made to the framework. This is done with Pearson correlations between metric pairs, combined with the amount of "Don't Know" answers for metrics. Pairs that are highly correlated are reduced to just one of its members to remove redundancy. Metrics that have a lot of "Don't Know" answers were removed to increase overall recognizability. Furthermore, some guidelines on how to translate this framework into a survey are also given, also based on obtained results. These guidelines provide information on the choice between scales, how radiobutton, slider and visual analog scales are all valid options but that each choice comes with its own caveats. This framework and guidelines provide a useful tool for practitioners and academics that want to measure the impact of an LSA transformation on organizational performance. The results obtained through the portfolio metrics of the framework show are more questionable. Overall the impact seems on the portfolio metrics smaller, no correlation between them and any maturity level was found either. As such more work on these portfolio dimensions is still required, likely with an entire new set of portfolio metrics.

When it comes to the impact of LSA transformations, the overall results of this thesis suggest that there is a positive impact across the performance dimensions collected from literature. There are some limitations such as the small number of respondents, as well as the questionable validity of the metrics where a negative answer would indicate a positive impact. It is not unlikely that these limitations impacted the obtained results: Stettina et al. [2021] were able to find more significant correlations, and there were no correlations found with the aforementioned metrics with questionable validity. At the same time, a pattern can be seen between the in thesis obtained results and previously obtained results. Collaboration and transparency are consistently amongst the most impacted metrics, how planned and organized work is amongst the least impacted. Additionally, when adjusting for scale the mean impact across metrics resembles those found in Laanti et al. [2011] and Stettina et al. [2021]. Various correlations between maturity levels and team and product metrics that are significant at the 0.01 level are found. Interestingly enough none of these include the team maturity level, despite some of these metrics affecting the team as well. Looking at the scatter-plot of such metrics, such as effectiveness of development and collaboration, reveals that there are some metrics that experience near immediate impact. The patterns that are found in the impact means add to the reliability of previously obtained results, while the findings from the correlation matrices increase insight on how LSA transformations impact performance. Future research could make a distinction between top-down and bottom-up transformations to see if this impacts the correlation between impact metrics and maturity.

To test the validity of the maturity model by Laanti [2017] a GRM was used. The results of this GRM are then compared against the maturity model and several discrepancies are found. For the portfolio and team level, minor discrepancies are found between the GRM results and the original model. Using the results of the GRM to further improve these levels only requires moving one or two practices to another maturity stage. The results of the program level contained much more discrepancies. Furthermore, the practices of the program level all score very similar in maturity, according to the results. This is an issue if one were to try to divide the practices into maturity stages (as they are presented in the original model) based on the obtained results. For the program level the removal of some practices seems like a better solution. Practices that score similarly in maturity and also have a low "Dscrmn" value are likely candidates for removal. To further test the model a suggestion for future research is done, where participants of a focus group are asked to categorize the practices and milestones that constitute the maturity model. The results of this focus group can then be compared against the model and the results obtained in this thesis. Suggestions for changes to the maturity model based on the results are also given.

In conclusion, despite the limitations of a small number of respondents and some metrics suffering from survey design choices, this thesis is able to contribute to the understanding of the impact of LSA transformations in various ways. Contributions are made by creating an organizational performance framework which adds to the reliability of previously obtained results, the correlation matrices reveal new insights on the workings of the relation between LSA transformation and organizational performance and the GRM results show areas of improvements of an established transformation maturity model.

7 Bibliography

Digital.ai Software Inc. 15th state of agile report, Jul 2021. URL https://explore.digital.ai/state-of-agile/15th-state-of-agile-report.

Tore Dyba and Torgeir Dingsoyr. What do we know about agile software development? *IEEE software*, 26(5):6–9, 2009.

Business Agility Institute. The business agility report: Responding to disruption, 3rd edition 2020, Sep 2020. URL https://businessagility.institute/learn/2020-business-agility-report-responding-to-disruption/487.

Christoph Johann Stettina, Victor van Els, Job Croonenberg, and Joost Visser. The impact of agile transformations on organizational performance: A survey of teams, programs and portfolios. In *International Conference on Agile Software Development*, pages 86–102. Springer, Cham, 2021.

Kim Dikert, Maria Paasivaara, and Casper Lassenius. Challenges and success factors for large-scale agile transformations: A systematic literature review. *Journal of Systems and Software*, 119:87–108, 2016.

Martin Kalenda, Petr Hyna, and Bruno Rossi. Scaling agile in large organizations: Practices, challenges, and success factors. *Journal of Software: Evolution and Process*, 30(10):e1954, 2018.

