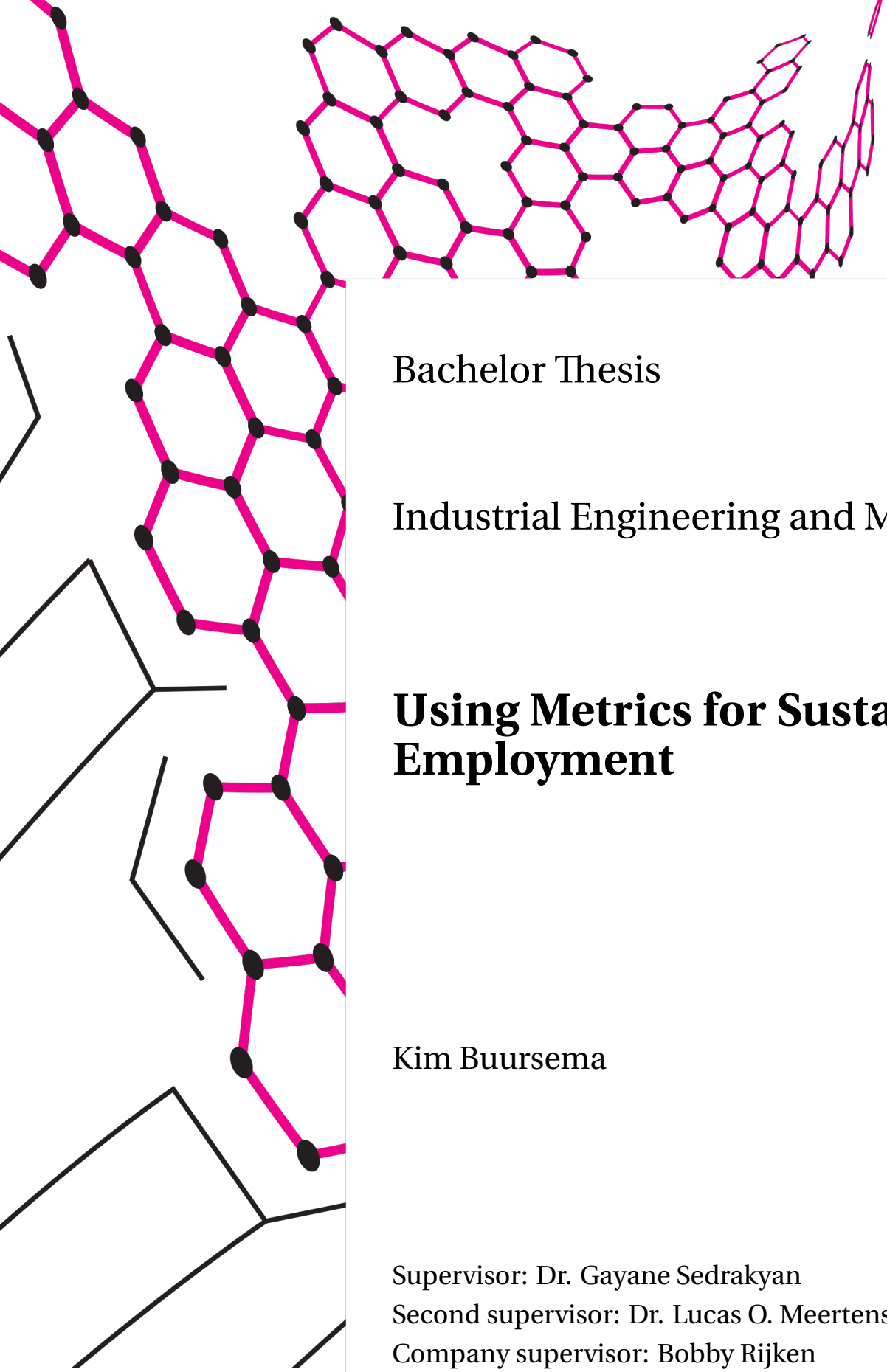

Bachelor Thesis

Industrial Engineering and Management

Using Metrics for Sustainable Employment

Kim Buursema

Supervisor: Dr. Gayane Sedrakyan
Second supervisor: Dr. Lucas O. Meertens
Company supervisor: Bobby Rijken

achmea

UNIVERSITY OF TWENTE.

August 18, 2023

UT/IEM-23.12-23/08/18.5

University of Twente.
Industrial Engineering and Management
PO Box 217
7500 AE, Enschede
Tel. +31(0)534899111

Using Metrics for Sustainable Employment
Bachelor Thesis

Kim Buursema

First supervisor: Dr. Gayane Sedrakyan
Second supervisor: Dr. Lucas O. Meertens
Company supervisor: Bobby Rijken

August 18, 2023
67 pages
6 appendices

This report was written as part of the thesis assignment of the Industrial Engineering and Management educational program.

Preface

In front of you lies my bachelor assignment, for which research is conducted at the department IT DM & BTV of Achmea. For the past six months, I have been working on this project. Working on a project of this scale independently has been an entirely new experience for me, but I have learned a lot and enjoyed the time I spent on it.

First, I would like to thank everyone who helped me at Achmea. All employees welcomed me and offered help where needed. A special thanks goes to Bobby Rijken, who has taken on the task of guiding me through the research. The ambience and help at Achmea made the experience enjoyable and motivated me to continue researching.

Secondly, my UT supervisors have also helped me with their feedback and suggestions. Therefore I would like to thank Gayane Sedrakyan for being my first supervisor and helping me from the start of the research. Furthermore, I would like to thank Lucas Meertens for being my second supervisor and finding the time to give feedback, even though the circumstances could have been better.

Lastly, I would like to thank friends and family who have helped me during the past months by answering small questions and checking parts of my thesis.

Kim Buursema
August 18, 2023

Management Summary

An implementation strategy is proposed to implement a set of metrics, as identified in section 4.1, to keep track of performance and well-being at work. With a foundation of knowledge from literature and analysis of existing problems and requirements, a set of Key Performance Indicators (KPIs) is identified, and two dashboards are designed. With these dashboards, an implementation strategy is written to maximize the effect and utilization of the metrics.

Problem Identification

The action problem of the research is formulated as: "The current employment should stay sustainable in the (near) future." The core problem is: "There is a lack of easy-to-use metrics to keep track of team performance and motivation." This core problem is a combination of three other problems, all resulting in the choice of not using metrics. In order to resolve these problems, the research seeks to answer the following research question: *"What is an effective set of metrics for the department IT Build Debiteuren Management & Betalingsverkeer to keep track of performance and well-being at work?*

KPIs

Two lists of KPIs are found using systematic literature reviews and interviews. Furthermore, these interviews uncovered issues with the used metrics during selection, design and implementation. Additionally, a list of inclusion and ranking requirements for new metrics was made based on these problems and interviews. The list of metrics is split up into strategic, tactical, and operational metrics to align with the varied objectives of department managers and employees. Among the operational and tactical categories, the six KPIs attaining the highest scores are selected, whereas three are selected for the strategic category. One set of well-being KPIs is selected, which is elaborated with additional metrics fitting the research objective. The selection of KPIs has been conducted using the multi-criteria decision-making method and the weighted sum model, using the ranking requirements. After scoring each KPI, validity is verified through discussion with the manager. With these two final selections, two dashboards are designed.

Implementation Strategy

An incremental implementation strategy is set up, including multiple evaluation loops and training, supporting the implementation of the dashboards to ensure long-term usage. Furthermore, management support and the prevention of using metrics to critique employees will maximize this effort. The operational dashboard is recommended to be used during daily stand-ups, whereas the tactical and strategic dashboards are recommended to be used for each planning interval. The well-being dashboard is intended to initiate team discussions on specific subjects during retrospectives. The results of an entire planning interval are discussed by scrum masters and the manager, after which results and feedback should be discussed with the teams.

Conclusion

In conclusion, this research identified three sets of metrics to measure performance and one set to measure well-being. An implementation strategy is proposed to ensure long-term use and to minimize problems, including an incremental implementation plan and recommendations for integrating the metrics into work processes. The selection of KPIs lead to the recommendation of developing three separate dashboards. Furthermore, two prototype dashboards are designed, which include recommendations on the visualization of KPIs to maximize the effectiveness of the dashboards further.

Table of Contents

List of Figures

List of Tables

Glossary

Abbreviation	Unabridged form
ART	Agile Release Train
C-TAM-TPB	Combination TAM and TPB (see TAM and TPB)
DSRM	Design Science Research Methodology
IEM	Industrial Engineering and Management
ISSA	International Social Security Association
IT Build DM & BTV	Department IT Build Debiteuren Management & Betalingsverkeer
KPI	Key performance Indicator
MCDM	Multiple Criteria Decision Making
PI	Planning Interval
QWC	Quality Work Competence
RTE	Release Train Engineer
SAFe	Scaled Agile Framework
SHW	Safety, Health and Well-Being
SLR	Systematic Literature Review
TAM	Technology Acceptance Model
TPB	Theory of Planned Behaviour
US	User Story
UT	University of Twente
UTAUT	Unified Theory of Acceptance and Use of Technology
VSM	Value Stream Mapping
VZ	Vision Zero
WSM	Weighted Sum Model

1| Introduction

This chapter aims to provide the background information necessary for the research and identify and explain the core and action problems. First, the company and department are introduced in section 1.1. Then, in section 1.2, the core and action problems are identified using a problem cluster. Section 1.3 defines the scope of the research, and the chapter finishes with the research design in section 1.4.

1.1 About Achmea

Achmea, founded in 1811, is a prominent insurance company in the Netherlands with subsidiaries such as Zilveren Kruis, Centraal Beheer, and Interpolis. While primarily operating in the Netherlands, Achmea also has a presence in Greece, Turkey, Canada, Australia, and Slovakia. Achmea is a quickly changing company; "To respond effectively to contemporary needs, we are evolving from an insurance company into a financial service provider" [4]. The company aims to assist clients, partners, and stakeholders resolve various health, income, mobility, and sustainability issues.

This research is conducted at the department IT Build DM & BTV (Debtors Management & Transactions). Within DM & BTV, the agile method, specifically SAFe (Scaled Agile Framework), has been used for six years. SAFe has brought increased flexibility and productivity within the organization. SAFe focuses on flexibility and structure within the organization. Using SAFe, the department or, in SAFe terms, Agile Release Train, ART, is split up into several teams, each led by a scrum master and product owner. The scrum masters facilitate the planning and functioning of their respective teams, while the product owners are responsible for the final product and deliverables. Regular meetings with the manager (in SAFe terms: Release Train Engineer, RTE) are held to discuss team and department progress. For a more elaborate explanation of SAFe, see [5].

1.2 Problem Identification

In pursuit of continuous improvement and innovation, Achmea's management acknowledges the need to stay ahead in an ever-changing and demanding environment. Furthermore, the company is committed to having sustainable employment and ensuring the well-being of its employees. Although the department has already implemented SAFe to facilitate continuous improvement, the management believes there is room for further improvement. The management attempted various initiatives, such as the 'Teambarometer' made in 2020. However, completing this dashboard took time and effort, resulting in less use and effectiveness.

1.2.1 Problem Cluster

The department's perceived problems were identified during introductory conversations with scrum masters and the manager. These conversations led to the first issues, but more in-depth knowledge and poking were required to find all problems. These were further explored during unstructured interviews (see 3.2). A problem cluster was made with all the found issues, which can be found in figure 1.1.

The problem cluster reveals several core problems that require attention (most left blocks). It is important to note that two of the identified problems, namely the onboarding program and the perception of metrics as an additional task, are outside the scope of this research. These issues fall under the purview of management, as they involve decisions regarding task allocation and the design of the onboarding program. However, the red-marked blocks must be prioritized and solved before the management addresses these problems.

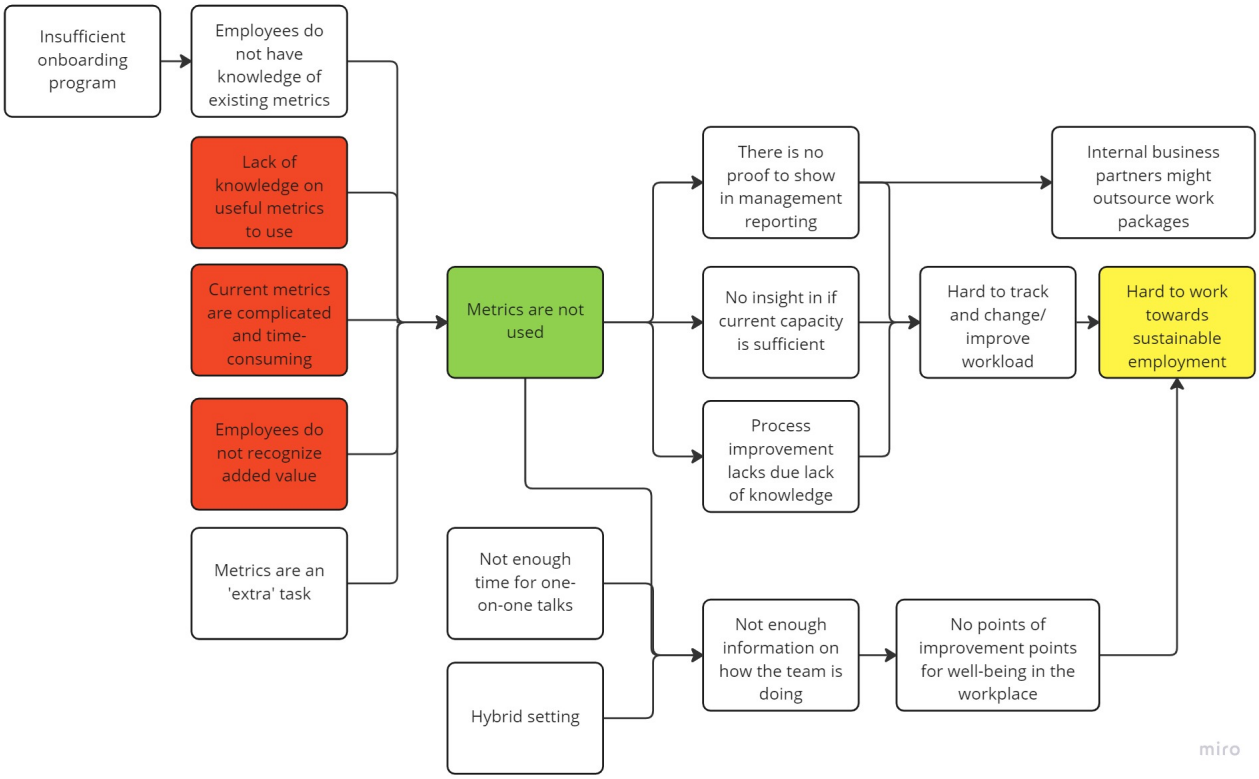

Figure 1.1: Problem cluster

The remaining three problems can be clustered together and addressed simultaneously. These problems, represented by the red blocks, merge into the green block, denoting that metrics are not used. The first problem is the complexity and time-consuming nature of existing metrics. Employees need to gain the necessary knowledge and time to use the metrics as they perceive this is nonexistent. Additionally, some employees do not recognize the added value of these metrics, leading to a decreased willingness to allocate time to these tasks. Furthermore, although the department already has metrics like the aforementioned 'team barometer', employees need to gain knowledge of the relevance of metrics and what information can be retrieved from them. Besides, employees need insight into possible additional useful metrics. Henceforth, the choice is made to desert the metrics.

The fact that metrics are not used has some consequences. The top right part of the problem cluster, following the green block, represents the 'hard/operational' side of the problem; there needs to be more insight into the team's performance. This insight is needed for management, information on capacity, and possible process improvements. The bottom right part represents a 'soft' side to the problem, which is a lack of knowledge of the team's well-being. This knowledge is needed to identify issues at an early stage and to ensure a healthy work environment. Both sides lead to the yellow block; it is hard to work toward sustainable employment. As stated in chapter 1.1, sustainability is a key point of Achmea and this department. The two aspects of sustainable employment (well-being and performance/workload) are currently hard to track. These aspects limit the management in improvement, which is wanted and possibly needed.

1.2.2 Core and Action Problem

Based on the problem cluster, the action problem can be formulated as follows: "The current employment should stay sustainable in the (near) future." While the management believes the employment is sustainable, there needs to be a measure or assurance for this claim. The norm is that the department will keep its employment sustainable, can show proof of sustainable employment, and can keep track of the road toward sustainable employment. The core problem is determined through the elimination of unrelated issues. The problems 'metrics are an extra task', 'onboarding program not sufficient', 'not enough time for

one-on-one talks', and 'hybrid setting' are too time-consuming or do not fit the scope. The first and last are eliminated because these problems require changing the general working structure. However, that is not necessarily desired. The second and third are not chosen because these relate to priorities which are not desired to be changed. Concluding, the core problem is identified as a combination of the left-over problems, which is as follows: "There is a lack of easy-to-use metrics to keep track of team performance and motivation."

1.3 Research Scope

This research focuses on the department IT Build DM & BTV within Achmea and specifically addresses the use of metrics regarding performance and well-being. The scope excludes issues unrelated to metrics, such as the onboarding program. The definition of performance included but was not limited to process performance, throughput time of work packages, capacity, and business value. Other definitions are only included if significance shows from literature or findings in the research. Similarly, the well-being aspects are limited to well-being that directly impacts the work environment. Furthermore, the scope of this research is partially defined by the time limit. There are only ten weeks to perform the research. Due to this limit, choices are made on aspects to include and exclude during research. These choices regard sample size during interviews as well as the design of prototypes.

1.4 Research Design

Based on the core and action problems in section 1.2.2, a research question is made: *"What is an effective set of metrics for the department IT Build DM & BTV to keep track of performance and well-being at work?"* This research aims to develop a supporting tool contributing to sustainable employment within the department. This type of research is categorized as design science [6]. It differs from natural science, as natural science looks at real phenomena, whereas design science focuses on technology helping to accomplish specific goals. Design science encounters challenges; the reliance of the dashboard on the department and its environment and the need for tailored evaluation specific to the research context. Additionally, in design science, the result is only successful if the result is more effective than the former [7].

To provide a clear overview and systematic approach for conducting design science research, Peffers et al. [8] introduced the Design Science Research Methodology. This methodology is a guideline for design science research, incorporating principles and guidelines. That methodology is translated into a table with all steps [2] and is shown in table 1.1. The DSRM serves as a comprehensive framework for guiding the research process in design science. It ensures that the research follows a systematic and iterative approach. In table 1.1, the arrows show the possible iterations in DSRM. Based on DSRM, the research question is divided into five sub-questions. This research design can be found in table 1.2. The design is structured with the DSRM principles and uses several other theories. These theories are explained in chapter 3. The design is explained below; the numbers of each question correspond with the numbers in table 1.2.