Maria Paasivaara, Benjamin Behm, Casper Lassenius, and Minna Hallikainen. Large-scale agile transformation at ericsson: a case study. *Empirical Software Engineering*, 23(5):2550–2596, 2018.

Anita Friis Sommer. Agile transformation at lego group: Implementing agile methods in multiple departments changed not only processes but also employees' behavior and mindset. *Research-Technology Management*, 62(5):20–29, 2019.

Kai Petersen and Claes Wohlin. The effect of moving from a plan-driven to an incremental software development approach with agile practices. *Empirical Software Engineering*, 15(6):654–693, 2010.

Maarit Laanti, Outi Salo, and Pekka Abrahamsson. Agile methods rapidly replacing traditional methods at nokia: A survey of opinions on agile transformation. *Information and Software Technology*, 53(3):276–290, 2011.

Abheeshta Putta, Maria Paasivaara, and Casper Lassenius. Benefits and challenges of adopting the scaled agile framework (safe): preliminary results from a multivocal literature review. In *International Conference on Product-Focused Software Process Improvement*, pages 334–351. Springer, 2018.

Maarit Laanti and Petri Kettunen. Safe adoptions in finland: a survey research. In *International Conference on Agile Software Development*, pages 81–87. Springer, Cham, 2019.

Roger Sweetman and Kieran Conboy. Portfolios of agile projects: A complex adaptive systems' agent perspective. *Project Management Journal*, 49(6):18–38, 2018.

Maarit Laanti. Agile transformation model for large software development organizations. In *Proceedings of the XP2017 Scientific Workshops*, pages 1–5, 2017.

Don Wells. Extreme programming: A gentle introduction, Oct 2013. URL http://www.extremeprogramming.org/.

Ken Schwaber and Jeff Sutherland. The scrum guide, Nov 2020. URL https://www.scrum.org/resources/scrum-guide/.

Sallyann Freudenberg and Helen Sharp. The top 10 burning research questions from practitioners. *Ieee Software*, 27(5): 8–9, 2010.

VersionOne Inc. 10th state of agile report, 2016. URL https://explore.digital.ai/state-of-agile/10th-annual-state-of-agile-report.

Inc. Scaled Agile. Safe homepage, Aug 2021. URL https://www.scaledagileframework.com/.

S Ambler. *Disciplined Agile Delivery: A Practitioner's Guide to Agile Software Delivery in the Enterprise*. IBM Press, 2012.

Jeff Sutherland. Inventing and reinventing scrum in five companies. *Cutter IT journal*, 14:5–11, 2001.

Daniel R Greening. Enterprise scrum: Scaling scrum to the executive level. In *2010 43rd Hawaii International Conference on System Sciences*, pages 1–10. IEEE, 2010.

Henrik Kniberg and Anders Ivarsson. Scaling agile@ spotify with tribes, squads, chapters & guilds. *Entry posted November*, 12, 2012.

Jochen Krebs. *Agile portfolio management*. Microsoft Press, 2008.

Craig Larman and Bas Vodde. Scaling agile development. *CrossTalk*, 9:8–12, 2013.

Ken Schwaber. Nexus guide: The definitive guide to scaling scrum with nexus: The rules of the game, Jan 2018. URL https://scrumorg-website-prod.s3.amazonaws.com/drupal/2018-01/2018-Nexus-Guide-English_0.pdf.

Edward D Arnheiter and John Maleyeff. The integration of lean management and six sigma. *The TQM magazine*, 2005.

Inc Cprime. Sage – solutions for agile governance in the enterprise, 2021. URL https://www.cprime.com/sage/.

Mashal Alqudah and Rozilawati Razali. A review of scaling agile methods in large software development. *International Journal on Advanced Science, Engineering and Information Technology*, 6(6):828–837, 2016.

Martin Fowler, Jim Highsmith, et al. The agile manifesto. *Software development*, 9(8):28–35, 2001.

Maarit Laanti. Characteristics and principles of scaled agile. In *International Conference on Agile Software Development*, pages 9–20. Springer, 2014.

Torgeir Dingsøyr and Nils Brede Moe. Towards principles of large-scale agile development. In *International Conference on Agile Software Development*, pages 1–8. Springer, 2014.

Torgeir Dingsøyr, Nils Brede Moe, Tor Erlend Fægri, and Eva Amdahl Seim. Exploring software development at the very large-scale: a revelatory case study and research agenda for agile method adaptation. *Empirical Software Engineering*, 23(1):490–520, 2018.

Pawel Paterek. Agile transformation in project organization-issues, conditions and challenges. 2017.

Ömer Uludağ, Martin Kleehaus, Niklas Dreymann, Christian Kabelin, and Florian Matthes. Investigating the adoption and application of large-scale scrum at a german automobile manufacturer. In *2019 ACM/IEEE 14th International Conference on Global Software Engineering (ICGSE)*, pages 22–29. IEEE, 2019.

Abheeshta Putta, Ömer Uludağ, Maria Paasivaara, and Shun-Long Hong. Benefits and challenges of adopting safe-an empirical survey. In *International Conference on Agile Software Development*, pages 172–187. Springer, Cham, 2021.

Marta Olszewska, Jeanette Heidenberg, Max Weijola, Kirsi Mikkonen, and Ivan Porres. Quantitatively measuring a large-scale agile transformation. *Journal of Systems and Software*, 117:258–273, 2016.

Tomas Gustavsson and Linda Bergkvist. Perceived impacts of using the scaled agile framework for large-scale agile software development. 2019.

Oktay Turetken, Igor Stojanov, and Jos JM Trienekens. Assessing the adoption level of scaled agile development: a maturity model for scaled agile framework. *Journal of Software: Evolution and process*, 29(6):e1796, 2017.

Miia Martinsuo and Päivi Lehtonen. Role of single-project management in achieving portfolio management efficiency. *International journal of project management*, 25(1):56–65, 2007.

Robert G Cooper and Scott J Edgett. Overcoming the crunch in resources for new product development. *Research-Technology Management*, 46(3):48–58, 2003.

Robert G Cooper, Scott J Edgett, and Elko J Kleinschmidt. New problems, new solutions: making portfolio management more effective. *Research-Technology Management*, 43(2):18–33, 2000.

Robert G Cooper and Anita Friis Sommer. New-product portfolio management with agile: challenges and solutions for manufacturers using agile development methods. *Research-Technology Management*, 63(1):29–38, 2020.

Teemu Lappi, Teemu Karvonen, Lucy Ellen Lwakatare, Kirsi Aaltonen, and Pasi Kuvaja. Toward an improved understanding of agile project governance: A systematic literature review. *Project Management Journal*, 49(6): 39–63, 2018.

Fred Niederman, Thomas Lechler, and Yvan Petit. A research agenda for extending agile practices in software development and additional task domains. *Project Management Journal*, 49(6):3–17, 2018.

Paul R Drake, Dong Myung Lee, and Matloub Hussain. The lean and agile purchasing portfolio model. *Supply Chain Management: An International Journal*, 2013.

Robert G Cooper, Scott J Edgett, and Elko J Kleinschmidt. Portfolio management in new product development: Lessons from the leaders—ii. *Research-Technology Management*, 40(6):43–52, 1997.

Robert G Cooper, Scott J Edgett, and Elko J Kleinschmidt. New product portfolio management: practices and performance. *Journal of Product Innovation Management: An International Publication of The Product Development & Management Association*, 16(4):333–351, 1999.

Robert G Cooper, Scott J Edgett, and Elko J Kleinschmidt. Benchmarking best npd practices—i. *Research-Technology Management*, 47(1):31–43, 2004a.

Robert G Cooper, Scott J Edgett, and Elko J Kleinschmidt. Benchmarking best npd practices—ii. *Research-Technology Management*, 47(3):50–59, 2004b.

Sascha Meskendahl. The influence of business strategy on project portfolio management and its success—a conceptual framework. *International Journal of Project Management*, 28(8):807–817, 2010.

Catherine P Killen, Robert A Hunt, and Elko J Kleinschmidt. Project portfolio management for product innovation. *International journal of quality & reliability management*, 2008.

Russell K Thornley. Sustainable strategic alignment of actual project portfolio execution: Application and exploratory case study. In *International Technology Management Conference*, pages 374–381. IEEE, 2012.

Ralf Müller, Miia Martinsuo, and Tomas Blomquist. Project portfolio control and portfolio management performance in different contexts. *Project management journal*, 39(3):28–42, 2008.

Jan Pries-Heje and Malene M Krohn. The safe way to the agile organization. In *Proceedings of the XP2017 scientific workshops*, pages 1–3, 2017.

Johan Linåker, Sardar Muhammad Sulaman, Rafael Maiani de Mello, and Martin Höst. Guidelines for conducting surveys in software engineering. 2015.

Gwanhoo Lee and Weidong Xia. Toward agile: an integrated analysis of quantitative and qualitative field data on software development agility. *MIS quarterly*, 34(1):87–114, 2010.