1. *What are the problems with the current metrics?*
   This question is derived from the initial activity of DSRM, which includes identifying the current problems and attempted solutions. This question is explanatory, as it aims to understand the current problems and their reasons for existence. The operationalization concepts are chosen as points of interest that have given an insight into the current situation. For the measurement of the concepts, data collection is necessary. Data is collected through interviews with employees and the management, as both have experience and knowledge regarding the current metrics and associated problems. For analysis, grounded theory is chosen. The grounded theory emphasizes the constant comparison of collected data. The constant comparison allows for constant emerging theories [9]. After each interview, the results are compared and analyzed to determine a new research path to maximize useful output. The outcomes are a list of current metrics and occurring problems. These

DSRM activities	Activity description	Knowledge base
Problem identification and motivation	*What is the problem?* Define the research problem and justify the value of a solution.	Understand the problem's relevance and its current solutions and their weaknesses.
Define the objectives of a solution	*How should the problem be solved?* In addition to general objectives such as feasibility and performance, what are the specific criteria that a solution for the problen defined in step one should meet?	Kowledge of what is possible and what is feasible. Knowledge of methods, technologies, and theories that can help with defining the objectives.
Design and development	*Create an artifact that solves the problem.* Create constructs, models, methods, or instantiations in which a research contribution is embedded.	Application of methods, technologies, and theories to create an artifact that solves the problem.
Demonstration	*Demonstrate the use of the artifact.* Prove that the artifact works by solving one or more instances of the problem.	Knowledge of how to use the artifact to solve the problem.
Evaluation	*How well does the artifact work?* Observe and measure how well the artifact supports a solution to the problem by comparing the objectives with observed results.	Knowledge of relevant metrics and evaluation techniques.
Communication	Communicate the problem, its solution, and the utility, novelty,and effectiveness of the solution to researchers and other relevant audiences.	Knowledge of the disciplinary culture.

Table 1.1: DSRM [2]

	Research population	**Research Method**	**Operationalization**	**Data Collection**	**Data Analysis**
1	The department	Explanatory	Current metrics, input data, outputs and opinions of employees	Semi-structured interviews	Grounded theory
2	Literature	Descriptive	KPIs of production	Secondary data collection	SLR
3	Literature	Descriptive	KPIs of well-being	Secondary data collection	SLR
4	The department	Qualitative	Complexity, needed output, available input	Interviews and observation	Multiple criteria decision-making
5	The department	Qualitative	Use intensity, analysis of output	Interviews and observation	User acceptance

Table 1.2: Research Design

outcomes are used further in setting requirements (question four), selecting effective KPIs, and testing the prototype.

2. *Which KPIs are effective to keep track of performance (production-wise)?*
   In order to proceed with the second activity of DSRM, a systematic literature review (SLR) is conducted to identify the possible KPIs as alternative solutions. This question is a knowledge problem for which secondary data collection is used to gather relevant information, of which the answer produced a list of possible KPIs. This list is then used with the requirements to select KPIs.

3. *Which KPIs are effective to keep track of performance (well-being)?*
   This question is similar to sub-question two; therefore, the same reasoning and methodology apply.

4. *What requirements should be met for a metric with the goal of measuring performance and well-being?*
   The next step of DSRM is to identify requirements for the solution. This knowledge, combined with the answers to the first three questions, is necessary for the final question. This qualitative research question aims to gather opinions and data from the department regarding their specific requirements. The focus is on complexity, desired output, available input, and time constraints, which are crucial in determining suitable KPIs. Data is collected through interviews and observations. The interviews provide insights into the requirements of both management and employees, while observations offer additional information on work processes and perceived knowledge of relevant programs. This data is analyzed and generates a list of requirements. The best solutions are selected using the multiple criteria decision-making model (see Chapter 3.2). This model is chosen because the choice for KPIs needs to be made using the list of requirements.

5. *How can the department easily use the metrics daily?*
   This question is also related to the third step of DSRM, which involves designing an implementation strategy. Since other metrics had problems and failed to endure, prevention of these problems should be researched, and the validity of this implementation needs to be ensured. Therefore, the implementation strategy is researched. Interviews and observations are conducted to gather insights from employees and management. The concepts related to user acceptance, as discussed in section 3.4.2, are utilized to validate the results. This validation is part of the fourth step of DSRM, which aims to demonstrate the effectiveness of the design.

Although not explicitly mentioned in the research design above, it is important to note that the third, fifth, and sixth activities of DSRM are part of the overall research process. The third activity involves the development of a prototype dashboard incorporating the selected KPIs, aligning with the research deliverables. Following, sub-question five can be answered. The fifth activity, evaluation, involves testing the design using existing data and conducting meetings with the department to gather feedback and assess its effectiveness. The final activity, communication, is done by writing a report and giving the stakeholders presentations on the design and relevant aspects.

2| Literature Review to identify KPIs

As already stated in section 1.4, this research focusses on performance and well-being. Therefore, the literature reviews are conducted to explore KPIs discussed in the academic literature that regard these topics. By conducting the literature reviews, this chapter answers the second and third questions outlined in section 1.4. The process of finding and selecting relevant articles can be found in appendix A (performance) and C (well-being-). The primary objective is acquiring knowledge about potential KPIs that can effectively assist the management in monitoring performance and well-being. The findings are used in the later stages of the research. Sections 2.1 and 2.2 discuss the results of the second and third questions, respectively.

2.1 Performance

In the first literature review, the second sub-question is answered, which is formulated as follows: *"Which KPIs are effective to keep track of performance (production-wise)?*". Six relevant articles were found and further analysed to gather information on KPIs. Most articles focused on KPIs applicable to departments or teams working with agile methods; the SAFe framework was not mentioned. The articles discuss various KPIs and their application in practice. Some highlight the most useful or frequently mentioned KPIs. Table A.4 provides an overview of the KPIs mentioned in the reviewed articles. The table shows that velocity is mentioned in all six articles. Sprint and release burndown are mentioned in four out of six articles. Important to note is that all these articles do not mention SAFe for specific KPIs. However, as multiple articles mention these KPIs, it is concluded that they effectively measure performance in an agile environment.

Additionally, as seen in table A.1, the search for KPIs related to SAFe yielded no results. The lack of results indicates a gap in the literature concerning KPIs for this specific framework. The statement of [10] supports this assumption by stating that KPIs for SAFe are goal-dependent. For example, a KPI regarding sale statistics would be effective in a sales department that wants to keep track of its sales goals, but this is ineffective for the department of this research. Therefore, no KPIs are selected as general performance KPIs for SAFe. Finally, considering that the KPIs measure different aspects, combining them can provide valuable insights to the department [11]. Particularly, the KPIs mentioned more than three times should be evaluated with the department's requirements, as these are academically accepted as KPIs that accurately measure performance.

2.1.1 Comparison with current metrics

The most evident finding in comparing the metrics in table A.4 and the metrics mentioned by employees (table 4.1) is that not all metrics correspond between both tables. There are metrics mentioned in the literature that are not found in the department, and there are metrics in use by the department that are not mentioned in the literature. The latter either indicates that the department uses metrics that are not effective for their goal or there is a gap in the literature. To ensure that all metrics used in this research fit the objectives, they must adhere to the inclusion criteria as stated in section 3.1. The metrics mentioned by employees but not found in the literature include the "team barometer", a well-being metric discussed in section 2.2. The "planning and realisation" metric is an Excel file in which teams keep track of the planned and realised number of story points. This results in an accuracy score in percentages. This metric does fit the department's goal, as it measures an aspect of performance. However, even though the specific approach used by the department is not mentioned in the literature, it can be translated into other metrics like the velocity deviation. Therefore, this metric does not adhere to the inclusion criteria, and thus this metric is not added to the list. The "working hours registration" metric is solely used for budget tracking.

This goal does not align with the goal of this research, and therefore the metric does not adhere to the inclusion criteria. Hence, this metric is not further included. Lastly, "Value Stream Mapping" (VSM) is defined as a tool utilized to identify weaknesses and strengths within a business process [12]. While this tool is not specifically an agile method, it is associated with lean practices and is mentioned together with lean in literature [13, 14]. Concluding, only VSM is included in the list of KPIs.

2.1.2 Other metrics

During the literature review, specific metrics were identified from SAFe 6.0 [5] not mentioned in the review articles. These are flow lead, flow distribution, flow efficiency and flow predictability. In addition, during the interview process (section 4.1.1), employees suggested other potentially interesting metrics not found in the literature, like user story ping pong. These metrics are listed in table B.1. Possible reasons for not finding these in the literature are due to differences in naming and the search terms used during the literature search. It should be noted that some metrics were found in the literature but with different applications. For instance, requirement coverage was mentioned as a metric to evaluate the quality of glossary extraction [15], instead of a metric to determine the percentage of requirements covered by testing. Secondly, some metrics identified during the interviews serve as a building block for metrics already found in the literature or describe alternative ways of representing similar data. For example, the number of user stories done contains the same information as the progress chart but has an alternative way of presenting this data. A more comprehensive set of performance metrics is developed by incorporating these metrics, which is used in further stages of this research. The final list of metrics can be found in table B.5.

2.1.3 Obtaining input data for KPIs

The performance KPIs all require input; for some, this input is available in the planning tool Azure DevOps. However, not all KPIs can be added to a dashboard in Azure DevOps. Therefore, a data collection method must be researched for the other KPIs to acquire the needed input data. As time and complexity are two problems for the department, this should be limited as much as possible. Therefore, automated data collection has the preference.

As this department works with Azure DevOps, using data from that tool is the least complex and time-consuming way. Azure DevOps stores data regarding user stories and other levels of tasks, to whom those are assigned, and other information regarding these tasks. This information can be extracted using analytics views. This view extracts a data set that can be constructed using filters. These data sets can be connected to PowerBI by importing them via the Azure DevOps server through PowerBI. A refresh button can be clicked when the data sets are added, making PowerBI look for new data using the same connection. Automatic updating without manually clicking a button is not possible. However, data is not collected in Azure DevOps for KPIs like downstream impact and standard violation. This data needs to be generated manually. Automated data collection is not possible. For example, standards need to be agreed upon for the metric 'standard violation', after which these standards need to be checked during every sprint. This check gives a numeric amount, but the standards do not have to be numeric or data-based.

2.2 Well-Being

In this literature review, the third sub-question is addressed: *"Which KPIs are effective to keep track of performance (well-being)?"*. Five articles were found and evaluated for relevant indicators. During analysis, it was determined that some articles do not provide indicators directly applicable to this research. First of all, The study by Wiseman et al. [16] suggests indicators mostly for organisational well-being rather than individual employee well-being, making them less relevant to this research. Additionally, the indicators proposed for individual well-being do not seem relevant to this research. Vayrynen and Kiema-Junes [17] focus on the difference in indicators between white- and blue-collar employees, emphasising safety climate and communication. Although this is a common focus among organisations [1], it does not directly address the focus of this research which is well-being at the workplace. Hence, the indicators mentioned

in Vayrynen and Kiema-Junes are not considered relevant. Adegbite et al. [18] identified three main indicators influencing well-being at work that are not work-related. The indicators mentioned are broad-scoped, entailing community relationships, security of life, and public trust [18]. However, security of life and public trust are broad indicators pertaining to an entire country or region. These indicators may not result in significant differences among employees. For example, all employees living in the Netherlands have similar circumstances resulting in their sense of security. A chance in these circumstances can change the sense of security for all employees, therefore not resulting in significant differences. The indicator related to community relationship focuses on the employee's relationship with their neighbourhood. This relationship can both positively and negatively affect an employee during work. However, the workplace cannot affect this indicator and therefore is not an effective measure for improving well-being at the workplace. Hence, this indicator is not added either.

The remaining two articles provided relevant indicators for further evaluation and selection. The article by Arnetz [3] uses indicators based on the quality work competence (QWC) questionnaire [19]. These indicators specifically address well-being at work for individual employees. An example of these indicators is mental energy, for which employees need to give a score based on several feelings. Hence, QWC is deemed relevant and included for further evaluation. The final article selects indicators to support Vision Zero (VZ) [1], which aims to ensure safety, health, and well-being (SHW). These indicators assist organisations in assessing their status regarding SHW and identifying areas for improvement to prevent SHW incidents. The indicators do not focus solely on well-being in terms of mental energy or related topics. However, they do encompass SHW, which focuses on well-being during analysis. An example of the VZ indicators is 'visible leadership commitment', which identifies if employees see that leaders (managers or other leading functions) commit to improving SHW [1]. Therefore, all VZ indicators are relevant for well-being at work. The list of indicators from VZ and the QWC can be found in C.6. Lastly, the department mentioned the "teambarometer". This metric is a dashboard designed by the department, displaying the well-being status of teams and the department. The dashboard did not yield the desired results, as the answers filled in by employees were what employees thought the desired or most easy answer was. Therefore, this metric is not included in the list of potential indicators.

In the studies [1, 3], the selection of indicators is driven by a specific goal of establishing a solution for measuring well-being at work. While some indicators can be considered individually, there arises a risk of overlooking important aspects of well-being. Therefore, when selecting indicators, it is essential to remember that a set of indicators collectively capture various aspects of well-being at work. Indicators should only be eliminated if unnecessary for the goal of this research and department. Additionally, a set of indicators could be expanded by incorporating additional relevant indicators.

3| Research Methodology

This section elaborates on several theories mentioned in section 1.4. These theories help to answer the sub-questions. Furthermore, they help to secure construct validity. Construct validity means that for results to be valid, they should adhere to some existing theories/models [20]. First, questions two and three are literature reviews and therefore are already based on theories. In sub-question four, this is done with the MCDM. Then, user acceptance is explained, which is used in sub-question five.

3.1 Identifying KPIs with SLR and interviews

To gain knowledge on KPIs regarding this research's goal and identify suitable KPIs, two SLRs are conducted. An SLR can be used when all scientific knowledge regarding a topic is needed, independent of any bias [21]. An SLR first identifies the study's objective and establishes the search terms, after which criteria are determined for in- and exclusion. Then, the sources which are searched are established. A search is conducted with these criteria and terms. This process is executed in appendices A and C.

For this research, the results of the SLR are expanded with results from interviews. The structure and types of these interviews are explained in section 3.2. The KPIs mentioned during interviews can vary from the subject of this research as strict search terms as used in an SLR cannot be applied to the subject. Therefore, inclusion criteria are set up to ensure that all interview results apply to this research. These criteria are listed below.

- The KPI contributes to the measurement of the performance or well-being of the department. This criterion is needed to ensure that all KPIs fit the research goal and limit the search's scope.
- The KPI should measure a new aspect that is not measured by any KPIs identified in the SLR, or it should have a significantly different measurement method for this same aspect. This criterion eliminates similar KPIs, resulting in a comprehensive and divergent list.

3.2 Interview types

In several sub-questions, interviews are selected as the data collection method. Interviews are a mostly qualitative data collection method that can be used when the data can not be collected by the use of other methods or results rely on interpretation and thus questions need clarification [22]. In this research, interviews are chosen because some data is needed, which was not expected to be uncovered in questionnaires or other collection methods. For example, during the initial round of interviews, problems needed to be identified. Superficial problems can be discovered, but in-depth interviews were needed to uncover all underlying issues, which can not be done using other methods.

Interviews have several techniques or approaches that can be used. The choice of interview type depends on the specific research needs and the stage of the research process. Three main structure divisions in interviewing are the structured interview, semi-structured interview, and unstructured interview [23]. Structured interviews are standardised so that each interviewee gets the same questions. Furthermore, there is minimal researcher involvement in discussions [24]. The generated outcomes are standardised because of this structure, making it suitable for research where this is desired. This structured line of questioning is transformed into questionnaires by some researchers [24]. Semi-structured interviews are used for an in-depth understanding of certain phenomena in the world [25]. There is a general guideline with questions, but probing is used to get below the surface. This type of interviewing is suitable when in-depth knowledge is needed on certain topics. Lastly, unstructured interviews resemble conversational exchanges without predefined questions; the researcher does keep the topic in mind [26]. There are no guidelines or questions prepared. This type is most suitable during the initial stages of research to gain

deeper insight into the topic and possible aspects.

3.2.1 Type of interview per sub-question

Choosing the appropriate type of interview is essential to utilise the interviews fully. For several stages of the research, interviews are used. The first sub-question of the research involves interviews to identify the metrics currently in use and problems with those metrics. The main problems resulting in metrics not being used are identified before this stage; in-depth knowledge of these problems and possible further problems is needed. Therefore, the semi-structured interview is most suitable. Likewise, in the fourth sub-question, aiming to gather requirements for metrics, in-depth knowledge and probing is needed to identify requirements and the reasoning behind those, and therefore semi-structured interviews are chosen. In the fifth sub-question, the data collected entails numerical values regarding user acceptance (section 3.4.2). As mentioned above, interviews are to be used when other methods can not collect the required data. Furthermore, structured interviews can also be conducted in the form of questionnaires, which is considered a different method according to Alshenqeeti [22]. Therefore, questionnaires are used as the data collection method for this sub-question. This approach enables the measurement of user acceptance aspects.

3.3 Multi-Criteria Decision Making method

As explained in section 1.4, with a list of requirements, a decision is made on the KPIs. The Multi-Criteria Decision-Making method is suitable when decision-making is complicated by multiple criteria or requirements that need to be considered simultaneously [27]. This research aims to select KPIs for two goals, measuring performance and well-being at work, that are easy to use on a daily basis. This aim comes with multiple requirements, like complexity and time intensity. Therefore, this method is used for the KPI selection. The MCDM process begins by defining the goal of the decision, which in this research is identifying effective KPIs for measuring performance and well-being. Subsequently, a comprehensive list of requirements is generated, as outlined in question four of the research design (section 1.4). These requirements serve as the criteria for evaluating the options. Next, a list of options is created; in this research, a list of KPIs follows from the literature research. A suitable method is selected to determine the weight or rank of the requirements.

The chosen method is the weighted sum model (WSM), a basic method for one-dimensional problems [28]. Furthermore, this method can incorporate the ranking requirements of section 4.1.2 so that it fits the goal and preferences of the department. WSM gives weight to the requirements. The weights are determined based on the goals and needs of the department. This weight is determined by evaluating if the requirements contribute to selecting KPIs fitting this goal. Next, a rubric is made for scoring each KPI per requirement. This rubric is made by assessing for each requirement what the ideal situation is, in which case the KPI scores a five. When the KPI is satisfactory but not ideal, it will score a three to four; in case of insufficient performance on this requirement, the KPI will score a one or two. After scoring all requirements per KPI, formula 3.1 is used to determine the final score per KPI. In this formula, $s_{c(n)}$ is the score of the KPI on criterion n and $w(n)$ is the weight of that criterion.

$$Score = s_{c(1)} * w(1) + s_{c(2)} * w(2) + ... + s_{c(N)} * w(N) \text{ , where } 1 \leq s \leq 5 \tag{3.1}$$

Finally, all KPIs are evaluated based on their score. This review results in a ranking or selection of the KPIs, with which the final decision is made.

3.4 Validation methods

During this research, multiple results are checked to ensure validity. Multiple methods are used based on the research method and data collection methods. First of all, the results from the interviews of sections

4.1.1 and 4.1.2 are validated based on methods mentioned in Dasom et al. [29]. Furthermore, the design and implementation plan are validated using a theory to measure user acceptance. In the sections below, these methods are explained.

3.4.1 Interview validation methods

Interviews can be validated using multiple methods. The first method used is checking the transferability of the results, an aspect of external validity, by discussing the results with the manager. This discussion is executed to check if the manager experienced the same results [29]. Transferability corresponds to the term equivalence [23]. Furthermore, the transferability of the research is increased by explaining the research process and the questions asked during the interviews. Other researchers can also use this explanation to check the methods and access the research process. With this addition, dependability can be checked. These measures together are used to validate the interview results or to provide an explanation for use in future validation.