Jan Recker, Roland Holten, Markus Hummel, and Christoph Rosenkranz. How agile practices impact customer responsiveness and development success: A field study. *Project management journal*, 48(2):99–121, 2017.

Robert Cooper, Scott Edgett, and Elko Kleinschmidt. Portfolio management for new product development: results of an industry practices study. *r&D Management*, 31(4):361–380, 2001.

Tomas Blomquist and Ralf Müller. Practices, roles, and responsibilities of middle managers in program and portfolio management. *Project Management Journal*, 37(1):52–66, 2006.

Linda Wallace, Mark Keil, and Arun Rai. Understanding software project risk: a cluster analysis. *Information & management*, 42(1):115–125, 2004.

Mohammad Abdur Razzak, John Noll, Ita Richardson, Clodagh Nic Canna, and Sarah Beecham. Transition from plan driven to safe®: Periodic team self-assessment. In *International Conference on Product-Focused Software Process Improvement*, pages 573–585. Springer, 2017.

Catherine A Roster, Lorenzo Lucianetti, and Gerald Albaum. Exploring slider vs. categorical response formats in web-based surveys. *Journal of Research Practice*, 11(1):D1–D1, 2015.

Fumiko Samejima. *The general graded response model*. Routledge, 2011.

Norbert Schwarz, Bärbel Knäuper, Hans-J Hippler, Elisabeth Noelle-Neumann, and Leslie Clark. Rating scales numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, 55(4):570–582, 1991.

Roger Tourangeau, Mick P Couper, and Frederick Conrad. Color, labels, and interpretive heuristics for response scales. *Public Opinion Quarterly*, 71(1):91–112, 2007.

Colleen Cook, Fred Heath, Russel L Thompson, and Bruce Thompson. Score reliability in webor internet-based surveys: Unnumbered graphic rating scales versus likert-type scales. *Educational and Psychological Measurement*, 61(4): 697–706, 2001.

Mick P Couper, Roger Tourangeau, Frederick G Conrad, and Eleanor Singer. Evaluating the effectiveness of visual analog scales: A web experiment. *Social Science Computer Review*, 24(2):227–245, 2006.

Leonard J Simms, Kerry Zelazny, Trevor F Williams, and Lee Bernstein. Does the number of response options matter? psychometric perspectives using personality questionnaire data. *Psychological assessment*, 31(4):557, 2019.

Fabian Fagerholm, Marko Ikonen, Petri Kettunen, Jürgen Münch, Virpi Roto, and Pekka Abrahamsson. Performance alignment work: How software developers experience the continuous adaptation of team performance in lean and agile environments. *Information and Software Technology*, 64:132–147, 2015.

Trent D Buskirk. Are sliders too slick for surveys? an experiment comparing slider and radio button scales for smartphone, tablet and computer based surveys. *methods, data, analyses*, 9(2):32, 2015.

Anna DeCastellarnau. A classification of response scale characteristics that affect data quality: a literature review. *Quality & quantity*, 52(4):1523–1559, 2018.

A Survey

Impact of Agile Transformations 2022

Start of Block: Contextual Questions

intro Dear participant,

Thank you for participating in this anonymous survey. This research aims to further understand the performance benefits of large scale agile transformations across various organizational levels. This survey is meant for anybody involved in the large scale transformation of a company: developers, managers, coaches, etc.

By participating in this study you are given the chance to cross-reference the impact of your organization's transformation with others. Additionally you can also gain insight on the maturity of your transformation.

This survey is fully anonymous and all data will be treated confidentially. There are no right or wrong answers, as such we would like to encourage you to answer all the questions. If you have any questions please contact me at t.poot@umail.leidenuniv.nl.

It is estimated that the survey will take around 10-15 minutes to complete.

Thank you very much for your participation!
Best regards,

Tim Poot
Student ICT in Business at Leiden University
t.poot@umail.leidenuniv.nl

Dr. Christoph J. Stettina
Professor at Leiden University

c.j.stettina@liacs.leidenuniv.nl

Page Break

role_t Which role best describes your current position?

- ○ Agile Coach (2)
- ○ Consultant/Trainer (4)
- ○ Development Team Member (6)
- ○ DevOps (8)
- ○ Executive (7)
- ○ Manager (10)
- ○ Product Manager/Product Owner (5)
- ○ Project/Program manager (3)
- ○ Release Train Engineer (9)
- ○ Scrum Master (1)
- ○ Other (11) _____

industry In what industry does your organization operate?