3.4.2 Validation with user acceptance

The final sub-question aims to advise on the implementation phase of the metrics. As depicted in the problem cluster (figure 1.1), the combined core problem indicates that metrics are currently not being utilised. This lack occurs at the individual level; employees do not fill in and update metrics, and at the management level, the metrics are perceived as complicated and challenging to interpret. Consequently, these problems result in low user acceptance among employees and management. A commonly used term for measuring if people are willing to use these metrics is user acceptance. User acceptance is defined as the willingness and intention of individuals or groups within the department to utilise the metrics as intended [30].

There are multiple theories and models to measure user acceptance. For this research, the model must measure acceptance of the new dashboard on a department level so it can be implemented correctly. Therefore, theories regarding a broader perspective (e.g. innovation diffusion theory) can not be used. Additionally, this part of the research covers employees actively choosing not to use the existing metrics and to find out how likely the new metrics will be accepted. Therefore, the psychological aspect of user acceptance is important [30]. Several theories, like the Theory of Planned Behaviour (TPB) [31], the Technology Acceptance Model (TAM) [32] and the Unified Theory of Acceptance and Use of Technology (UTAUT) [33], are suited for this type of research.

There is a degree of overlap in the variables employed in various theories. UTAUT combines certain aspects from other theories to unify other theories used to measure user acceptance. This unification results in a theory that includes organisational context. Although organisational context can influence the acceptance of technology, in this research, acceptance within one department of a large organisation is measured. Furthermore, UTAUT uses demographic characteristics and time aspects, causing it to be a model suitable for analysing technology utilisation over longer periods [33]. This aspect makes the theory applicable to technology meant for long-term use by individuals. Although in this research, the technology is aimed at use by teams during a long period, team member turnover ensures that this factor is eliminated. Furthermore, one of the variables, 'Facilitating Conditions', regards the user's perception of whether the company has the appropriate technology and support to ensure successful implementation. However, metrics have already been used within the department, and other departments have successfully implemented similar metrics. Consequently, the necessary technology is assumed to be in place, rendering the "Facilitating Conditions" variable less relevant for this particular context.

TAM is a model designed for measuring technology, specifically in the context of information systems. It is meant to determine the use of the technology during work [33]. This goal overlaps the goal of this research, which is finding a set of metrics that the department will use. TPB has a focus on individual behaviour instead of acceptance of technology. It is aimed to measure the intention to perform a certain

behaviour. Whereas TAM measures the acceptance of a specific technology, TPB aims to measure if a certain wanted behaviour is conducted. This research aims to determine if employees will use the metrics and dashboards presented. Taylor and Todd present a combination of TAM and TPB [34], including all constructs from both theories. In the present study, this combination, which is called C-TAM-TPB, was selected. This selection is based on the goal of this research, which is to find a solution that will be used in the long term by employees. To validate the solution, C-TAM-TPB measures aspects which predict the likelihood of employees using the technology, therefore suiting this goal. By addressing four constructs, the core problems as identified in section 1.2.2 are evaluated. This identification leads to a conclusion on whether these problems are expected to be resolved with the proposed solution.

Combination TAM and TPB (C-TAM-TPB)

The problems that need to be addressed with the model are the complexity of the metrics, time consumption, recognised value, and required knowledge for the metrics, which together form the core problem. These are aspects that are part of the theories mentioned above. However, their explanation and use in the models make C-TAM-TPB the most applicable theory to this research.

The TAM model measures acceptance by two variables; perceived usefulness and perceived ease of use. In a later model (TAM2), the subjective norm is also added [33]. Perceived usefulness is defined as the degree to which the user perceives the technology would enhance their performance. This construct aligns with the problem that employees do not recognise the added value of metrics. By measuring the perceived usefulness of the proposed dashboards, it can be checked if employees do recognise the value of the new dashboards, which would result in more use of the dashboards [35]. Therefore, this is an appropriate measure for this research. Perceived ease of use reflects user's belief that the technology is user-friendly and requires minimal effort, thus addressing the concerns related to time consumption and metric complexity. This construct measures a dashboard's subjective complexity and time consumption. The assumption employees make regarding these aspects is an important factor in their choice to use the dashboard, as the problems show that employees chose to desert metrics in the past because of perceived complexity and time consumption (section 1.2.1). The last measure, the subjective norm, examines the influence of significant others or other important people on the user's perception of the importance of using the new technology. This measure can provide insights into the impact of management pressure on technology utilisation, possibly resulting in the choice (not) to use technology. In the context of this research, the subjective norm can measure the impact of both department management and other employees on the choice to use technology.This construct does not align with one of the problems mentioned in section 1.2.1. However, the influence of others can be an underlying issue in the choice of not using a technology [36].

TPB also has the variable subjective norm, along with attitude towards behaviour and perceived behavioural control. Attitude towards behaviour refers to the user's feelings regarding the desired behaviour. This construct complements perceived usefulness by focusing on the user's feelings regarding using metrics rather than only assessing their perception of the metric value. By assessing the attitude towards behaviour, it is checked if employees are willing to use dashboards, regardless of the quality of the dashboard. Therefore, if this is positive, employees are likelier to use any dashboard [37]. This construct is an appropriate addition to this research as it can show that the use of dashboards, in general, is accepted. Secondly, perceived behavioural control refers to what the user sees as a constraint for using the metrics. These constraints can be both internal and external [31]. Hence, this variable can regard the time constraint; employees' perception of their ability to make time for the use of the complexity; the perception of the ability to learn the needed knowledge.

Table 3.1 shows an overview of the used constructs. In this research, the constructs are measured using an online questionnaire. In this questionnaire, nine questions are posed for the performance dashboards and the same nine for the well-being tool. The questions are derived from literature [38]. Per variable, at least two questions are posed with a 5-point Likert scale (1 = strongly agree to 5 = strongly disagree).

Construct	Explanation
Perceived usefulness	The degree to which a user perceives the technology will enhance their performance
Perceived ease of use	The degree to which a user believes the technology is user-friendly and requires minimal effort
Subjective norm	The influence of other important people on the user's perception of the importance of using the technology
Attitude towards behaviour	The user's feeling towards the desired behaviour
Perceived behavioural control	The degree to which users see constraints for using the technology

Table 3.1: C-TAM-TPB model

Furthermore, two questions are added for further analysis. These are regarding the team and function. The questions can be found in appendix D.

4| KPI Selection, Design and Implementation

In this chapter, first, in section 4.1 KPIs are selected based on current problems and requirements. Next, the design and implementation of these KPIs are presented in section 4.2.

4.1 KPI selection

In this section 4.1.1, the first sub-question from section 1.4 is answered: *"What are the problems with the current metrics?"*. Next, in subsection 4.1.2, the fourth sub-question is answered: *"What requirements should be met for a metric with the goal of measuring performance and well-being?"*. Following this, subsection 4.1.3 explains a dashboard framework based on an idea generated during an interview. Then, subsection 4.1.4 explains the Weighted Sum Model (WSM). The KPIs identified in chapter 2 are selected using the requirements outlined in section 4.1.2 and the dashboard framework in subsections 4.1.5 (performance) and 4.1.6 (well-being).

4.1.1 Current metrics and problems

To answer the second sub-question, semi-structured interviews (section 3.2 were conducted to gain insight into the department's existing and previously used metrics. The analysis followed the grounded theory approach, which involves constant comparison and analysis of the collected data [9]. Grounded theory is based on an iteration process during the research. An initial interview is held, after which results are analysed. Based on this analysis, a new interview is conducted with possibly a changed and improved structure. The use of this theory results in allowing theories to emerge and to improve the research path and final results constantly. Iterations of this theory are stopped when new results are not found anymore. This theory allowed for a change in the structure of the interviews where needed to explore emerging theories and ideas in depth. All interviews followed a similar format with open-ended questions, beginning with questions regarding the current and past metrics employed by the department. For example, all interviewees were asked what metrics are currently used. Following, interviewees were asked for their opinions and perspectives on those metrics. Some initial problems were identified during these questions; for instance, it became clear that the department has an ongoing discussion about using a metric regarding velocity. Furthermore, using a metric regarding well-being was perceived as mostly negative. Probing was needed to gain further insight into why certain metrics were not considered useful or were not being utilised. Additional questions were added to subsequent interviews based on insights from the first interviews. These include questions about the mentioned problems and the interviewee's previous experiences with metrics in their former jobs or departments. For the well-being metric, this probing led to the discovery that some employees found it can be a useful metric, whereas others said it should not be used at all.

The interviews yielded a list of metrics that the department is familiar with, which is presented in table 4.1. The third column represents the experience employees have with those metrics. The last column indicates the required input for the metrics. One metric that should be highlighted is the teambarometer. This metric was the first well-being metric within the department, used to track how teams are doing. This metric was introduced in 2022. However, within half a year, almost no team used the metric anymore because it was tedious or because new scrum masters did not know that this needed to be used. Other scrum masters mentioned that discussing the teambarometer every two weeks made the retrospectives more tedious, as when they could discuss new topics every meeting. Another metric to highlight is the sprint burndown. All interviewees mentioned this metric, and they found this a useful metric. This metric was mainly used as a conversation starter during the daily stand-up. Despite this positive view, most teams do not use this metric regularly anymore, which is a surprising discovery. Interviewees could not mention specific reasons for this, even though it got discussed less over time.

Metric	Elaboration	Experience	Required Input
Teambarometer	This metric was used to get an insight into the well-being of the team members.	Useless if the results are not discussed with the team, people just filled in straightforward answers to get it done because it is a tedious task, some found it useful to get a quick insight into how the team is doing	Employees fill in a form with questions about their week/sprint
Velocity	This metric determines the amount of work done in a period.	Teams use this differently, and there is a discussion on the right way.	The number of story points planned during a period (sprint, PI)
Sprint Burndown	Number of story points remaining during a sprint	Was a conversation starter, but slowly faded out of daily routine	The number of unfinished story points throughout a sprint
Planning and realisation	At the end of periods, the teams look back at their planning and realisation and compare these results	Is deemed helpful because the new planning can be adapted based on the results of the last period	Amount of story points planned and amount of story points realised
Business value derived	Generally, this metric is used by taking all costs and revenue to look at the value of the user story. Qualitative aspects are valued manually.	This metric is only used by management but is helpful to determine whether a work package is worth the effort put in	Costs, revenue, and qualitative value of a user story
Working hours registration	All hours employees work are registered and plotted against the budget and work packages.	Only used by management, it is an effective check on the budget and if costs weigh up against results	Worked hours, budget and value of work packages
Value Stream Mapping	This metric is meant to determine the value of each step in the department's process is	This metric is unfinished due to being complex. A start has been made in PowerBI	A mapped-out process and perceived value of each step in the process

Table 4.1: Current metrics

Of the experiences in table 4.1 and some follow-up questions during the interviews, general problems are identified. These are problems currently occurring, like the first two regarding employees, or that employees have experienced in the past, like metrics fading out of daily routine. The core problems found in the problem cluster overlap with this list of problems, as lack of knowledge is related to the problem of drawing conclusions and making metrics is complicated, and time-consuming metrics is related to updating and time and discipline, and not recognising added value relates to metrics being useless and lacking motivation.

- Employees
  - find filling forms for well-being metrics tedious
  - find metrics useless if the results are not discussed
- Metrics fade out of daily routine
  - Motivation for using metrics is lacking
  - Updating metrics costs time and discipline
- Making metrics can be complex
- Metrics are not updated frequently
- Concluding metrics is hard

Next to the problems obtained from employees' experiences, there are also some input issues. First of all, teams use the planning tool differently. Therefore, some metrics like the sprint burndown can give a result that differs from reality. Secondly, there is a discussion on how the metric for velocity should be used. The discussion is unremitting; the department can not agree. Lastly, it is important to note that the required input for some metrics can be obtained from Azure DevOps, but some need to be added manually.

These need to be resolved to ensure that the metrics presented in this research do not encounter similar problems. Most problems need to be taken into account during the implementation phase. For example, the differences in ways of using metrics and databases can be solved by writing a process on how they should be handled. However, metrics' complexity and time intensity should be considered during the design phase. These aspects are used in the requirements for selecting metrics. Some problems found are not described in the problem cluster. These are the problems regarding required input. These problems are important during the design phase, as these can deliver time-intensity issues for the department.

During the second round of interviews, it became clear that there are several goals for the metrics, which differ per employee. Employees emphasise metrics they can use to check if they are on schedule and to use during planning, like the sprint burndown, whereas scrum masters prefer metrics that measure performance on a tactical or strategic level, like velocity or even business value derived. Higher-ups generally wanted strategic or tactical metrics, whereas other employees were more eager for operational metrics that address day-to-day issues. One scrum master pitched having metrics devoted to one of the three levels (strategic, tactical, and operational) and separating them using a line or spacing in a dashboard design. This idea is used when selecting metrics and valuing the metrics with the criteria and implementation, as this changes the general use (section 4.1.3).

4.1.2 Requirements for new metrics

The fourth sub-question is about the requirements for a metric. These requirements can be derived from the list of problems. First of all, the complexity of the metric should be taken into account, as employees mentioned that this had been a problem in the past. Next to that, employees mentioned that time-consuming metrics resulted in not using metrics, and therefore a requirement should be about the time intensity of a metric. These requirements still needed to be more specific. Therefore, another round of interviews was conducted to get a clear view of these two requirements and to add other missed requirements. These interviews were structured like the interviews mentioned in section 4.1.1. First, employees were asked to give their requirements for metrics. These questions resulted in the requirement of regular

usage of metrics and the ability to become part of the work process. Then, an elaboration was asked regarding the metrics' time intensity and complexity. This elaboration resulted in a clear definition of those requirements and added one new requirement; the automatic updating of metrics.

These requirements are based on employees' opinions and analysis from previously conducted interviews regarding the problems with metrics. These requirements do not consider the academic reason for including a metric that would be excluded based on these criteria. For example, time-intensive metrics can provide more detail and insights into the department's process and improvement points. The same reasoning can be included for complex metrics; a complex and hard-to-read metric can still provide much information and details regarding the department's performance when employees have the right knowledge to read them. However, the department already has experience with metrics and has given employees responsibility for gaining knowledge regarding metrics and dashboards. In section 1.2, it can be read that these previous attempts to work with dashboards have not succeeded due to the problems identified in section 1.2.2. Therefore, for this research, the choice is made to consider requirements regarding these problems and employees' opinions, resulting in the requirements including time and complexity.

Next, the requirements are split into two categories. These categories are based on employees' definitions and the department's needs; some requirements are mandatory, and others are favourable. First, there are exclusion/inclusion criteria. A metric must comply with these to be included because it is not an effective metric for the department. For example, the required input should be available; otherwise, the department cannot use the metric. Secondly, some requirements give a preference to certain metrics. These requirements help rank metrics and make a final choice. This category, for example, contains the time requirement, as it gives a preference to metrics that take much time to update and analyse. Additionally, the note was made that there should be a limit on the number of metrics; otherwise, there could be a loss of overview on all metrics. Furthermore, multiple articles mention that a single view dashboard is most effective [39]. Therefore, a limit has to be set on the number of metrics compared to the use of the metrics.

Inclusion requirements

- Data should be updated automatically for daily used metrics
- Metrics should be used regularly (frequency is based on the user and the goal)
- Metrics should be able to become part of the work process
- The required input should be available

Ranking Requirements

- Complexity; A metric must be comprehensible and applicable to all organisational personnel. Herefore, it should be simple. The simplicity is twofold: updating and analysing should be doable. Complexity is the measure in which a metric is straightforward and simplistic
- Time; the time and effort put into a metric should weigh against the frequency of utilisation based on its context (a metric used every day cannot take too long to update)
- Perceived necessity; during the interviews, some metrics were named as particularly necessary for the department. Perceived necessity is defined as the necessity for a metric to be in the dashboard, meaning a dashboard cannot be made without this metric if it is perceived necessary [40, 41]
- Relation to goal; the metric should give insight into aspects contributing to the final goal: maintaining sustainable employment

4.1.3 Splitting metrics into categories

From the interviews, it became clear that there are different goals for employees with different functions. This recovery has raised the idea of designing three dashboards with different levels: strategic, tactical, and operational. Operational is related to day-to-day tasks and covers the same period; an example is the sprint burndown. Tactical means the overarching goal of using operations to advance [42]. Tactical decisions usually span a year. Strategic is the theory of becoming the best, which can be reached using tactics including operational attributes [42]. The strategic period is two to five years. The department works SAFe and thus focuses on being agile and flexible. Therefore, the general periods for the three levels are adapted to fit the department. Strategic will regard periods of a year or longer, tactical will be one PI, and operational will stay on a day-to-day basis.

Horkoff et al. [11] translate this definition to metrics (or indicators). Metrics are a tool to give measure to an organization's goal. They can be constituted of multiple other metrics and create levels. Higher-level metrics are linked to strategy, whereas low-level metrics are linked to operations. Strategy metrics evaluate a particular goal, tactical metrics measure processes contributing to the goal, and operational metrics measure aspects of the processes. The three levels are a pyramid framework, as in figure 4.1. This framework for metrics can be used for the department so that the different goals of different functions are satisfied with the designed dashboard. The KPIs are first classified as operational, tactical, or strategic metrics to find a design that fits this research. Within the levels, a choice is made for final metrics, after which a dashboard is designed where the three levels are connected. Within each level, the choice for metrics is made using the MCDM method (section 3.3). The classification of the metrics is based on automated updates, frequency of use, and the department's goal. The list below shows the requirements for each classification.

- Operational
  - Automated updates
  - use at least once per sprint
  - Track progress and performance for at most a sprint
- Tactical
  - Updating is straightforward and takes at most five minutes per moment of use
  - Use at least once per PI
  - Track progress and performance for at most a PI
- Strategic
  - Updating can take time and can be done manually
  - Use at most once per PI
  - Track progress and performance over a period longer than a PI

4.1.4 Ranking method for KPI selection

The raking method for KPI selection is explained in section 3.3. First, the weights of the requirements are determined. These can be found in table 4.2. For complexity, during initial interviews and requirement interviews, it became clear that this is an important aspect of selecting KPIs. Employees find some KPIs too complex to understand and use, resulting in less use. The same reasoning determined the weight for time based on analysis and updating the KPIs. As both are equally important and have the same effect, they have the same weight. The score of perceived necessity was determined by evaluating the department's goals. This requirement does not contribute to the goals but is included to give preference to some requirements based on employees' opinions. Therefore, this has the lowest score. Relation to goal has the highest score, as it contributes to finding KPIs that align with the set goals, helping the selection of KPIs

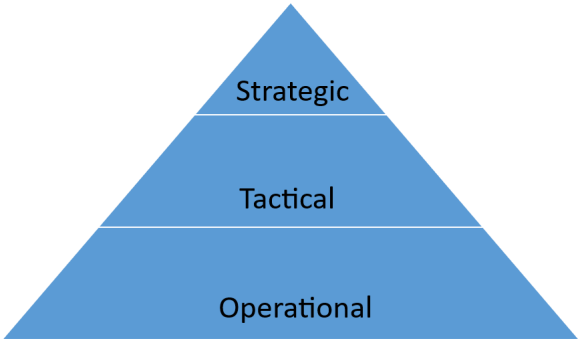

Figure 4.1: Category pyramid

Requirement	Weight
Complexity	2.5
Time	2.5
Perceived necessity	1
Relation to goal	4

Table 4.2: weights requirements

that effectively keep track of performance and well-being. However, it is decided that this requirement is not so important that it can weigh half of the total weight. Furthermore, a requirement should always count for at least 0.1 of the total weight (where the sum equals 1) to ensure it can still influence the outcome. With these reasons in mind, the final weights are determined.

Additionally, a rubric is made, which is used to score each KPI. The rubric can be found in table 4.3. Next, formula 4.1 is used to determine the final score per alternative. The ranking requirements are time (s_{time}), complexity (s_{com}), perceived necessity (s_{pn}) and relation to goal (s_{rtg}).

$$Score = s_{com} * 0.25 + s_{time} * 0.25 + s_{pn} * 0.1 + s_{rtg} * 0.4 \text{ , where } 1 \leq s \leq 5 \tag{4.1}$$

For example, velocity deviation scored a 3 on complexity because it requires some background knowledge, and updating requires a concise explanation. This KPI scored a 4 on time, as updating takes less than five minutes, and analysis can also be done within five minutes. On perceived necessity, this scored 2, as it was not mentioned by employees but was perceived as more necessary than other not mentioned KPIs scoring 1. In relation to the goal, this KPI scored a 5, as velocity deviation is a direct indication of performance over a more extended period, as high deviation indicates insufficient performance and planning. Applying these scores in the equation, the final score of velocity deviation is 4,0.

4.1.5 KPI performance selection

First, with the manager, it was decided that all test-related KPIs were not used for the rest of this research. The reason is that testing is a separate group within the department, and all testers are divided into teams. KPIs regarding testing would indicate the performance of testers but not of the SAFe scrum teams. Therefore, the indicators related to this group and this type of work are out of the scope of this research. These KPIs are indicated in table B.5 with '*'.

Furthermore, the indicators need to apply to the inclusion criteria. For operational metrics, the data should be updated automatically. This requirement is met by all KPIs on the operational level. Furthermore, metrics should be based on the metric level; they should be used regularly and be able to become part of the work processes. For operational metrics, this is daily; tactical metrics are (bi)-weekly up to each PI, and for strategic metrics, it is each PI up to once a year. All KPIs meet this requirement. The last inclusion requirement is about required input. For most KPIs, the required input is available. However, there also are some for which the input currently does not exist. This input can be generated, meaning

Score	1-2	3-4	5
Complexity	The KPI is very hard to understand, even with expert knowledge, updating requires extensive training	The KPI is understandable with for a person with background knowledge, updating can be done after concise explanation	The KPI is understandable by a person without background knowledge, updating can be done without background knowledge
time	Updating the KPI costs more than half an hour, analysis of the KPI takes more than twenty minutes	Updating the KPI takes five minutes at most, analysis of the KPI is done within ten minutes	Updating the KPI is automated and does not take time, analysis of the KPI is done within two minutes
Perceived Necessity	The KPI is not mentioned by employees at all	The KPI is mentioned at least twice as being useful	The KPI is mentioned at least four times as being useful
Relation to goal	The KPI does not measure any aspect of the goals set	The KPI measures an aspect of one of the goals set, but is not a direct measure of the goal	The KPI measures the goals directly

Table 4.3: Rubric ranking requirements

it should be added manually. These KPIs are indicated in table B.5 with '**'. An example of such a KPI is team member turnover. Input for this metric can be generated by tracking the number of employees leaving and starting at the department for periods. However, there is no automated way to keep track of these numbers; therefore, this should be kept track of and updated in the metric manually.

The classification can be found in table B.5. Next, the KPIs are ranked based on the ranking criteria (see section 4.1.4). The complete ranking can be found in table B.5. This ranking is divided per level. For each level, a maximum number of KPIs is chosen. The maximum for operational, tactical, and strategic is six, six, and three. These numbers were chosen because, during the initial stages of this research, employees mentioned the loss of overview due to too many metrics and possibilities. Furthermore, a limited number of metrics is preferred for design and visual purposes, further elaborated upon in section 4.2.1. Lastly, due to strategic metrics scoring low compared to other metrics, it is chosen only to include three metrics to limit the risk that problems arise regarding time and complexity issues. Therefore, the top six or three in the rankings are chosen as the final KPIs. These selected KPIs are listed in table 4.4.

Three of the chosen operational KPIs represent a numerical value indicating the current performance within one glance. These are '#User stories done', '#Stories' and 'Flow load'. First, the number of stories shows the total number of stories in all states for the current sprint. That way, the team knows what their workload is. By determining their average number based on historical data, the team can decide upon values within which this KPI should be to be on track and perform well. The number of stories done shows what part of this total amount is finished. This metric gives a different indication but can be used during reporting to management. The flow load can be used to see how many story points are in progress. This metric is beneficial to keep track of the total workload during each moment in time to prevent that teams from taking on too many user stories simultaneously. This metric can be a driver for maintaining an adequate workload, contributing to sustainable employment. Another driver of this set of metrics is the sprint burndown. This graph shows the remaining work for the current sprint, including a trend line which shows what trend should be followed to finish all work while maintaining a steady workload. Within agile methods, this graph is well-known. The department has used this graph and identified some patterns already. The burndown can be used to identify those patterns, e.g. finishing all story points during the

KPI	Score
Operational	
# User stories done	5.0
# Stories	4.9
Sprint burndown	4.8
Release burnup/down	4.8
Flow load	4.7
Velocity	4.7
Tactical	
User story ping pong	4.3
# User stories pushed	4.1
True sprint length or cycle time	4.0
Cancellation rate	4.0
Team member turnover	4.0
Lead time	3.9
Strategic	
Velocity deviation	4.0
Flow predictability	3.8
Flow efficiency	3.6

Table 4.4: Final performance KPIs selected

last day instead of spreading over the sprint, which can help process improvement within the teams and department.

The tactical and strategic KPIs are essential for longer-term improvements regarding workload and process improvement. Of these KPIs, cycle time and velocity deviation can be the main drivers for these points of interest. Cycle time, or true sprint length, measures the time user stories, on average, are in the phases that started to be completed. Within SAFe, user stories need to be finished within one sprint. When the cycle time is below the total days in a sprint, this indicates that the team's performance is adequate. However, when cycle time is higher, the team cannot deliver user stories within one sprint. This cycle time can indicate that the workload needs to be lowered, and the teams underestimate the workload of user stories or possible other problems. In order to find these problems, conversations need to take place, and finally, to resolve them, actions need to be taken. When this is done, the team's performance will increase, the workload will be tracked better, and thus sustainable performance is one step closer. Velocity deviation is tracked over a more extended period and requires historical data. Velocity deviation measures the deviation from the average velocity of a team over the past years. If the deviation is high, this means that a team does not have a steady workload or that a team's plans need to be revised. In both cases, this can result in employees being unable to manage all the tasks and, therefore, not delivering what was promised to management. To conclude, these indicators can drive this research's final goal and action problem, contributing to sustainable employment within the department.

4.1.6 KPI well-being selection

First, the KPI selection for well-being is mostly influenced by the problem that employees find filling in forms for metrics tedious; repeatedly answering the same questions in a form results in average answers so that employees do not have to spend much time on the form. However, subjective well-being is mostly measured using quantitative methods (e.g. interviews or surveys) [43]. Therefore, the problem arises that the goal and measurement method conflict with the problems found. For both the QWC and VZ indicator, a questionnaire needs to be filled in by employees. However, the VZ indicators can be used with only yes or no questions, taking up less time. Another way of reducing the conflict is by implementing a set of indicators. Instead of filling in an entire survey during each PI or sprint, a part of the survey can be highlighted. The whole set of indicators will come by during a PI, but each period has a different focus point

	Complexity	Time	Perceived necessity	Relation to goal	Total score
QWC	4	3	1	4	3.45
VZ	4	4	1	4	3.7

Table 4.5: Well-being indicators scores

and, thus, different questions. Employees still need to fill in a form, but the number of questions is significantly reduced, and the questions differ every period. With this implementation strategy in mind, a set of indicators that fits the department and still solves or reduces the mentioned problems can be chosen. In section 4.2.3, an explanation can be found.

In section 2.2, it is noted that the indicators are a set which should be considered during the selection procedure. Therefore, first, a choice is made for one of the sets of KPIs by using WSM (section 4.1.4) for the complete set. The sets are scored based on the WSM as explained in section 4.1.4. QWC scores 3.45, whereas VZ scores 3.7 (scores per variable can be found in table 4.5).

Therefore, VZ is the chosen indicator set. VZ is set up around safety, health, and well-being. For this research, well-being is the most important aspect. Therefore, the indicators will only be used on this part; thus, all safety and health aspects are disregarded. Furthermore, some of the indicators of VZ are only applicable in certain situations or after certain events. Therefore, these are excluded from the set for this research. Four indicators excluded are regarding the department's onboarding program (3.1, 6.1) or hiring requirements (1.2, 5.2). These indicators can be considered if the department decides to work on the core problems mentioned in section 1.2.2. Three other indicators measure if evaluation is done after implementing well-being measures and programs (2.1, 5.1, 3.2). These indicators are not applicable for iterative use of the indicators, as they can only be answered when these measures and programs have been put up. This results in seven indicators. For each indicator, questions are provided. However, after reviewing these questions with the scrum masters, it became clear that some terms were unclear or needed additional explanation. Therefore, the questions are reviewed and changed to be clear, or only a basic explanation needs to be added. The final questions can be found in table 4.6. The first column gives the number of the indicator the question belongs to, corresponding with the number in figure C.1.

A few indicators are highlighted in this section. The first indicator is regarding leader commitment (1.1). This indicator is based on the principle that leaders, in the case of this department scrum masters, product owners and the manager (RTE), must show commitment to improve and stimulate well-being to stimulate their employees to work on well-being actively. This indicator is not included with the goal to 'grade' leaders but can be used as a check for leaders to see if their commitment is noticed so that they can improve when necessary. Secondly, indicator 4.1 regarding discussing well-being is an effective measure for the teams to see if the team can discuss this topic during meetings and to check if it is taken into account during discussions, for example, regarding work division. This aspect is important, as not discussing well-being can indicate employees not feeling comfortable or safe or well-being is overlooked during the general work process. Lastly, indicator 7.1, regarding follow-ups on suggestions, is crucial for this department, as employees mentioned that the previous attempts regarding well-being felt useless due to no follow-up conversations. When this indicator is insufficient, this might lead to employees not actively joining to prevent well-being issues, which can affect the department's sustainable employment.

In this table, the last three questions are not part of VZ. Because VZ is set up to determine if organizations prevent SHW issues well, aspects like mental energy, how an employee is doing, and other related aspects are not indicated. For this, different indicators from QWC are added. These are mental energy, work climate, and work tempo. With the addition of these questions, the department can measure its position in preventing well-being and the current situation.

Indicator	Question
1.1	Do leaders (PO, SM, RTE) visibly demonstrate their commitment to well-being in their work processes and behaviour?
2.2	Are reported unplanned well-being events followed up by leaders for investigation, learning/improvement, and feedback to those directly involved? Under these events fall any unexpected events regarding well-being (e.g. burnout, mental breakdowns)
4.1	Is well-being an integrated part of discussions in work meetings?
4.2	Is the department systematically considering well-being when planning and organizing work?
6.2	Is well-being discussed in refresher events or training?
7.1	Are employees' suggestions for improving well-being followed up adequately?
7.2	Do employees get recognition for excellent well-being performance? Excellent performance is defined as when employees look out for their own well-being, support their colleagues and notify leaders when they see possible incidents or emerging problems
QWC	Is the workload sustainable?
QWC	Is the work climate good to work in? A good work climate entails clear purpose, expectations and achievable goals for employees together with a fun mood
QWC	Is your mental energy at a level such that you can focus on work during an entire workday?

Table 4.6: Final list of questions

4.2 Design and implementation guidelines

In this section, a dashboard is designed using the chosen KPIs of sections 4.1.5 and 4.1.6. The design is followed by an implementations strategy, which answers the last sub-question: "*How can the department easily use the metrics on a day-to-day basis?*".

4.2.1 Performance Dashboard design

The department uses Azure DevOps as an agile planning tool. Azure DevOps has the option to make custom dashboards. As this tool is used regularly during the day to keep track of work and planning, using this tool will end in easier integration into the work process of the department. Furthermore, by using the dashboard of Azure DevOps, data will be automatically updated in the dashboard without the needed interference of employees. Unfortunately, Azure DevOps does have drawbacks in its dashboard function. Firstly, a drawback is the limited design options available. Only pre-designed widgets can be added, as well as custom queries that sort planned work based on filters. For the pre-designed widgets, no choice is available for changing the design. Furthermore, the overall design of the dashboard cannot be changed. This design includes the title, background and possibly other additions like parting lines. Queries have more design options. First, when choosing a numeric tile, the tile only shows the total number of results for the query. The background colour of this tile can be conditionally formatted. Furthermore, several types of charts can be chosen for queries, like pie- and line charts. Lastly, the size and placement of widgets can be changed; widgets need to be placed in squares and can have the size of AxB blocks.

With these advantages and disadvantages in mind, the choice is made to make at least the operational dashboard in Azure DevOps, as all of the chosen operational KPIs can be configured in Azure DevOps. As a result, visual analytics, as researched by, amongst others, Sedrakyan et al. [44], cannot be used to determine appropriate visualisations of the KPIs. Furthermore, due to the three KPIs levels, the dashboard's design needs to consider these levels. One dashboard can contain KPIs from different levels, but the sep-

aration of KPIs at different levels needs to be considered. Therefore, tactical KPIs can be integrated into the Azure DevOps dashboard, but separation needs to be kept in mind. Of the selected tactical KPIs, only lead time and cycle time are possible to add to the dashboard. These are both pre-designed widgets; thus, the design cannot be changed if used in Azure DevOps. The other metrics need to be made using another platform, like PowerBI. Lastly, the strategic metrics can not be added to Azure DevOps either. During the time scope of this research, not all dashboards can be designed. Therefore, it is chosen only to make the operational dashboard, including the tactical KPIs, possible in Azure DevOps.

The dashboard's design can be found in figure 4.2. It is important to note that in this figure, the graphs are based on data currently in Azure DevOps. However, as that is a tool not used by the department yet, the data is copied from another tool and not kept up to date. Therefore, the graphs and numbers do not represent the usually expected numbers. Sprint burndown, release burndown, velocity, cycle time, and lead time are pre-designed widgets in this design. Herefore, the design of these metrics must remain the same. The number of stories, the number of US done, and the flow load are numeric query tiles. These can be conditionally formatted such that the background colour changes. These tiles are based on passively presenting information unless action is necessary [45]. This way, the attention is directed towards the tiles only when necessary, giving the dashboard a clearer overview and making it easier to grasp quickly. This option can be used for the teams to keep track of possible situations where the numbers are too high or too low. Multiple colours are available in that case, but "traffic light visualisations" [45] are recommended to attract immediate attention and quickly indicate the situation.

As mentioned before, separation must be considered when combining operational and tactical metrics in one dashboard. Because cycle time and lead time are available in Azure DevOps, it was chosen to add these to the operational dashboard as this ensures the automatic update of these metrics. Due to limited design options, there is no possibility to separate the widgets from the others by a line or different colouring. Therefore, to ensure that it is clear that these widgets have a tactical function, this is discussed during a presentation for the entire department as well as with the scrum masters of each team. Furthermore, three boxes for explanation are added to the bottom of the dashboard, including a link to a more elaborate explanation document. This document includes an explanation of the configuration of all metrics so that teams can adjust these to their needs and goals. The separation of operational and tactical metrics is explained in this dashboard as well.

As eight of the fifteen metrics are included in the operational dashboard, seven must be added to a dashboard in PowerBI. These metrics can be added together in one dashboard to limit the total number of dashboards for the department. However, in the design, a dashboard should fit on one page. Therefore, separate views are needed for the tactical and strategic metrics, or they fit on one page and should be separated clearly by a line or different design.

4.2.2 Well-Being tool design

The International Social Security Association (ISSA) has set up a set of indicators for VZ. With this set of indicators, a guideline is set up for their use. Within this guideline, three options can be chosen with different time intensities. It is advised to start with the first option, which is the least time-intense. This option requires that employees answer one question per indicator with either yes or no. Then, based on the percentage of employees filling in yes or no, it can be evaluated how the department scores on that indicator. When all indicators are filled in, the department can see where their improvement points are to prevent well-being issues. The decision is made to make the well-being tool only for group discussions because employees have mentioned that filling in forms is tedious. That way, the questions can stimulate group discussions on well-being topics, but employees are not asked to fill in forms.

For the design of the well-being tool, the decision was made to keep it minimalistic. The research of Bomström et al. supports this choice [45], who found that dashboards regarding agile environments should have a passive minimal view that only causes reactions when necessary. It needs to be noted that this research was done regarding software development dashboards. Janes et al. [46] state that designs should

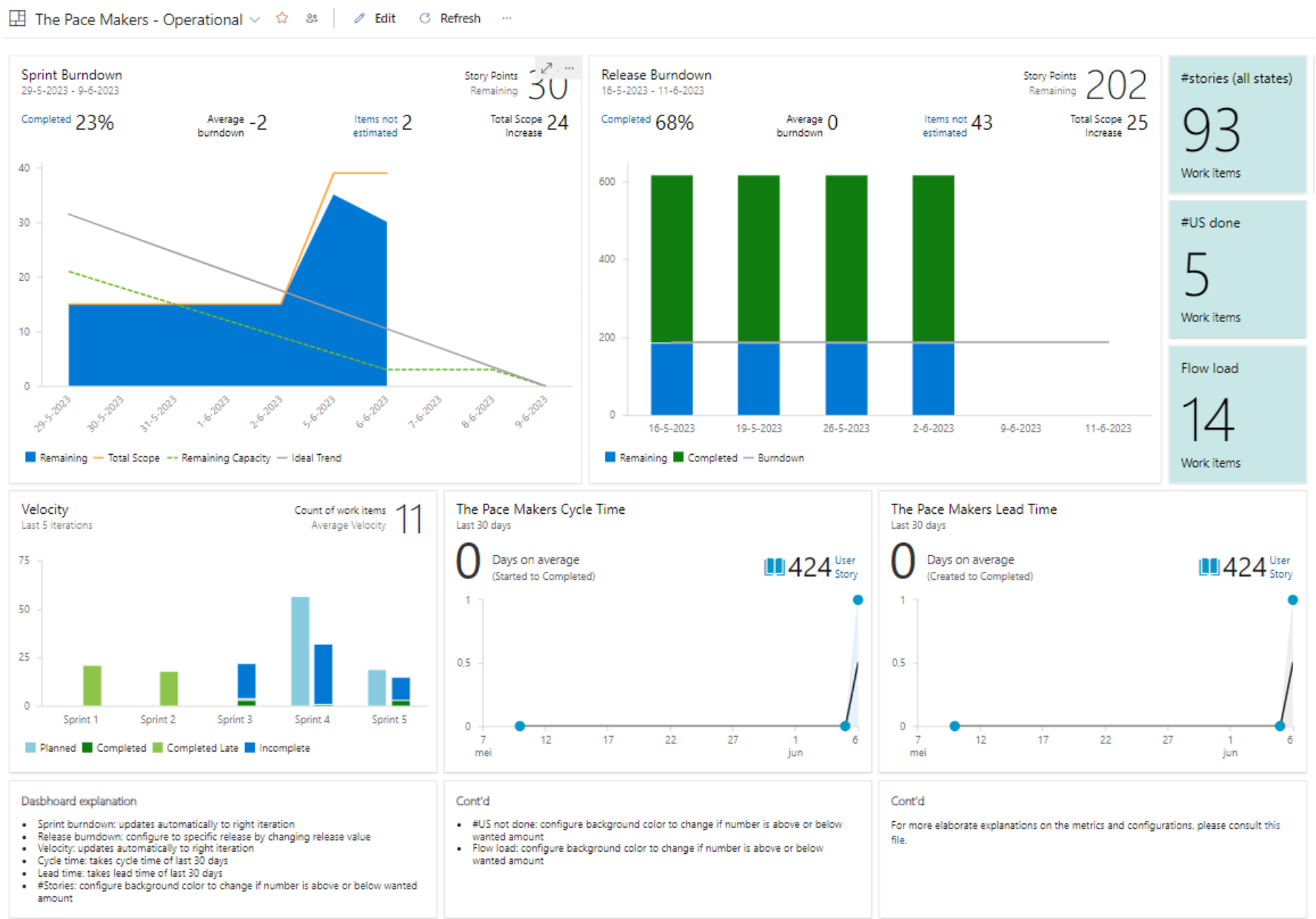

Figure 4.2: Operational performance dashboard.

also be minimalistic. However, they propose to choose between a 'pull' or 'push' method for dashboards. The choice between these strategies depends on the goal of a dashboard. Pull dashboards are meant to be interactive; the user needs to put effort into analysis but can obtain more information from this dashboard. Push dashboards are designed the other way around; all necessary information is shown to the user, so pushed. This type of dashboard is most useful if a user needs to receive information regarding issues and unplanned situations. Therefore, this push method is more suitable for this research.