- ○ Agriculture (12)
- ○ Chemical (10)
- ○ Consulting (2)
- ○ Education (9)
- ○ Energy (8)
- ○ Entertainment (5)
- ○ Financial Services (3)
- ○ Healthcare (7)
- ○ Hospitality (6)
- ○ IT (1)
- ○ Manufacturing (4)
- ○ Public Sector (13)
- ○ Transport (11)
- ○ Other: (14) _____

org_size How many employees are working in your organization?

- ○ 201 - 1.000 (1)
- ○ 1.001 - 5.000 (2)
- ○ 5.001 - 20.000 (3)
- ○ 20.001 - 50.000 (4)
- ○ >50.000 (5)

Page Break

scope_% What percentage of employees is within the scope of this transformation?

- ○ 0-25% (1)
- ○ 26-50% (2)
- ○ 51-75% (3)
- ○ 76-100% (4)

---

scope_departments Which departments are included in the scope of this transformation?

- ☐ Finance (4)
- ☐ HRM (6)
- ☐ IT (5)
- ☐ Marketing (3)
- ☐ Production (1)
- ☐ Research and Development (2)
- ☐ Others: (7) _____

---

framework What framework(s) is being used in the transformation?

- [ ] Agile Portfolio Management (APM) (5)
- [ ] Disciplined Agile (DA) (6)
- [ ] Enterprise Scrum (3)
- [ ] Large Scale Scrum (LeSS) (7)
- [ ] Lean Management (9)
- [ ] Nexus (8)
- [ ] Recipes/Solutions for Agile Governance in the Enterprise (RAGE/SAGE) (10)
- [ ] Scaled Agile Framework (SAFe) (1)
- [ ] Scrum Scale/Scrum of Scrums (SoS) (2)
- [ ] Spotify Model (4)
- [ ] Own framework (13)
- [ ] Don't know (11)
- [ ] Other (12) _____

confidence_team How familiar are you with your organization's transformation **on team level?**

- Not familiar at all (2)
- Slightly familiar (3)
- Somewhat familiar (4)
- Fairly familiar (5)
- Completely familiar (6)

---

confidence_product How familiar are you with your organization's transformation **on product level**?

- Not familiar at all (2)
- Slightly familiar (3)
- Somewhat familiar (4)
- Fairly familiar (5)
- Completely familiar (6)

---

confidence_portfolio How familiar are you with your organization's transformation **on portfolio level**?

- Not familiar at all (2)
- Slightly familiar (3)
- Somewhat familiar (4)
- Fairly familiar (5)
- Completely familiar (6)

End of Block: Contextual Questions

Start of Block: Practices - Team

intro_practices
Thank you for completing the first section of this survey. You will now be asked several questions about the adaptation of agile practices and the achievement of agile milestones.

If you are uncertain about a specific practice or achievement, feel free to use the 'don't know' option.

Page Break

practices_team For each agile practice or agile milestone on **team level**, indicate how often your organization uses a practice or accomplishes a milestone.

	Never (1)	Seldomly (2)	Sometimes (3)	Frequently (4)	Always (5)	Don't know (6)
Fast fixes are done as needed (1)	○	○	○	○	○	○
Scrum is in use (2)	○	○	○	○	○	○
Existence of a dedicated build environment (3)	○	○	○	○	○	○
Automatic testing, integration and deployment efforts (4)	○	○	○	○	○	○
Test first approach (5)	○	○	○	○	○	○
Systematically removing impediments (6)	○	○	○	○	○	○
No errors released, production code is practically error-free (7)	○	○	○	○	○	○
Production releases multiple times per day (8)	○	○	○	○	○	○
Version control is in use (9)	○	○	○	○	○	○

End of Block: Practices - Team

Start of Block: Practices - Product

practices_product For each agile practice or agile milestone on **program/ART/cross-team level**, indicate how often your organization uses a practice or accomplishes a milestone.

	Never (1)	Seldomly (2)	Sometimes (3)	Frequently (4)	Always (5)	Don't know (6)
Products/programs are agile (1)	○	○	○	○	○	○
Incremental planning and execution (2)	○	○	○	○	○	○
Ability to embrace change (3)	○	○	○	○	○	○
Agile release trains are in use (4)	○	○	○	○	○	○
Agile roles are in use, are defined and carry responsibility (5)	○	○	○	○	○	○
Incremental demos guide future development (6)	○	○	○	○	○	○
Organized for lean agile way-of-working (7)	○	○	○	○	○	○
Value stream thinking (8)	○	○	○	○	○	○
Agile budgeting and cost follow-up (9)	○	○	○	○	○	○
Networked leadership (10)	○	○	○	○	○	○
Systematically speeding up production releases (11)	○	○	○	○	○	○
Agile metrics are in use (12)	○	○	○	○	○	○

Acceptance test is planned first before a feature (13)	○	○	○	○	○	○
Continuous positive feedback from customers from last deliveries (14)	○	○	○	○	○	○
Ability to create systems and services previously impossible (15)	○	○	○	○	○	○
Ability to respond rapidly to changing customer needs (16)	○	○	○	○	○	○
Organization is networked, empowered, self-controlled and adaptive (17)	○	○	○	○	○	○

End of Block: Practices - Product

Start of Block: Practices - Portfolio

practices_portfolio For each agile practice or agile milestone on **portfolio level**, indicate how often your organization uses a practice or accomplishes a milestone.