The design of the dashboard can be found in figure 4.3. The dashboard is made in PowerBI. The questions are split up; two questions are discussed in each sprint. The team can discuss the questions together and fill in either yes or no for the questions by filling in a form. Figure F.18 shows an example of the form for sprint one. This form is linked to powerBI. Therefore, the dashboard automatically updates the answers when a form is filled in. As the questions are answered with either "yes" or "no", only data is available regarding the number of times this was answered. This data is a single numerical number. Evergreen [47] states that this type of data can be used when other data is not present to evaluate. In that case, presenting one number is sufficient to make the user see and understand the data. Furthermore, it is important to see for what indicator "yes" was answered for analysis. Therefore, only numeric values of the number of times "yes" is filled in are shown. That way, when this value is more than a certain value, it can easily be seen that that indicator is lacking. As with the operational dashboard, conditional formatting with traffic light visualisations attracts attention to the indicators not performing as needed. When clicking on an indicator, a table appears as shown in figure 4.4. This table gives an overview of all responses regarding that indicator.

Furthermore, the PI score is shown in a gauge chart. This chart type is useful when a numeric number can be within a specific range, where a goal needs to be met [48]. This score is determined by dividing the number of times "yes" is filled in by the total number of answers. This graphic uses traffic light visualisations, gradually changing colour from red to green depending on its score. On the meter, a target value of 0.8 is shown to help analyse the overall well-being of the department. When only one department needs to be evaluated, the filter on the left side can select single or multiple departments. Below all indicators, a concise explanation is provided to ensure that all employees understand the dashboard.

4.2.3 Implementation strategy for the dashboards

In multiple other pieces of research, it was found that metrics are used. However, most programs, including metrics, do not last longer than a year, which is considered an implementation failure [49]. There are numerous reasons for this failure, some corresponding with the problems mentioned in section 4.1.1. In this section, reasons for failure are identified based on the literature. Then, an incremental implementation plan is made for all dashboards. Finally, a work process is suggested to incorporate the dashboards into the department's existing processes.

Reasons for implementation failure

In research on the implementation of metrics programs, Hall and Fenton [50] identified fifteen expert recommendations. These recommendations substantially overlap with issues identified in other articles like Pfleeger [51]. In the implementation strategy, these problems need to be taken into account.

First, one problem is eliminated by choice of KPIs and requirements of section 4.1.2. This elimination is the fact that data collection should be automated as much as possible so that the extra effort employees need to put into metrics is minimized. This problem is mentioned multiple times in literature [49, 50, 51, 52, 53, 54]. Secondly, two problems are mentioned as ranking requirements. These requirements limit the problems as much as possible, though they still need to be evaluated during implementation to ensure they are limited or eliminated. These are complexity and perceived necessity. Complexity relates to the problem of employees not understanding what a metric depicts or not understanding how to analyze them. Keeping metrics accessible makes it more likely to be used in the long-term [49, 50, 51]. Perceived necessity relates to users finding the metric useful and needed. This aspect needs to be in-

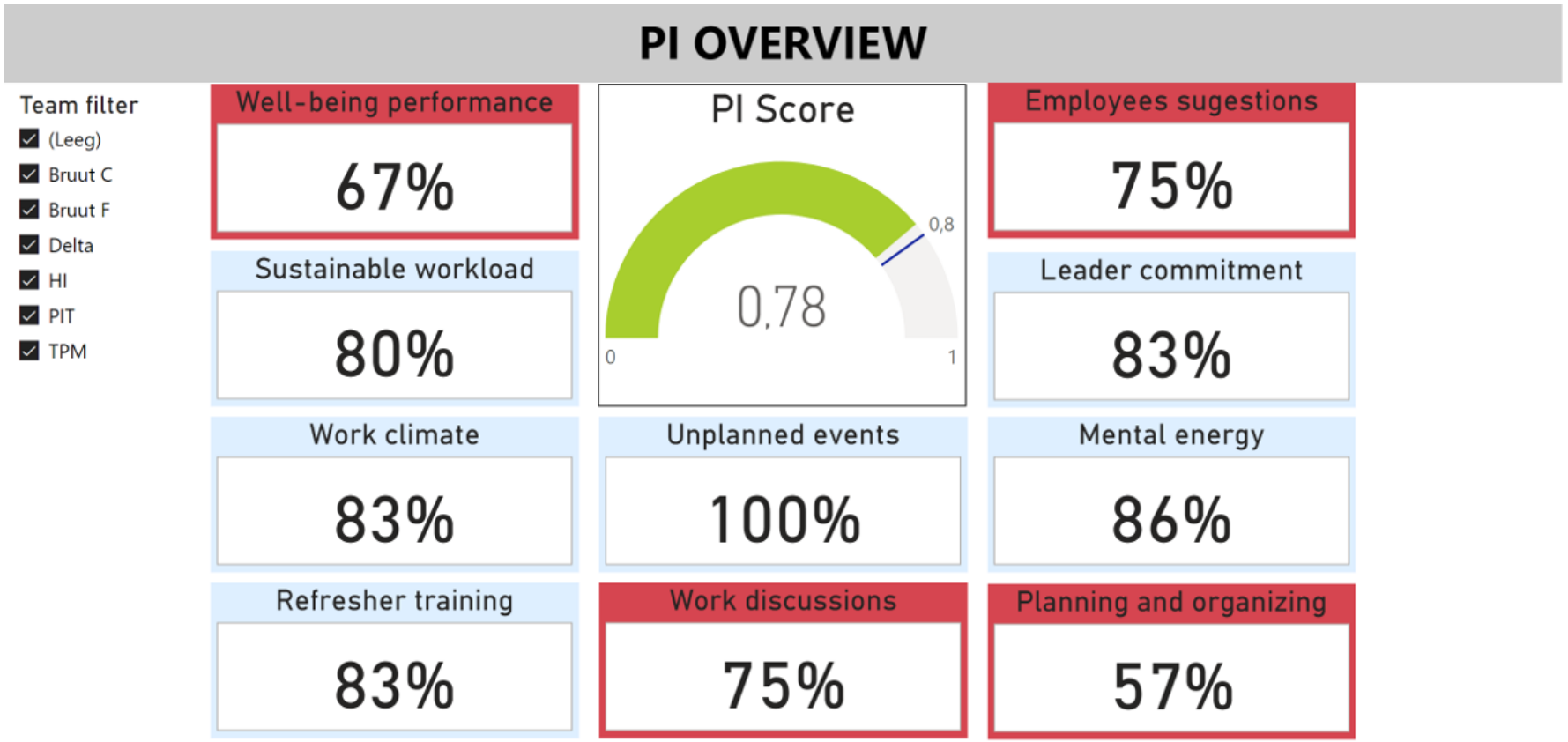

Figure 4.3: Well-being dashboard

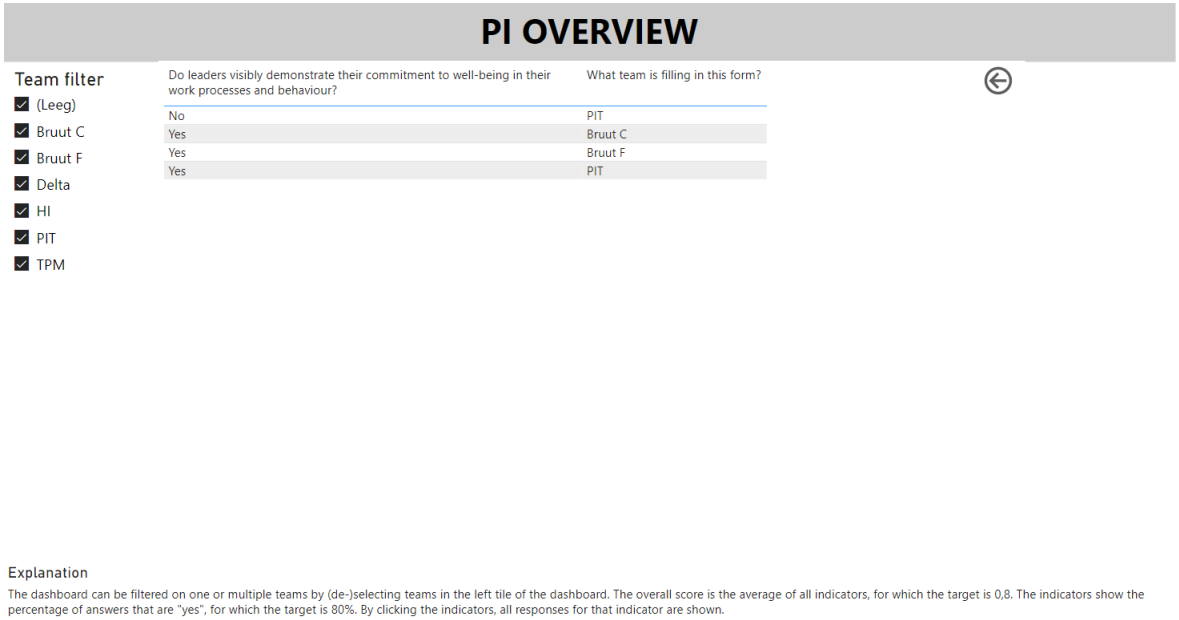

Figure 4.4: Well-being dashboard indicator responses

creased as much as possible, as otherwise, employees need to use metrics against their beliefs [50]. Using the selection method, including complexity and perceived necessity as ranking requirements, these are automatically increased positively as much as possible.

Further problems need to be solved through the implementation strategy. The first problems arise early in the process, during the design of the metrics. Metrics that fit the goal of the department are important to the implementation. The goals are already included in the choice of KPIs in this research. Furthermore, management and other important stakeholders with a higher function than the general metrics users need to commit to metrics and support the implementation [52, 53, 54]. This commitment will improve the subjective norm (section 3.4.2, together with the opinion of peers [49]. During the implementation design phase, further first problems arise. When change is necessary, it is important to include all stakeholders in the process [50, 51]. During the implementation of metrics, the same holds. Employees need to be included from problem identification until the last evaluation to increase the chance of implementation success. The reason is that by including employees, their reasons for using metrics are included, and possible problems can be identified in the early stages. This way, the metrics chosen are easier to relate to the process and its problems. When this relation is clear to all employees, the likelihood of metrics utilization increases significantly. Furthermore, by including them, employees feel that their opinion is valued and are more likely to have a positive attitude towards the metrics. Lastly, the last important problem before implementation is the required knowledge of employees. The knowledge needs to be adequate so all employees can understand straightforward metrics. Adequate knowledge can be arranged by training for all employees and in later stages for new employees [49, 52].

The following problems arise in the initial stages of implementation. Implementing new technology requires employees to adapt to new processes and structures. The adapting process entails that implementing metrics means employees must get used to using them, analyzing them, and regularly including them in the work process. This process should not be done at once, as the change suddenly impacts employees, resulting in a negative view. Therefore, incremental implementation is advised [50, 55]. This incremental implementation plan can include implementing metrics per project, starting with a project where metrics are viewed positively and are expected to yield significant results [51]. The first projects or teams starting with metrics can evaluate the process and, if positive, spread enthusiasm.

The last set of problems relates to the process of using metrics. Firstly, as mentioned by employees during

this research, motivation for using metrics decreases when results are not discussed [55, 54]. Furthermore, during that discussion, feedback and evaluation on the design and use of the metrics should be brought up so that metrics stay relevant to the department's goals [50, 55]. Secondly, metrics should be integrated into an existing work process, or a new transparent and regularly used process should be developed [53, 54]. A final problem regarding the use of metrics, causing negative views, is the assessment of metrics. Employees can think or feel that their work is criticized or evaluated based on the metrics [51, 53, 54]. This feeling will lead to them not using metrics or incorrectly using metrics. When the aim of metrics is precise, and the focus is on team productivity or product quality, individuals are not criticized, and employees can be assured they are not evaluated based on the metrics. Furthermore, the way of using and the goal of analysis can increase the results gathered from the metrics [52, 55]. In the case of measuring team productivity, it needs to be clear to a team what the aim of the metric is and focus on identifying bottlenecks or problems to improve, not on evaluating the team.

Incremental Implementation

First, before implementing a new dashboard, management must be on board with the implementation. This commitment improves the subjective norm and the feeling of 'needing' to use the dashboard, and as a result, employees are more likely to use the dashboard. When the dashboard and implementation are fully supported by management, the incremental implementation plan can be started. In order to give a clear overview, all steps of the strategy, accompanied by a concise explanation, are presented in table 4.7. Steps 5a and b can take place at the same time. Furthermore, within all evaluation steps, in case of feedback suggesting the dashboard works insufficiently, improvement has to take place before going to the next step.

By implementing the feedback steps, the teams can address their problems, the metrics and ways of using them can be improved, and employees are included in the implementation process. That eliminates the risk that employees feel the change is forced upon them or that the dashboards are not working optimally. Further, implementing feedback and changing the dashboard to the team's needs ensures it will relate to the team's needs. Furthermore, the perceived usefulness will likely increase by implementing feedback, and employees' attitudes towards using the dashboard will become more positive. Additionally, starting with one team that can try out the first dashboard can eliminate the first issues. Furthermore, they can get used to the new dashboard. When the first team has done that, they can tell other employees, which will positively influence the subjective norm.

The second possible implementation problem mentioned is the complexity of the dashboard. The dashboard is designed straightforwardly and has a consice explanation and an elaborate explanation file. Training employees is needed before using the dashboards to ensure that complexity still does not become an issue. The training intensity depends on the dashboards and their platforms/goals. In the case of the dashboards designed for this research, one explanatory presentation and the provided explanation file are satisfactory. Additionally, the perceived necessity can be increased during this training by explaining the use of the dashboards and the goals. Even though this is done by the first communication step of the implementation plan, repeating this during training can improve the perceived necessity as all users understand the technology better during and after training.

Lastly, it is important to note the problem addressed by [51, 53, 54] regarding the critique on products instead of employees. The dashboards and metrics are meant to improve teams' performance, with the end goal of sustainable employment. Therefore, the metrics measure team performance and analysis of the metrics can show points of improvement for the team. Management and teams themselves should never use metrics to discuss persons or to compare teams on their performance. Such discussions will lead to employees getting a negative attitude towards the metrics and will work counterproductive for the final goal.

	Step	Explanation
1	Communicate	Communicate the implementation plan, goals, and reasons to all stakeholders
2	Training	Train employees who use or change the dashboard in how to do that
3	First team implementation	Implement only one dashboard in the first team
4	Evaluate	Evaluate the use of the dashboard with the team
5a	Expand to other teams	Implement the first dashboard in all teams
5b	First team expansion	Implement all dashboards in the first team
6	Evaluate	Evaluate the use of all dashboards with the teams
7	Expand to other teams	Implement all dashboards in all teams
8	Evaluate	Evaluate the use of all dashboards with the teams

Table 4.7: Incremental implementation

Plan for the use of Performance Dashboard

The performance dashboards have three levels; operational, tactical, and strategic. This separation of levels results in particular moments of use. First of all, the operational dashboard should be utilized most often. This dashboard is meant as a tool to keep track of team performance daily. Therefore, in SAFe teams, this dashboard should be discussed during daily stand-ups. The metrics sprint burndown, flow load, and number of stories done update continuously and give an update on the team's progress during that sprint. The velocity and release burndown can be discussed less often but should still be discussed at least once every sprint.

Second, a tactical dashboard should be discussed every sprint or PI during, for example, a retrospective. The retrospective is a meeting focused on reviewing the last sprint to evaluate and find improvement points. The tactical dashboard can help guide this discussion with the selected metrics. Lastly, the strategic dashboard should be discussed once per PI at most. This dashboard is meant to measure aspects of the departments' strategy. At the end of each PI, the strategic dashboard can be consulted to look at the metrics and analyze those compared to the set goals. In section 4.2.1, it is discussed to combine the tactical and strategic dashboards. When this is the case, it needs to be made clear during the communication and training (steps 1 and 2 of the incremental implementation) that these are metrics of different levels and therefore need to be discussed in different intervals.

Plan for the use of the Well-Being Dashboard

The well-being dashboard gives an overview of the well-being of the entire train. Each team uses a form during the sprint retrospective to fill in the questions corresponding to that sprint. The team discusses and fills in these. Therefore, it is not an individual tool. At the end of a PI, an overview is given in the dashboard, showing the position of the teams concerning well-being and well-being prevention. When this overview shows that indicators are not up to a satisfactory level, the scrum masters take this to a meeting together with the manager. There, these problems are discussed, and further actions are planned. These actions can be on a team level. However, when multiple teams show the same problems, department-wide action can be taken. When these discussions have taken place, scrum masters always need to discuss this with the team, such that teams feel their input is used and their satisfaction with the discussions stays on a sufficient level.

5| Validation of Results

In this chapter, results found in chapter 4 are validated. First, all interviews are validated in section 5.1. Then, the validation of the dashboard designs is explained in section 5.2.

5.1 Validation of interviews

As interviews have reliability issues in stability, equivalence, and internal consistency [23] (corresponding to the issues mentioned in Dasom et al. [29]), it is important to validate the results. Therefore, the list of problems is discussed with the management, also called a member check [56]. All problems were known, and management either experienced or recognised that employees experienced them. Therefore, the list is validated. The same discussion has taken place regarding the requirements for selecting new metrics. All requirements have been accepted. Further validation is ensured by an explanation of the research process in section 4.1.1, corresponding to the concept of dependability as mentioned in Dasom et al. [29].