	Never (1)	Seldomly (2)	Sometimes (3)	Frequently (4)	Always (5)	Don't know (6)
Portfolio backlog is prioritized (1)	○	○	○	○	○	○
Work is being identified as Epics and Features (2)	○	○	○	○	○	○
Backlog tool support is in use (3)	○	○	○	○	○	○
Portfolio work is continuous (4)	○	○	○	○	○	○
Systematic and fast rolling decision making is being used (5)	○	○	○	○	○	○
Agile metrics are in use (6)	○	○	○	○	○	○
Options thinking in portfolio decision making (7)	○	○	○	○	○	○
Measuring feedback guidance based on data collected and trends (8)	○	○	○	○	○	○
Detecting and utilizing fast business opportunities (9)	○	○	○	○	○	○
Agility is part of the values and company strategy (10)	○	○	○	○	○	○

Ability to innovate new businesses that increase client competitiveness (11)	○	○	○	○	○	○

End of Block: Practices - Portfolio

Start of Block: Practices - Model

model_picture Please look at the following transformation maturity model

model_question Based on the maturity model above, please rank the maturity of your organization's transformation.
The phases of this model build up on each other, that is to say that in order to be at the fluent level you also have to have implemented all the practices at the novice and beginner level.

	Beginner (1)	Novice (2)	Fluent (3)	Advanced (4)	World-class (5)
Portfolio level (1)	○	○	○	○	○
Program level (2)	○	○	○	○	○
Team level (3)	○	○	○	○	○

End of Block: Practices - Model

Start of Block: Metrics - Team and Product

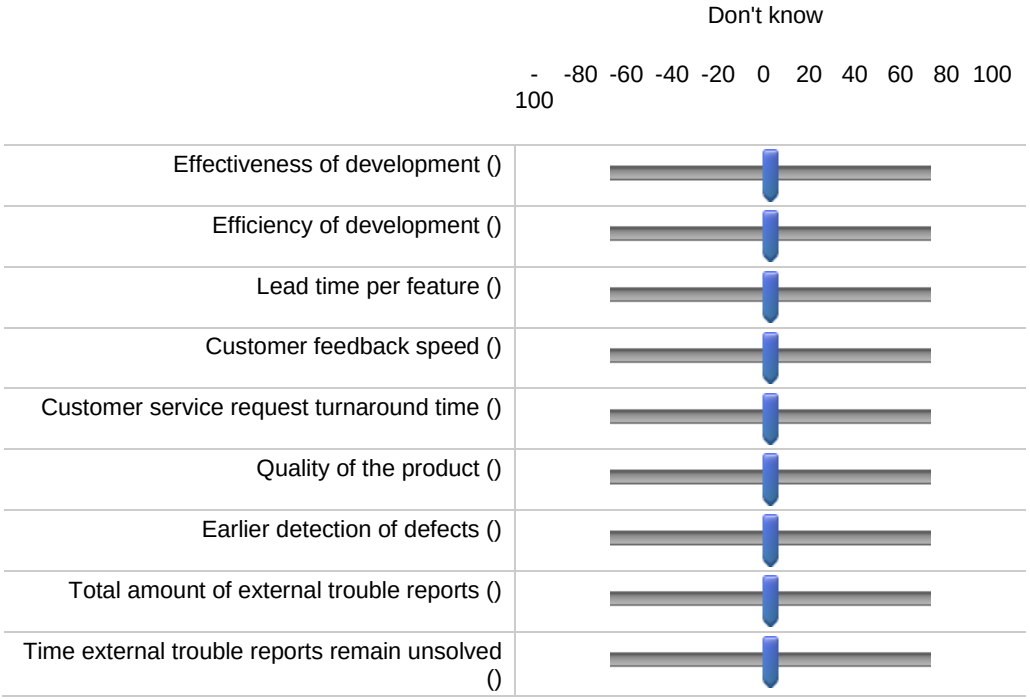
intro_metrics
There are two more sections in this survey. This section is about the various performance benefits of large scale agile transformations.

Per statement, indicate how big a certain aspect of your organization has been impacted by its transformation. If you are unsure about a specific aspect, please use the 'Don't know' option.

Page Break

metrics_tp_1
In your opinion, from -100% to 100%, how has your organization's large scale agile transformation impacted the following aspects?

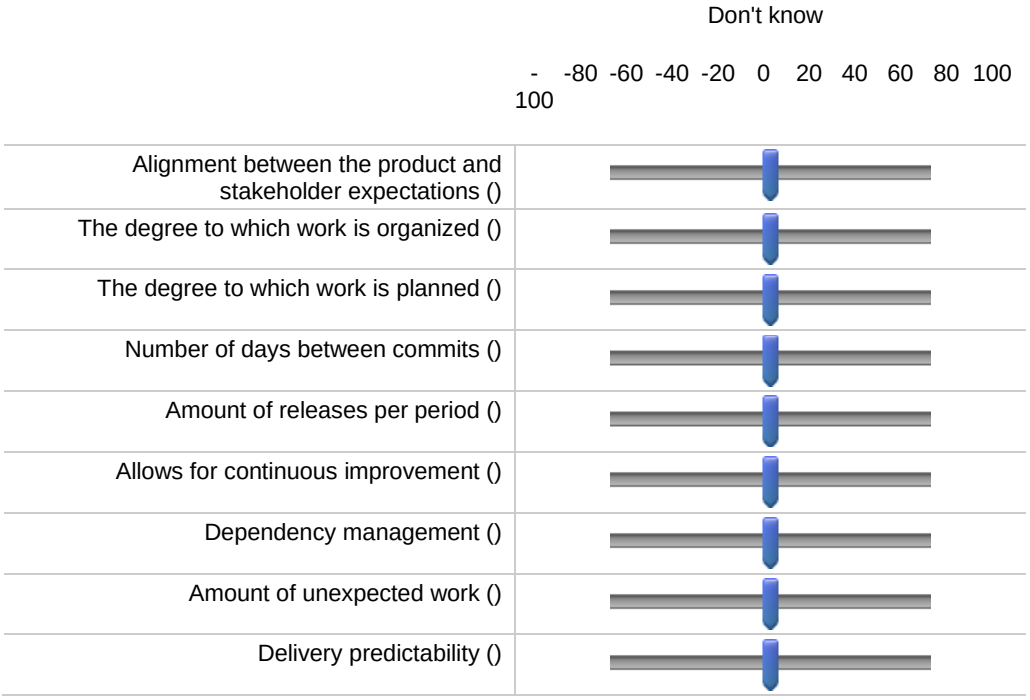
For example: for the question "Effectiveness of development" an answer of -20 would imply a 20% decrease in development effectiveness. An answer of 20 would imply a 20% increase in effectiveness.



Page Break

metrics_tp_2
In your opinion, from -100% to 100%, how has your organization's large scale agile transformation impacted the following aspects?

For example: for the question "Effectiveness of development" an answer of -20 would imply a 20% decrease in development effectiveness. An answer of 20 would imply a 20% increase in effectiveness.



Page Break

metrics_tp_3
In your opinion, from -100% to 100%, how has your organization's large scale agile transformation impacted the following aspects?

For example: for the question "Effectiveness of development" an answer of -20 would imply a 20% decrease in development effectiveness. An answer of 20 would imply a 20% increase in effectiveness.

	Don't know -100 -80 -60 -40 -20 0 20 40 60 80 100
The degree to which work is fun ()	
The degree to which work is hectic ()	
Autonomy of development teams ()	
Collaboration ()	
Transparency of development ()	
Employee turnover ()	
Ability to attract new employees ()	
Overall customer satisfaction ()	