5.2 Validation of user acceptance

As discussed in section 3.4.2, user acceptance is a measure of the willingness of people to use metrics as intended [30]. In this study, a questionnaire was used to validate the design of the operational and well-being dashboards. The questionnaire aimed to assess the likelihood of the dashboards' adoption by determining people's willingness to use the dashboards. Of 63 employees, 32 responses were received, representing a response rate of approximately 51%. Table 5.1 provides the distribution of responses across different functions. It is important to note that management roles were slightly over-represented, constituting 25% of the responses. Additionally, one respondent outside the department, categorised as a stakeholder, completed the questionnaire. The results, related to functions and teams, can be found in appendix E.

The first results are regarding the operational dashboard. In figure 5.1, these results can be found. In the results, a score of 1 means that the responder strongly agrees with the statement, whereas 5 means that the responder strongly disagrees. Firstly, the attitude towards behaviour is measured. The graph reveals that 53% of the responses for this indicator were rated as 2, and in total, 91% of the responses fell within the range of 1 to 3. As for perceived behavioural control, 79% of the responses were rated between 1 and 3, with the peak at 2 (30%). Although relatively more responses indicated insufficient perceived behavioural control compared to attitude towards behaviour, nearly three-fifths of the responses are deemed sufficient. The subjective norm is the only indicator where less than 10% of the answers scored 1. However, 84% of the responses fell within the range of 1 to 3, with the highest peak at 2 (39%). Finally, for perceived usefulness, 87% of the responses within the interval of 1 to 3, with 36% of the responses rating it as 2. It should be noted that, as can be seen in figure 5.2, the responses of the latter three shift towards the middle range, while attitude towards behaviour has a significant peak at 2.

Function	**Number of responses**
Test specialist	7
Business analyst	5
Sap CD specialist	9
Stakeholder	1
Abap specialist	2
Management	8

Table 5.1: Employees functions

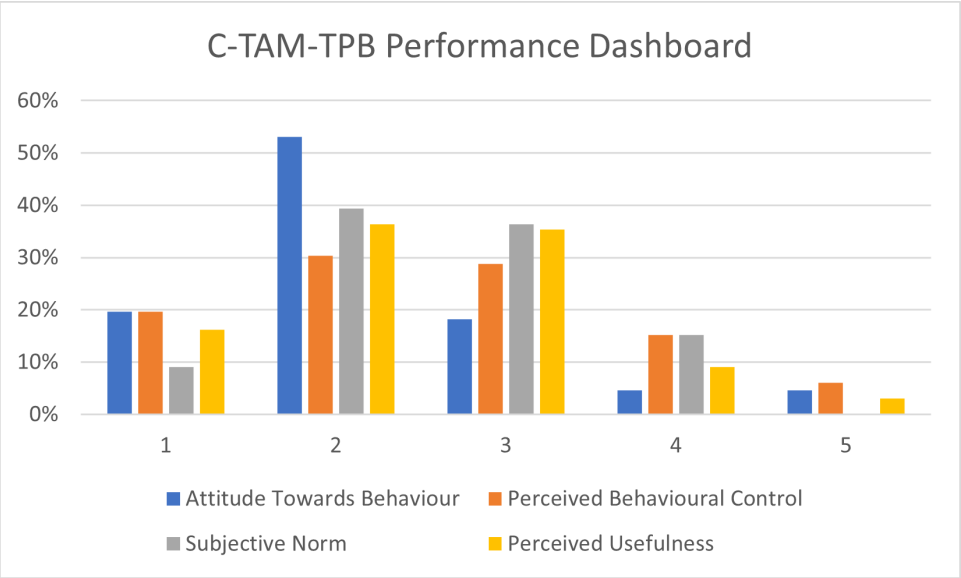

Figure 5.1: User Acceptance results performance dashboard

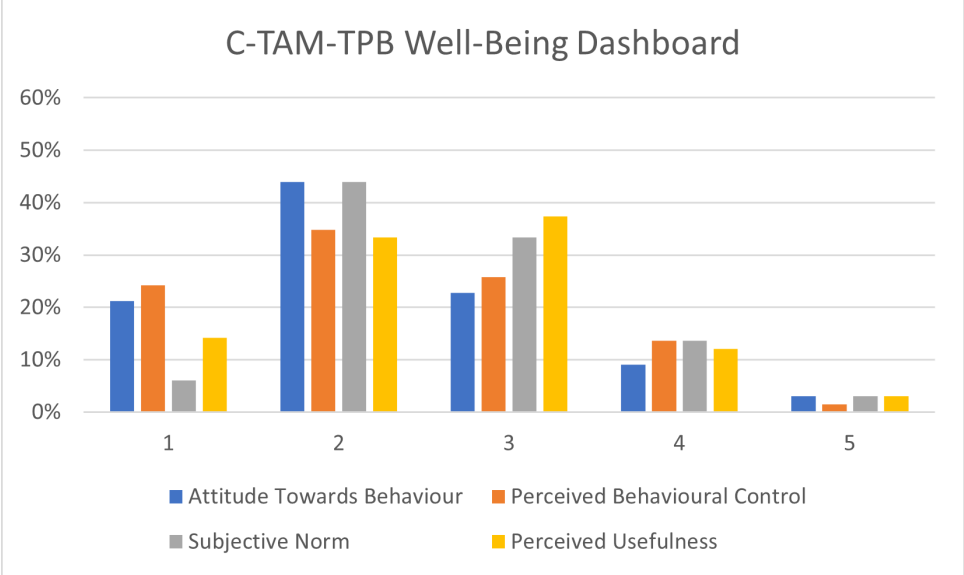

Figure 5.2: User acceptance results well-being dashboard

The second set of results shows the user acceptance of the well-being dashboard which can be found in figure 5.2. Attitude towards behaviour and subjective norm have similar results, peaking at 2 with 44%, with 88% and 83%, respectively, within the range of 1 to 3. Perceived behavioural control has a flatter curve, peaking at 35%, with 85% of the responses falling within the range of 1 to 3. Perceived usefulness for well-being is the only construct with the peak at 3 (37%), with 84% of the responses within the interval of 1 to 3.

No outliers were observed when considering the breakdown by functions, indicating a specific function scoring significantly below the average for both dashboards. The same holds for the division of teams. Additionally, all four indicators received positive scores. Therefore, the acceptance of both dashboards appears promising, providing evidence of the validity of the results.

6| Discussion

This chapter discusses the results (section 6.1). Furthermore, recommendations are made regarding these results and possible further research (section 6.3).

6.1 Discussion

In this study, we aimed to address the main research question of how to develop effective metrics for measuring performance and well-being in the context of sustainable deployment. We conducted interviews, literature reviews, and implementation research to achieve this. In this section, we critically reflect on our findings, establish links with our research questions, summarize the main findings and conclusions, and provide arguments and justifications for our choices.

The first sub-question is *"What are the problems with the current metrics?"* This question was researched using interviews, where employees mentioned multiple problems. Time and complexity were the primary issues discouraging employees from using dashboards. Moreover, the absence of a well-developed implementation plan led to metrics fading out of routine, not being discussed, or lacking motivation. It is important to note that the quality of interviews greatly depend on the interviewer, the setting, and other environmental influences. These influences impact the internal and external validity of the interviews [20]. For instance, an inexperienced interviewer's ineffective use of probing may yield superficial answers instead of in-depth conversations. As a result of influences, the interview results might be incomplete, as other problems could have gone unmentioned. Validity is checked to ensure that these influences do not negatively impact the results. This check is done by discussing the results, as proposed in Dasom et al. [29] and Koelsch [56]. This discussion has taken place with management and checked if the results found in section 4.1.1 are viable and, if known, well-reflected. This discussion had a positive outcome, ensuring credibility. Further validity is ensured by transparency on the methods in section 4.1.1. However, the results could still lack some other aspects not mentioned by employees and management as confirmability and transferability are not checked. Additionally, the interviewees were sampled by asking management and emailing one team. This sampling method resulted in interviews with management, one interested employee, and one employee with experience with metrics at a previous department. This sample does not completely represent the whole department, as the division of management/employees is not representative. Thus, it is crucial to consider that the results lean more towards management's perspective, potentially leading to employees not fully agreeing with the presented problems and requirements. Furthermore, the employees spoken to volunteered for an interview. These employees have a strong opinion on using dashboards compared to other employees. Therefore, the sample of interviewees might not accurately represent the entire department.

Following the interviews to identify current issues, several methods have been used to identify KPIs used in similar contexts. The first sub-question for this part is: *"Which KPIs are effective to keep track of performance?"* This question resulted in a literature review and the addition of some metrics mentioned by the department's employees. This research gave a list of forty KPIs. The second sub-question relates to the well-being aspect: *"Which KPIs are effective to keep track of well-being?"* This question was also answered using a literature review, resulting in a set of indicators that prevent well-being issues. Consistent inclusion and exclusion criteria were applied for both literature reviews. The inclusion criteria ensured that all articles were understandable by the researcher and that the subject regards the subject of the literature review. The exclusion criteria were added to ensure the quality and relevance of the selected articles. An example is the criteria that articles need to be cited by other articles. However, by using these exclusion criteria, possibly articles with valuable knowledge have been excluded from the research. Additionally, three databases were selected for this research. Other databases could have relevant articles for this research, which currently are not included. Therefore, the found KPIs can be elaborated in further research.

The well-being indicators researched focus on well-being at work. However, well-being aspects that the work environment does not influence can also still affect the well-being of employees. Therefore, it is important to remember that this set of indicators only keeps track of the work-environment well-being. KPIs may be mentioned in literature for both literature searches but not added in this research due to the inclusion requirements, like a minimum of 1 citation and the language. Therefore, when used in other research or departments, this list can be used as a base but can be expanded.

The fourth sub-question is: *"What requirements should be met for a metric with the goal of measuring performance and well-being?"* Interviews were conducted, and the list of problems was analyzed. For this set of interviews, as with the interviews of section 4.1.1, internal and external validity issues occur. A member check is conducted to ensure validity by discussing the requirements with management. However, further validation on confirmability and transferability is lacking. Therefore, the final list of requirements might be incomplete or partially invalid. With these requirements, it was made sure that the chosen KPIs relate to the department's goal and exclude KPIs that will result in problems the department experienced before. Because of previous experience and problems, a conscious decision was made to include requirements like time and complexity that exclude KPIs that could give valuable insight to the department. This exclusion supports a part of the objective of this research by excluding complex or time-intensive metrics. However, another goal is to find effective metrics to measure performance and well-being. Other metrics can possibly measure these aspects more effectively. In this research, preference is given to the first objective. This decision needs to be re-evaluated if this research is used in other contexts. Part of the requirements needed to be met to include KPIs, whereas others were preferences. Therefore, the requirements were split into inclusion requirements and ranking requirements. The KPIs were split into three categories; operational, tactical, and strategic based on updating and frequency of use. Based on these requirements and the weighted sum model, we selected fifteen performance KPIs and one set of indicators for well-being to effectively track the department's progress. To ensure alignment with the department's goals, we discussed the scores assigned to each KPI with the manager, following a pre-defined rubric. Therefore, the validity of these choices is strengthened with a member check and explanation of the process, ensuring credibility and transferability [29].To provide a comprehensive assessment of well-being, we expanded the set of well-being indicators by incorporating three additional metrics to monitor the department's current well-being. Furthermore, some indicators were excluded because they were not important to the department. The exclusion of these indicators does provide the risk that not all well-being issues are included. Furthermore, including new indicators and changing the questions of existing indicators implies that this new set of indicators need to be evaluated in further research to establish validity and reliability. For the well-being set, a dashboard was made in PowerBI. The choice was made for the performance indicators only to deliver a prototype of the operational dashboard to the department.

The last sub-question is: *"How can the department easily use the metrics daily?"* This question was answered by conducting implementation research, which resulted in an implementation strategy. First, reasons for implementation failures were researched. Reasons for failure were found in six articles. Other relevant articles might have been missed due to the choice of databases or search terms. Then, with these reasons and recommendations in mind, a strategy was proposed. However, it is important to note that further evaluation is necessary to confirm the effectiveness of the implementation strategy in practice. Based on this implementation strategy, the first communication about the dashboards has already occurred. Furthermore, the user acceptance of both dashboards is measured to validate the prototype dashboards and to ensure that long-term use is more likely than in the past. The responses show that employees score the dashboards positively on all four indicators, validating the effectiveness prototypes.

6.2 Limitations

In section 6.1, the results are discussed, revealing some limitations of this research. Firstly, the interviews' results regarding the current metrics and problems and the requirements are not validated on confirmability and transferability. Confirmability regards the researcher's bias and its effect on the research [29],

and transferability is the level at which the results can be generalized. In order to use these results for further research, these aspects of validity need to be checked. Furthermore, the sample of interviewees is only a small part of the department and thus not a complete representation of the department, which might have affected the results.

Secondly, the literature reviews have limitations regarding the exclusion criteria and chosen databases. The exclusion criteria exclude possible relevant articles from this research, which might have limited the final list of KPIs. Additionally, three databases are chosen, of which two are extensively searched. The exclusion of other databases could have led to missing relevant articles. The same databases have been used during the search for reasons for implementation failure, leading to the same limitation. Furthermore, the concepts searched for might have synonyms unidentified in this research and other concepts might be of relevance to this research.

Thirdly, during the research, the MCDM method has been used. In order to score KPIs, a rubric has been made based on the research objectives. However, this rubric is determined by the researcher and thus based on the researcher's perceptions. Furthermore, it is not scientifically proven to be valid. Therefore, this rubric limits the validity of the scoring process for the KPIs. The same limit applies to the weights of the WSM as determined in section 4.1.4.

Fourthly, the implementation strategy is proposed based on researched implementation failure. This strategy is not validated, as the research is finished before the completion of the implementation strategy. Therefore, further research is needed to evaluate the strategy. Additionally, KPIs are categorized based on available data input and the possibility for automated updates. This input is based on available data from the department and the existing infrastructure. Herefore, the technical infrastructure of the department has limited the research.

Lastly, a choice is made based on time and scope during the research. This choice was only to provide a prototype dashboard of the operational performance KPIs and well-being indicators. Therefore, this research lacks tactical and strategic performance dashboards. However, an implementation strategy is written for those dashboards. Therefore, when the choice is made to design these dashboards, the implementation strategy can be used.

6.3 Recommendations

First of all, the scope of this research leaves gaps around this research that are recommended to fill. These include the core problems found outside of the scope of this research because they fall under the purview of management. It is recommended that the department conducts further research on the problems and solutions in order to achieve the set goal. The problems that need attention are identified in section 1.2.1 as not-selected core problems. These are the insufficient onboarding program and the fact that metrics are an extra task. Management is recommended to look into the onboarding program and see what is needed to increase the knowledge of new employees concerning metrics and other related processes. Additionally, the VZ indicators are identified by the ISSA as a complete set, helping in preventing well-being issues. Three VZ indicators are excluded from the well-being dashboard because these are only applicable during the hiring and onboarding process. However, based on VZ, these indicators effectively prevent well-being issues. Therefore, the management should include the three VZ indicators regarding onboarding and hiring new employees as mentioned in section 4.1.6 when doing further research into the hiring and onboarding process. Furthermore, due to the change in the set of VZ indicators, evaluating the new set of indicators and their effectiveness in measuring well-being in practice is recommended.

Furthermore, even though the implementation strategy makes the metrics part of work processes, the fact that metrics are an extra task should be evaluated and discussed within the department. For the strategic metrics, which are not used regularly, management is recommended to guide the department in their use

and take responsibility for those as unclear responsibilities often result in issues [57, 58]. Though less regular, they should also become part of a reoccurring work routine. That way, they become part of the work processes and are not seen as extra effort. Additionally, a part of the tactical and strategic metrics selected in section 4.1.5 are not in a dashboard design, due to the time spanning this research. These metrics are of use to the department, based on the ranking and selection in sections 4.1.5 and 4.1.6, and therefore it is recommended that the tactical and strategic dashboards are developed.

Additionally, two recommendations are made for further research. Firstly, employees influence the data based on the working method with Azure DevOps, for example, when story points are put on 'completed'. The performance dashboards are affected by this influence. Thus, the effectiveness of the dashboards might be influenced. This effect should be further investigated. Secondly, in order to evaluate the dashboards, it is recommended to research options for automated data collection regarding interaction with the dashboards. This data can be used to evaluate the dashboards, their use and further validate the results of this research.

Lastly, several limitations are mentioned in section 6.2. It is recommended to expand the research based on these limitations. The first aspect is the validity of the interview results. The confirmability and transferability of the interviews are recommended to be evaluated to use the results in further research. Next, the literature reviews and research regarding implementation failure are recommended to be expanded by searching other databases and removing the exclusion criteria. Furthermore, it is recommended to evaluate the proposed implementation strategy in order to validate this strategy.

6.4 Scientific Contribution

This research makes valuable contributions to the field of KPIs and metrics, specifically in the context of sustainable deployment within the scaled agile framework (SAFe). Firstly, we identified and addressed the prevalent problems associated with the current metrics used within the department. While these problems are well-known in technology implementation [50], our findings validate their occurrence in our specific department, thereby reinforcing their significance.

Next, the research results add knowledge to the research area of KPIs and metrics, with a specific focus on SAFe. While research on possible KPIs to measure performance and well-being in agile environments exists (e.g. [1, 59]), there is a knowledge gap exists on metrics specifically for the scaled agile framework. Some metrics are recommended; however, for most metrics, it is said these are specific to the goal for which they are used [10] (e.g. sales metrics for sales environments). Therefore, by using the MCDM method and the WSM to select KPIs that align with SAFe, this research adds to this area of scientific knowledge. The connection between the KPIs and SAFe is established by the requirements and goals used during the research, which are incorporated in the MCDM method and WSM. For example, the research identifies KPIs that are useful to keep track of performance on a day-to-day basis, which supports the high flexibility and iteration structure of SAFe. For instance, the sprint burndown is found to be an effective measure of performance and planning in agile environments [60] and, intending to measure performance and be straightforward in use, is also applicable to SAFe. This knowledge can be generalized to other companies and departments that want to start using KPIs to measure their performance and well-being. For example, other departments of Achmea can adopt this framework, as the working method is SAFe, and the departments' goals are similar due to company-wide goals.

In section 4.2.3, results are presented regarding implementation problems. These problems are found in more departments and companies and are described in multiple articles [51, 50, 49]This research contributes to scientific knowledge by using MCDM to prevent one of these problems; the requirements are set up to automate data collection as much as possible, and to decrease further manual actions to update data in the dashboards. In further research, similar requirements can be used in order to prevent implementation failure due to data collection and updating issues. Furthermore, the incremental implementation plan is a recommended framework for implementation and can be generalized to other departments

and companies. It provides a framework for all situations regarding these problems. For example, the framework emphasizes the gradual adoption of dashboards to make the change effective and smooth. This research adds to the scientific knowledge regarding the implementation of dashboards by proposing an implementation framework which can be applied for dashboard implementation. Therefore, all departments wanting to start using dashboards as an integrated part of their work process can use the framework regardless of the dashboard or goals of the dashboard. For the framework to be generalized to technology implementation, further research is necessary to find possible additions or changes.

For both these contributions, it is important to note that the research focuses on finding a solution for a specific department. If the results are applied to a broader academic field, results need to be generalized. However, the results depend on the department's goals as the department works with an agile method. Therefore, generalization to agile or other flexible, iterative methods would be most effective. The results are likely not applicable to sectors where agile methods are not used. However, the framework for selecting KPIs and the incremental implementation strategy can be used more generally and adapted to different goals.

The research findings focus on one department's goals. These goals are regarding sustainable deployment. Enhancing performance and preventing well-being issues are chosen as the tools to work on sustainable deployment. Other departments within the company could use the results of this research when they choose to work on this goal, as other departments within Achmea have the same structure and work with SAFe as well. However, the list of problems and requirements should be validated by the management of that department to ensure that the research results align with the department. When other companies want to use these findings, the goals and work processes must be reviewed, after which problems and requirements can be deduced. When these are not similar, results must be adapted based on these goals or work processes. However, the research method could be applied in all fields.

7| Conclusion

An implementation strategy is proposed to implement a set of metrics, as identified in section 4.1, to keep track of performance and well-being at work. With that proposal, the research objective is met, and the research question is answered. This research aimed to identify metrics that effectively keep track of performance and well-being. With a basis of knowledge from literature and analysis of existing problems and requirements, a set of KPIs is identified, and two dashboards are designed. With these dashboards, an implementation strategy is written to maximize the effect and utilization of the metrics.

The department had multiple issues regarding metrics, one of which was a need for more knowledge of effective KPIs. Therefore, a literature search was conducted to ensure a knowledge basis for further research. This research was expanded with interviews, adding to the problems mentioned during the problem identification. An unexpected result of the interviews was that the department currently or in the past used metrics that are not found in the literature. Three KPIs mentioned by the employees were not found in the literature. In total, 42 KPIs are found, of which 40 are included in the list of KPIs for performance. This list of possible KPIs can be used in other departments and companies, as it is a general list of KPIs for measuring performance in an agile work environment. During the search for well-being indicators, an unexpected result was the identification of sets of indicators by multiple researchers. This finding has led to the choice of selecting a set of indicators as opposed to individual indicators.

For selecting KPIs, a list of requirements was composed of interviews and an analysis of the problems. Because some requirements need to be met and others only give preference to KPIs, the requirements were split into inclusion and ranking requirements. Furthermore, during the interviews, it became clear that the department has multiple goals with the metrics. Therefore, the KPIs were split into operational, tactical, and strategic. This categorization resulted in selecting three sets of KPIs for performance and the recommendation of designing two separate dashboards and a prototype for the operational dashboard. Though the requirements are department specific, the selection method can be generalized, and the split into categories can be used in a broader scope. The chosen well-being indicators were adapted to the department's needs, and the dashboard was designed. In the final set of indicators, the VZ indicators identify improvement points for workplace well-being. In contrast, the added indicators identify current problems with the team's well-being. These are two causal problems in figure 1.1. However, in the problem cluster, it was assumed that the identification of improvement points was caused by information on the current team's well-being. The result, the combination of VZ and QWC indicators, shows that these run side-by-side and together can prevent and solve problems, leading to more sustainable employment.

Finally, the last part of the listed problems was to be solved with an adequate implementation strategy. In the literature, some problems mentioned by the employees were found. Additionally, other problems are described that generally result in implementation failure. Therefore, the implementation strategy is based on the problems mentioned and the literature recommendations. The implementation strategy has two aspects; the incremental implementation plan and a recommendation for including the solution in the work process. The incremental implementation plan can be generalized. However, the recommendation for inclusion in work processes is specific to the work process of the department and, therefore, should be reviewed before being used by other departments or companies. By using this implementation strategy and the design of the operational and well-being dashboard, the department can enhance performance, manage workload, and foster sustainable employment. Teams can easily keep track of their performance regarding their planning and current task load. That way, teams can start discussions regarding differences in workload over specific periods or the total workload. Performance can be increased, the workload can be managed, and the entire process can be maintained sustainably by using the dashboard as an indicator and discussion starter. Furthermore, the well-being tool helps start discussions on the prevention of well-being issues and helps identify possible ongoing well-being issues. If the department's management

actively acts upon the discussions and provides answers, this will support the department in achieving its goal. To conclude, the proposed KPIs to measure performance and well-being, the prototype dashboards and the implementation strategy will positively contribute towards the goal of sustainable employment.

References

[1] Zwetsloot, G., Leka, S., Kines, P., and Jain, A. Vision zero: Developing proactive leading indicators for safety, health and wellbeing at work. *Safety Science*, 130:104890, (2020), https://www.sciencedirect.com/science/article/pii/S0925753520302873.

[2] Geerts, G. L. A design science research methodology and its application to accounting information systems research. *International Journal of Accounting Information Systems*, 12(2):142–151, (2011), https://www.sciencedirect.com/science/article/pii/S1467089511000200. Special Issue on Methodologies in AIS Research.

[3] Arnetz, B. B. Subjective indicators as a gauge for improving organizational well-being. an attempt to apply the cognitive activation theory to organizations. *Psychoneuroendocrinology*, 30(10):1022–1026, (2005), https://www.sciencedirect.com/science/article/pii/S030645300500096X. Stress, sensitisation and somatisation: A special issue in honour of Holger Ursin.

[4] Achmea. About us. https://www.achmea.nl/en/about-us. [Accessed 19-Apr-2023].

[5] SAFe 6.0 Framework — scaledagileframework.com. https://scaledagileframework.com/#. [Accessed 28-Apr-2023].

[6] March, S. T. and Smith, G. F. Design and natural science research on information technology. *Decision Support Systems*, 15(4):251–266, (1995), https://www.sciencedirect.com/science/article/pii/0167923694000412.

[7] Hevner, A. and Chatterjee, S. Design science research in information systems. In *Integrated Series in Information Systems*, pages 9–22. Springer US, Boston, MA, (2010).

[8] Peffers, K., Tuunanen, T., Rothenberger, M., and Chatterjee, S. A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3):45 – 77, (2008), https://www.scopus.com/inward/record.uri?eid=2-s2.0-65249190803&partnerID=40&md5=3c54e220e271e7d58a5c58b3ce8afeb9.

[9] Suddaby, R. From the editors: What grounded theory is not. *The Academy of Management Journal*, 49(4):633–642, (2006), http://www.jstor.org/stable/20159789.

[10] Measure and Grow - Scaled Agile Framework — scaledagileframework.com. https://scaledagileframework.com/measure-and-grow/. [Accessed 26-May-2023].

[11] Horkoff, J., Barone, D., Jiang, L., et al. Strategic business modeling: representation and reasoning. *Software & Systems Modeling*, 13(3):1015–1041, (2014), https://doi.org/10.1007/s10270-012-0290-8.

[12] Naeemah, A. J. and Wong, K. Y. Sustainability metrics and a hybrid decision-making model for selecting lean manufacturing tools. *Resources, Environment and Sustainability*, 13:100120, (2023), https://www.sciencedirect.com/science/article/pii/S2666916123000130.

[13] Poppendieck, M. and Poppendieck, T. *Lean software development: an agile toolkit*. Addison-Wesley, (2003).

[14] Tripathi, V., Chattopadhyaya, S., Bhadauria, A., et al. An agile system to enhance productivity through a modified value stream mapping approach in industry 4.0: A novel approach. *Sustainability*, 13(21), (2021), https://www.mdpi.com/2071-1050/13/21/11997.

[15] de Oliveira Neto, F. G., Horkoff, J., Svensson, R., et al. Evaluating the effects of different requirements representations on writing test cases. In Madhavji, N., Pasquale, L., Ferrari, A., and Gnesi, S., editors, *Requirements Engineering: Foundation for Software Quality*, pages 257–274, Cham, (2020). Springer International Publishing.

[16] Wiseman, J., McLeod, J., and Zubrick, S. R. Promoting mental health and well-being: integrating individual, organisational and community-level indicators. *Health Promotion Journal of Australia*, 18(3):198–207, (2007), https://onlinelibrary.wiley.com/doi/abs/10.1071/HE07198.

[17] Vayrynen, S. T. and Kiema-Junes, H. K. Exploring blue- and white-collar employees' well-being at work system. *International Journal of Sociotechnology and Knowledge Development*, 10(2):14–34, (2018), https://doi.org/10.4018/ijskd.2018040102.

[18] Adegbite, W. M., Bawalla, O. G., and Adedeji, O. Measuring employees' well-being among nigerian bankers: Exploring the socio-cultural indicators. *Journal of Workplace Behavioral Health*, 35(4):279–304, (2020), https://doi.org/10.1080/15555240.2020.1834866.

[19] Arnetz, B. B. Staff perception of the impact of health care transformation on quality of care. *International Journal for Quality in Health Care*, 11(4):345 – 351, (1999), https://www.scopus.com/inward/record.uri?eid=2-s2.0-0032816947&doi=10.1093%2fintqhc%2f11.4.345&partnerID=40&md5=9ce18a06dedeed498ddecdca3f4213e4. Cited by: 81; All Open Access, Bronze Open Access.

[20] Broniatowski, D. A. and Tucker, C. Assessing causal claims about complex engineered systems with quantitative data: internal, external, and construct validity. *Systems Engineering*, 20(6):483–496, (2017).

[21] Nightingale, A. A guide to systematic literature reviews. *Surgery (Oxford)*, 27(9):381–384, (2009), https://www.sciencedirect.com/science/article/pii/S0263931909001707. Determining surgical efficacy.

[22] Alshenqeeti, H. Interviewing as a data collection method: A critical review. *English linguistics research*, 3(1):39–45, (2014).

[23] Schindler, P. S. *Business research methods*. (2018).

[24] Fontana, A. and Frey, J. H. The interview. *The Sage handbook of qualitative research*, 3:695–727, (2005).

[25] Kvale, S. and Brinkmann, S. *Interviews: Learning the craft of qualitative research interviewing*. sage, (2009).

[26] Zhang, Y. and Wildemuth, B. M. Unstructured interviews. *Applications of social research methods to questions in information and library science*, pages 222–231, (2009).

[27] Aruldoss, M., Lakshmi, T. M., and Venkatesan, V. P. A survey on multi criteria decision making methods and its applications. *American Journal of Information Systems*, 1(1):31–43, (2013).

[28] Triantaphyllou, E., Shu, B., Sanchez, S. N., and Ray, T. Multi-criteria decision making: an operations research approach. *Encyclopedia of electrical and electronics engineering*, 15(1998):175–186, (1998).

[29] Im Dasom, Pyo Jeehye, L. H. J. H. O. M. Qualitative research in healthcare: Data analysis. *J Prev Med Public Health*, 56(2):100–110, (2023), http://www.jpmph.org/journal/view.php?number=2270.

[30] Dillon, A. User acceptance of information technology. *Encyclopedia of human factors and ergonomics*, 1:1105–1109, (2001).

[31] Ajzen, I. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2):179–211, (1991), https://www.sciencedirect.com/science/article/pii/074959789190020T. Theories of Cognitive Self-Regulation.

[32] Marangunić, N. and Granić, A. Technology acceptance model: a literature review from 1986 to 2013. *Universal Access in the Information Society*, 14(1):81–95, (2014), https://doi.org/10.1007/s10209-014-0348-1.

[33] Davis, F. and Davis, F. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13:319–, (1989).

[34] Taylor, S. and Todd, P. Assessing it usage: The role of prior experience. *MIS Quarterly*, 19(4):561–570, (1995), http://www.jstor.org/stable/249633.

[35] Davis, F. D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340, (1989), http://www.jstor.org/stable/249008.

[36] Schepers, J. and Wetzels, M. A meta-analysis of the technology acceptance model: Investigating subjective norm and moderation effects. *Information Management*, 44(1):90–103, (2007), https://www.sciencedirect.com/science/article/pii/S0378720606001170.

[37] Saraih, Ummi Naiemah, Zin Aris, Ain Zuraini, Abdul Mutalib, Suhana, et al. Examining the relationships between attitude towards behaviour, subjective norms and entrepreneurial intention among engineering students. *MATEC Web of Conferences*, 150:05011, (2018), https://doi.org/10.1051/matecconf/201815005011.

[38] Ruiz-Herrera, L. G., Valencia-Arias, A., Gallegos, A., et al. Technology acceptance factors of e-commerce among young people: An integration of the technology acceptance model and theory of planned behavior. *Heliyon*, 9(6):e16418, (2023), https://doi.org/10.1016/j.heliyon.2023.e16418.

[39] Zainuddin, Z. Q. M., Yahya, F., Moung, E. G., et al. Effective dashboards for urban water security monitoring and evaluation. *International Journal of Electrical and Computer Engineering (IJECE)*, 13(4):4291, (2023), https://doi.org/10.11591/ijece.v13i4.pp4291-4305.

[40] Thompson, V. A. Interpretational factors in conditional reasoning. *Memory & Cognition*, 22(6):742–758, (1994), https://doi.org/10.3758/bf03209259.

[41] Verschueren, N., Schroyens, W., Schaeken, W., and d'Ydewalle, G. The interpretation of the concepts 'necessity' and 'sufficiency' in forward unicausal relations. *Current psychology letters*, (14, Vol. 3, 2004), (2004), https://doi.org/10.4000/cpl.433.

[42] Khalifa, A. S. Strategy and what it means to be strategic: redefining strategic, operational, and tactical decisions. *Journal of Strategy and Management*, 14(4):381–396, (2021), https://doi.org/10.1108/JSMA-12-2020-0357.

[43] King, M. F., Renó, V. F., and Novo, E. M. L. M. The concept, dimensions and methods of assessment of human well-being within a socioecological context: A literature review. *Social Indicators Research*, 116(3):681–698, (2014), https://doi.org/10.1007/s11205-013-0320-0.

[44] Sedrakyan, G., Mannens, E., and Verbert, K. Guiding the choice of learning dashboard visualizations: Linking dashboard design and data visualization concepts. *Journal of Computer Languages*, 50:19–38, (2019).

[45] Bomström, H., Kelanti, M., Annanperä, E., et al. Information needs and presentation in agile software development. *Information and Software Technology*, 162:107265, (2023), https://www.sciencedirect.com/science/article/pii/S0950584923001192.

[46] Janes, A., Sillitti, A., and Succi, G. Effective dashboard design. *Cutter IT Journal*, 26:17–24, (2013).

[47] Evergreen, S. D. H. Effective data visualization: The right chart for the right data. *SAGE*, (2017).

[48] Franken, D. Designing a dashboard for the sales department of company x, (2022).

[49] Umarji, M. and Emurian, H. Acceptance issues in metrics program implementation. In *11th IEEE International Software Metrics Symposium (METRICS'05)*, pages 10 pp.–20, (2005).

[50] Hall, T. and Fenton, N. Implementing effective software metrics programs. *IEEE Software*, 14(2):55–65, (1997).

[51] Pfleeger, S. Lessons learned in building a corporate metrics program. *IEEE Software*, 10(3):67–74, (1993).

[52] Gopal, A., Krishnan, M., Mukhopadhyay, T., and Goldenson, D. Measurement programs in software development: determinants of success. *IEEE Transactions on Software Engineering*, 28(9):863–875, (2002).

[53] Jeffery, R. and Berry, M. A framework for evaluation and prediction of metrics program success. In *[1993] Proceedings First International Software Metrics Symposium*, pages 28–39, (1993).

[54] Offen, R. and Jeffery, R. Establishing software measurement programs. *IEEE Software*, 14(2):45–53, (1997).

[55] Iversen, J. and Mathiassen, L. Lessons from implementing a software metrics program. In *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, pages 11 pp. vol.1–, (2000).

[56] Koelsch, L. E. Reconceptualizing the member check interview. *International Journal of Qualitative Methods*, 12(1):168–179, (2013), https://doi.org/10.1177/160940691301200105.

[57] Costello, T. Raci—getting projects "unstuck". *IT Professional*, 14(2):64–63, (2012).

[58] Elonen, S. and Artto, K. A. Problems in managing internal development projects in multi-project environments. *International Journal of Project Management*, 21(6):395–402, (2003), https://www.sciencedirect.com/science/article/pii/S0263786302000972. Selected papers from the Fifth Biennial Conference of the International Research Network for Organizing by Projects. Held in Renesse, Seeland, The Netherlands, 28-31 May 2002.

[59] Greening, D. R. Agile enterprise metrics. In *2015 48th Hawaii International Conference on System Sciences*, pages 5038–5044, (2015).

[60] Lai, S.-T., Susanto, H., and Leu, F.-Y. Project management mechanism based on burndown chart to reduce the risk of software project failure. In Barolli, L., editor, *Advances on Broad-Band Wireless Computing, Communication and Applications*, pages 197–205, Cham, (2022). Springer International Publishing.

[61] Uludag, O., Philipp, P., Putta, A., et al. Revealing the state of the art of large-scale agile development research: A systematic mapping study, (2020).

[62] Literature guide for Industrial Engineering and Management | Service Portal | University of Twente — utwente.nl. https://www.utwente.nl/en/service-portal/university-library/find-access-literature/guides-per-discipline/industrial-engineering-and-management. [Accessed 06-Apr-2023].

[63] Kurnia, R., Ferdiana, R., and Wibirama, S. Software metrics classification for agile scrum process: A literature review. In *2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pages 174–179, (2018).

[64] Almeida, F. and Carneiro, P. Performance metrics in scrum software engineering companies. *International Journal of Agile Systems and Management*, 14(2):205, (2021), https://doi.org/10.1504/ijasm.2021.118061.

[65] Mohsen, W., Aref, M., and ElBahnasy, K. Software metrics for cooperative scrum based ontology analysis. In *2017 2nd International Conference on Knowledge Engineering and Applications (ICKEA)*, pages 60–70, (2017).

[66] Mahnic, V. and Zabkar, N. Measuring progress of scrum-based software projects. *Elektronika ir Elektrotechnika*, 18(8):73–76, (2012), https://eejournal.ktu.lt/index.php/elt/article/view/2630.

[67] Agarwal, M. and Majumdar, R. Tracking scrum projects tools, metrics and myths about agile. *International Journal of Emerging Technology and Advanced Engineering*, 2(3):97–104, (2012).

[68] Getting started with a Definition of Done (DoD) — scrum.org. https://www.scrum.org/resources/blog/getting-started-definition-done-dod. [Accessed 10-May-2023].

8| Appendices

A Systematic Literature Review Performance

A systematic literature review is done to find an answer to the question '*Which KPIs are effective to keep track of performance (production-wise)?*'. In this appendix, the process of finding and selecting valuable articles can be found.

A.1 Search Terms

First, the key concepts need to be found to identify the search terms. From the research question, already two concepts can be found. These are KPI and performance. However, due to the specification that the KPIs need to be effective for the department, a different concept has been added, which already covers this. The added key concept is the Scaled Agile Framework, as the department works with this framework. Uludağ et al. [61] have already done a mapping study regarding this framework and selected related search terms. These have been reviewed, and some are selected for this literature review as related, narrow or broader terms for this framework. In table A.1, all the search terms linked with the key concepts can be found. These terms are used in the search query to find different articles. After finding helpful articles, reference tracking is used to get more on the topic.

A.2 Criteria

After using all the search terms, the articles need to be selected. For this, inclusion and exclusion criteria are used. The inclusion criteria are things the articles need to contain. If this is not the case, they will be disregarded. The exclusion criteria work the other way around; if an article does contain that, they are disregarded. The criteria are in table A.2.