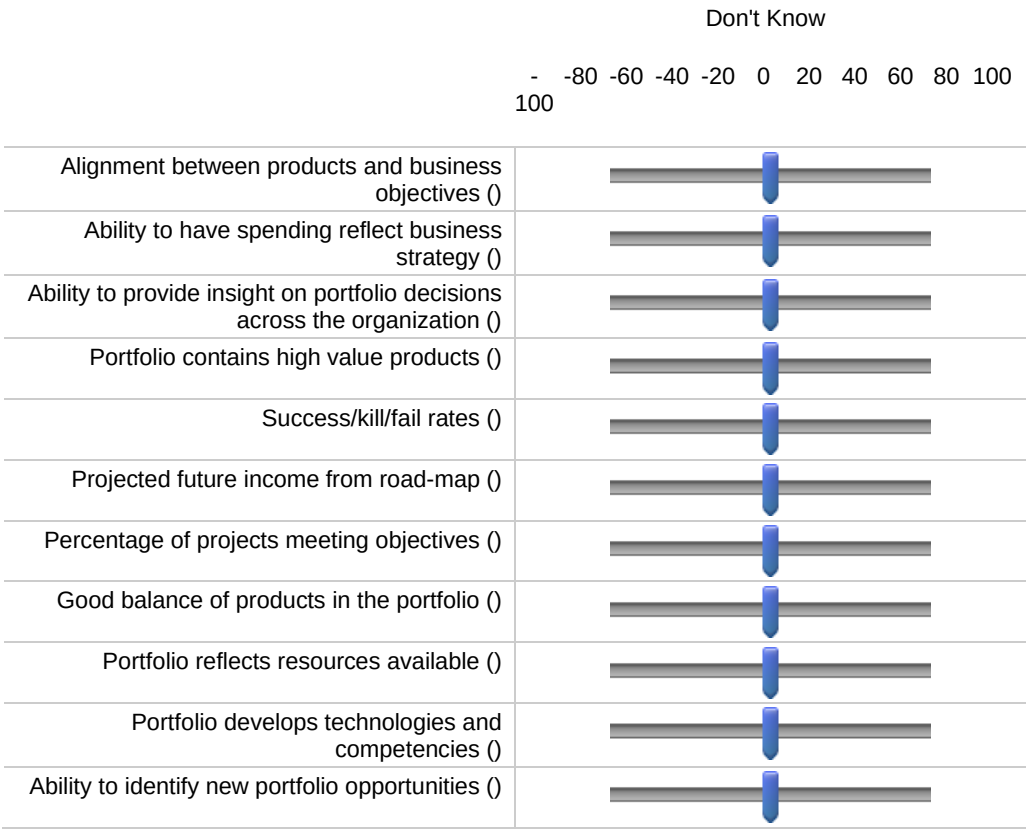
End of Block: Metrics - Team and Product

Start of Block: Metrics - Portfolio

metrics_p
In your opinion, from -100% to 100%, how has your organization's large scale agile transformation impacted the following aspects?

For example: for the question "Alignment between products and business objectives" an answer

of -20 would imply a 20% decrease in alignment. An answer of 20 would imply a 20% increase in alignment.



End of Block: Metrics - Portfolio

Start of Block: Cluster Variables

intro_cluster
You are almost done with the survey, there is one final section. The questions in this final section are about how your organization operates in various aspects.

We would like to remind you that there are no wrong answers and as such we would like to

encourage you to answer every question.

Page Break

measuring_check Do you measure the impact of your agile transformation?

- Never (1)
- Seldomly (2)
- Sometimes (4)
- Frequently (5)
- Always (6)

---

measuring_how Do you steer the course of your transformation based on those measurements?

- Never (5)
- Seldomly (7)
- Sometimes (6)
- Frequently (8)
- Always (9)

---

line_and_matrix Does your organization maintain/have a line organization next to its team structure?

- Yes (1)
- No (2)

---

in/out_source What percentage of teams in your organization is outsourced?

- ○ 0-25% (1)
- ○ 26-50% (2)
- ○ 51-75% (3)
- ○ 76-100% (4)

---

story_size What is your average story size?

- ○ 1 (1)
- ○ 2 (2)
- ○ 3 (3)
- ○ 5 (4)
- ○ 8 (5)
- ○ 13 (6)
- ○ 20 (7)
- ○ 40 (8)
- ○ 100 (9)

---

scrum_master_role What kind of role is the Scrum Master in your organization?

- ○ We don't have a Scrum Master (1)
- ○ Scrum Master is a full-time role (7)
- ○ Scrum Master is a part-time role - fulfilling other tasks next to it (e.g. being an engineer) (9)
- ○ Scrum Master is a part-time role - supporting 2 or more teams (3)

Display This Question:

If What kind of role is the Scrum Master in your organization? = Scrum Master is a part-time role - fulfilling other tasks next to it (e.g. being an engineer)

scrum_master_distrib What is the work distribution like for part-time Scrum Masters?

- ○ 25% Scrum Master 75% other roles (1)
- ○ 50% Scrum Master 50% other roles (2)
- ○ 75% Scrum Master 25% other roles (5)

End of Block: Cluster Variables

Start of Block: End

Email
Thank you for completing our survey and helping us with our research!
If you would like to receive updates about the results of the research you can leave your e-mail address here. We will then get back to you as soon as possible.

_____

End of Block: End