A.3 Sources

To select usable sources, the UT library, particularly the page for Industrial Engineering and Management [62], was consulted. Scopus and Web of Science are selected as the most significant databases. These are selected because these sources are mentioned as important databases, and they are broad databases with peer-reviewed articles. Therefore, these can give many results. If these results are insufficient or contain too many articles unrelated to this research, Business Source Elite will also be used. This database is more specific to the area of research and is in the list of relevant databases of the UT library for this study.

	Key concepts	**Related terms**	**Narrower terms**	**Broader terms**
1	KPI	core indicator	-	metric, indicator, analytics, tracker
2	Performance	efficiency, functioning, efficacy, productivity	production rate, performance rate	value
3	Scaled Agile Framework	scaling agile frameworks, large scale scrum, scrum at scale	scrum of scrums, large solution scaled agile framework, essential scaled agile framework	lean agile, scaled agile lean development, agile

Table A.1: Search terms

Inclusion criteria	Exclusion criteria
Written in English or Dutch	Not peer-reviewed
Articles should be about KPIs or one of the related terms mentioned in table A.1	Published before the year 2000
	Not cited by other articles

Table A.2: Inclusion and exclusion criteria

A.4 Search Log

Source	Search string	Total hits	Remarks
Scopus	(TITLE-ABS-KEY (kpi) AND TITLE-ABS-KEY (performance))	3660	This are too many hits; the search string needs to be narrowed
Scopus	(TITLE-ABS-KEY (kpi) AND TITLE-ABS-KEY (performance) AND TITLE-ABS-KEY (agile))	124	A quick scan of the titles shows that some articles are promising, but still, many articles do not seem relevant
Scopus	((TITLE (kpi) OR TITLE (”key performance indicator”)) AND TITLE-ABS-KEY (agile))	**12**	By having KPI in the title, the search is narrowed. A large sample of the titles seems usable, and thus these will be further evaluated.
Scopus	(TITLE-ABS-KEY (metric) AND TITLE-ABS-KEY (performance) AND TITLE-ABS-KEY (agile))	424	This are too many results, and not all seem helpful, so the search string needs to be narrowed
Scopus	(TITLE (metric) AND TITLE-ABS-KEY (performance) AND TITLE-ABS-KEY (agile))	**46**	The titles seem promising, so these articles are further evaluated
Scopus	(TITLE (metric) AND TITLE-ABS-KEY (performance) AND TITLE-ABS-KEY (”scaled agile”))	0	
Scopus	(TITLE (metric) AND TITLE-ABS-KEY (performance) AND TITLE-ABS-KEY (”large scale scrum”))	0	
Scopus	(TITLE (metric) AND TITLE-ABS-KEY (performance) AND TITLE-ABS-KEY (”scrum”))	**9**	The titles all seem interesting, so these are further evaluated
Scopus	(TITLE (metric) AND TITLE-ABS-KEY (performance) AND TITLE-ABS-KEY (scrum) AND TITLE-ABS-KEY (scale))	0	

Continued on following page

Table A.3, continued.

Source	Search string	Total hits	Remarks
Web of Science	(TI=(kpi) OR AB=(kpi) OR AK=(kpi) OR TI=(key performance indicator) OR AB=(key performance indicator) OR AK=(key performance indicator)) AND (TI=(performance) OR AB=(performance) OR AK=(performance))	15872	Too many results
Web of Science	(TI=(kpi) OR TI=(key performance indicator)) AND (TI=(performance))	1142	Still too many results
Web of Science	(TI=(kpi) OR TI=(key performance indicator)) AND (TI=(performance)) AND TI=(agile)	**1**	The article seems useful and will be further evaluated
Web of Science	(TI=(kpi) OR TI=(key performance indicator)) AND (TI=(performance)) AND TI=(scrum)	1	Based on the title, this article is not valuable for the research

Table A.3: Inclusion and exclusion criteria

A.5 Article Selection

The search found that Web of Science contains a limited range of articles on an agile work environment in combination with KPIs, as this search only resulted in one article. Scopus did have this; most of the results come from that database. In total, 68 hits are selected for further evaluation. There are seven double titles of these hits, so 61 unique articles are further evaluated. Based on the inclusion and exclusion criteria, 36 articles are left. Of these articles, eight were selected based on their abstracts, of which four have accessible full versions. With reference tracking, two more useful articles were found. Therefore, these 6 are the final selected articles to compare [63, 59, 64, 65, 66, 67].

A.6 Performance KPIs mentioned in selected articles

KPI	Explanation	[63]	[59]	[64]	[65]	[66]	[67]
Velocity	This metric determines the amount of work done in a period.	x	x	x	x	x	x
Velocity deviation	Standard deviation of velocity divided by expected velocity		x				
Sprint burndown	Number of story points remaining during a sprint	x		x	x	x	
Release burndown	Number of story points remaining for a release	x		x	x	x	
Standard violation	The total amount of standards (in design, processes or anything else) violated during a sprint			x			x
Defects per iteration (sprint)	The total amount of defects in a sprint			x			x

Continued on following page

Table A.4, continued.

KPI	Explanation	[63]	[59]	[64]	[65]	[66]	[67]
Defect density	The number of defects over the size/complexity of the project			x			
Level of automation	Percentage of automated tests compared to the total amount of tests			x			x
# stories	The number of story points in a sprint	x		x			x
# tests	The number of tests in a sprint			x			
Progress chart	Tasks are listed in three columns, 'To do', 'in progress' and 'done', to keep track of progress			x			
Work in progress	The number of story points currently in progress			x			
Business value derived	The value a company derives from a user story	x		x			x
Earned value management	Performance is measured using the set budget, planning and the outcomes	x				x	
Sprint goal success	The frequency in which sprint goals are met			x			
Forecast horizon	The sum of story points in the backlog down to the first user story without an estimation of story points		x			x	
Lead time	The total time it takes to release an epic or user story		x	x	x		
True sprint length or cycle time	The time it takes to release a sprint increment		x				
Downstream impact	There are multiple metrics to determine downstream impact, showing the dependency within an organization or department		x				

Table A.4: Performance KPIs mentioned in selected articles

B Performance KPIs

B.1 Metrics mentioned by employees

Metric	Explanation
#US pushed	Total number of US pushed to the next sprint
Cancellation rate	Percentage of cancelled US compared to US in the sprint backlog
Changed or new requirements	Number of changed or new requirements added after refinement plotted against time
Critical test cases automated	The percentage of critical test cases that are automated compared to all critical test cases
Customer satisfaction	The customer, end-user of businesses satisfaction with the released product
Cycle time	Per phase or per US
Defect fixing capacity	Percentage of capacity spent on defect fixing compared to the total capacity
Escaped defects	Defects experienced by users after testing, either total amount, type or list
First time pass rate	Percentage of tests passed on first execution for new requirements compared to all tests for new requirements
Requirements coverage	Percentage of requirements successfully covered by tests, can be split per associated risk level
Right first time	Percentage of US that are accepted by the business or other customers the first time
#US done	The amount of US done (Definition of Done [68]) at the end of a sprint
Team member turnover	The number of team members replaced during a certain period
Test automation failure	Percentage of test cases planned for test automation not completed successfully
Test execution rate	Number of tests executed, can be split per test variety
Test pass rate	Percentage of tests that pass, can be split per test variety
User story ping pong	Number of times US go back on the backlog or to other teams/team members due to unclear requirements

B.2 All performance KPIs

KPI	Explanation	Level	Score
Sprint burndown	Number of story points remaining during a sprint	Operational	4.8
Release burndown	Number of story points remaining for a release	Operational	4.8
# stories	The number of story points in a sprint	Operational	4.9
# tests *	The number of tests in a sprint	Operational	4.6
Progress chart	Tasks are listed in three columns, 'To do', 'in progress' and 'done', to keep track of progress	Operational	4.1
Work in progress	The number of story points currently in progress	Operational	4.2

Continued on following page

Table B.5, continued.

KPI	Explanation	Level	Score
#US done	The amount of US done (Definition of Done [68]) at the end of a sprint	Operational	5.0
Test execution rate*	Number of tests executed, can be split per test variety	Operational	3.7
Velocity	This metric is meant to determine the amount of work done in a period.	Operational	4.7
Escaped defects *	Defects experienced by users after testing, either total amount or type	Operational	3.3
Flow load	Number of work items currently in progress (active or waiting)	Operational	4.7
Flow distribution	Percentage of work items in the system per type	Operational	4.1
Defects per iteration*	The total amount of defects in an iteration	Tactical	4.0
Standard violation **	The total amount of standards (in design, processes or anything else) violated during an iteration	Tactical	3.1
Defect density*	The amount of defects over the size/complexity of the project	Tactical	2.8
Forecast horizon	The sum of story points in the backlog down to the first user story without an estimation of story points	Tactical	3.7
Lead time	The total time it takes to release an epic or user story	Tactical	3.9
True sprint length or cycle time	The time it takes to release a sprint increment	Tactical	4.0
#US pushed	Total number of US pushed to the next sprint	Tactical	4.1
Cancellation rate	Percentage of cancelled US compared to US in the sprint backlog	Tactical	4.0
Changed or new requirements	Number of changed or new requirements added after refinement plotted against time	Tactical	3.8
Defect fixing capacity*	Percentage of capacity spent on defect fixing compared to the total capacity	Tactical	3.3
First time pass rate*	Percentage of tests passed on first execution for new requirements compared to all tests for new requirements	Tactical	3.9
Requirements coverage*	Percentage of requirements successfully covered by tests, can be split per associated risk level	Tactical	3.5
Right first time*	Percentage of US that are accepted by the business or other customers the first time	Tactical	4.3
Team member turnover**	The number of team members replaced during a certain period	Tactical	4.0
Test automation failure*	Percentage of test cases planned for test automation not completed successfully	Tactical	3.5
Test pass rate*	Percentage of tests that pass, can be split per test variety	Tactical	4.2
User story ping pong	Number of times US go back on the backlog or to other teams/team members due to unclear requirements	Tactical	4.3
Level of automation*	Percentage of automated tests compared to the total amount of tests	Tactical	4.2

Continued on following page

Table B.5, continued.

KPI	Explanation	Level	Score
Critical test cases automated*	The percentage of critical test cases that are automated compared to all critical test cases	Tactical	4.2
Velocity deviation	Standard deviation of velocity devided by expected velocity	Strategic	4.0
Business value derived**	The value a company derives from a user story	Strategic	3.3
Earned value management**	Performance is measured using the set budget, planning and the final outcomes	Strategic	3.1
Sprint goal success**	The frequency in which sprint goals are met	Strategic	3.5
Value stream mapping**	For this metric, processes need to be mapped out. The subprocesses are valued, which results in an overview of value per step in the process	Strategic	3.3
Customer satisfaction**	The customer, end-user or businesses satisfaction with the released product	Strategic	3.0
Downstream impact**	There are multiple metrics to determine downstream impact, showing the dependency within an organization or department	Strategic	3.0
Flow efficiency**	Percentage of the time spent on work that adds value compared to the total amount of work	Strategic	3.6
Flow predictability	How predictable a team or department is, do they measure up to their promises	Strategic	3.8

Table B.5: All performance KPIs

*Test-related KPI, out of scope. **Input need to be added manually

C Systematic Literature Review Well-Being

A systematic literature review is done to find an answer to the question '*Which KPIs are effective to keep track of performance (well-being)?*'. In this appendix, the process of finding and selecting helpful articles can be found.

C.1 Search Terms

First, the key concepts need to be found to identify the search terms. From the research question, already two concepts can be found. These are KPI and performance. However, because performance, in this case, is about the well-being of employees, another key concept is added: well-being. The search terms linked with the key concepts can be found in table C.6. These terms are used in the search query to find different articles. After finding helpful articles, reference tracking is used to get more on the topic.

C.2 Criteria

After using all the search terms, the articles need to be selected. For this, inclusion and exclusion criteria are used. The inclusion criteria are things the articles need to contain. If this is not the case, they will be disregarded. The exclusion criteria work the other way around; if an article does contain that, they are disregarded. The criteria are in table C.7.

C.3 Sources

To select usable sources, the UT library, specifically the page for Industrial Engineering and Management [62], was consulted. Scopus and Web of Science are selected as the most important databases. These are selected because these sources are mentioned as important databases, and they are broad databases with peer-reviewed articles. Therefore, these can give many results. If these results are insufficient or contain too many articles unrelated to this research, PsycINFO (EBSCO) will also be used. This database is more specific to the area of research and is in the list of most significant databases of the UT library for psychology.

Key concepts	Related terms	Narrower terms	Broader terms
KPI	core indicator	-	metric, indicator, analytics, tracker
Well-being	positive mental health	happiness, health, satisfaction, joy, peace of mind, fulfilment, energy	mental health, flourishing

Table C.6: Search terms

Inclusion criteria	Exclusion criteria
Written in English or Dutch	Not peer reviewed
Articles should be about KPIs or one of the related terms mentioned in table C.6	Published before the year 2000
	Not cited by other articles

Table C.7: Inclusion and exclusion criteria

C.4 Search Log

Source	Search string	Total hits	Remarks
Scopus	(TITLE-ABS-KEY (well-being) OR TITLE-ABS-KEY (well-being)) AND TITLE-ABS-KEY (kpi)	20	After a quick scan of the titles, the articles do not seem to be relevant
Scopus	TITLE-ABS-KEY (kpi) AND TITLE-ABS-KEY (mental) AND TITLE-ABS-KEY (health)	6	The articles do not seem relevant based on their titles
Scopus	TITLE-ABS-KEY (kpi) AND TITLE-ABS-KEY (flourishing)	1	This article does not seem relevant
Scopus	TITLE-ABS-KEY (well-being) AND TITLE-ABS-KEY (metric)	2047	The search string is too broad, and the first few titles do not seem relevant
Scopus	TITLE-ABS-KEY (well-being) AND TITLE-ABS-KEY (metric) AND TITLE-ABS-KEY (work)	287	Too many results, most titles do not seem relevant
Web of Science	(Ts=(wellbeing) OR Ts=(well-being)) and Ts=(kpi)	11	These articles do not seem relevant
Web of Science	(TI=(indicator) OR AB=(indicator) OR AK=(indicator)) AND (TI=(wellbeing) OR AB=(wellbeing) OR AK=(wellbeing))	12060	Too many results
Web of Science	(TI=(wellbeing) OR TI=(well-being)) and TI=(indicator)	468	Too many results
Web of Science	(TI=(wellbeing) OR TI=(well-being)) and TI=(indicator) and Ts=(work)	**58**	Some articles seem relevant, so they are selected
Scopus	TITLE (indicator) AND (TITLE (well-being) OR TITLE (well-being))	513	The search term is too broad; most titles do not seem relevant
Scopus	(TITLE (well-being) OR TITLE (well-being)) AND TITLE (indicator) AND TITLE-ABS-KEY (work)	**71**	Some articles seem relevant, so they are selected

C.5 Article selection

In total, 129 articles are selected for further evaluation. Based on the in- and exclusion criteria, 37 articles are discarded. Then, 33 more results were discarded due to double titles. Therefore, the selection is a total of 59 articles. The titles of these articles were reviewed, after which 20 were selected for further evaluation. The abstracts of these articles were read, and based upon that, five articles are the final selection [3, 16, 17, 18, 1].

C.6 Sets of Indicators found

In figure C.1 and table C.9, the sets of indicators mentioned in section 2.2 are listed. These are cited from the sources mentioned in the captions.

No.	PROACTIVE LEADING INDICATOR	AIM (Short description; see details in section Fact sheets)
1.1	Visible leadership commitment	Through visible leadership commitment, leaders demonstrate their commitment to SHW and actively promote SHW improvement.
1.2	Competent leadership	Committed and intrinsically motivated SHW leadership is essential to drive the development processes of VISION ZERO.
2.1	Evaluating risk management	Evaluation of the effectiveness of SHW risk management shows leadership focus and commitment to improving SHW, and supports organizational learning and continuous development.
2.2	Learning from unplanned events	Learning from unplanned events (incidents, events, cases) contributes to preventing similar undesirable events from (re)occurring.
3.1	Workplace and job induction	Integrating SHW in induction processes demonstrates that SHW are an integral part of each job and each business process.
3.2	Evaluating targeted programmes	Evaluating targeted SHW programmes (for example temporary campaigns) helps to verify that they are implemented as intended, and improvement goals are met.
4.1	Pre-work briefings	Integrating SHW in pre-work briefings allows for the identification of context specific hazards, risks and prevention measures prior to work.
4.2	Planning and organization of work	Planning and organization of work is essential for the success of every organization and for ensuring SHW.
5.1	Innovation and change	Technological, organizational and personnel changes occur frequently in organizations and should be considered proactively to improve SHW from the start in the design phase.
5.2	Procurement	Procurement can determine SHW risks for a long period. The indicator aims to trigger the systematic use of procurement for SHW improvement.
6.1	Initial training	Initial training is key to ensuring good SHW and to qualifying leaders and workers before they start their jobs.
6.2	Refresher training	Refresher training ensures that leaders and workers' knowledge and skills on SHW remain up to date.
7.1	Suggestions for improvement	When suggestions for SHW improvements are welcomed and are taken seriously, it stimulates active commitment and contributes to SHW improvement.
7.2	Recognition and reward	Recognition and reward for SHW involves showing appreciation for engaging in desired SHW behaviours.

Figure C.1: Vision Zero indicators, cited from [1]

Indicator	Explanation
Mental energy	Employee-ratings of: feelings of restlessness, irritability, worry, feeling low, moodiness, difficulty concentrating during the last month (4-point scale)
Work climate	Atmosphere at work, cohesion among coworkers, supportive atmosphere among coworkers
Work tempo	Time for planning work duties, sufficient time to execute tasks, time to reflect upon/consider how tasks had been carried out, time to consider how work processes could be improved in one's department
Performance feedback	Clear work directives from the immediate supervisor, feedback from supervisor when a task has been done well and poorly, respectively
Skills development	Professional skills development in one's work, immediate supervisor provides an employee with opportunities for skills development, opportunities for a more advanced position within health care, one's skills are utilized in current position, current job tasks offer
Goal clarity	Workplace goals are: well-defined, realistic, influenceable, accessible
Participatory management	Opportunity to influence workplace decisions, actual influence over workplace decisions concerning desire, latitude for deciding how work should be done, latitude for deciding what tasks should be done, sufficient influence concerning responsibilities, access to adequate information to carry out work duties efficiently, information from immediate supervisor sufficiently concrete to be useful in one's work professional development
Efficacy	Planning of work duties, employees strive toward the same goals, resources are used optimally at work, and the decision-making process is functional
Leadership	Immediate supervisor: clear in their communication, act consequently, has described how to achieve departmental goals, provide opportunities to develop employee's professional skills, open for change in workplace organization and work habits
Internal communication	Adequate information to carry out work duties efficiently, information from immediate supervisor sufficiently concrete to be useful in one's work, immediate supervisor is clear in his/her communication style, employee opportunity to comment information from immediate supervisor

Table C.9: QWC indicators, cited from [3]

D C-TAM-TPB Questionnaire

Question	Variable
In which team are you? What is your function?	General
I think positive about working with the dashboard/tool I would like to work with the dashboard/tool	Attitude towards behaviour
I have the resources, skills and knowledge to use the dashboard/tool I have the intention to use the dashboard/tool	Perceived behavioural control
I am more likely to use the dashboard/tool if my coworkers recommend using it I am more likely to use the dashboard if management recommends using it	Subjective norm
I think using the dashboard is useful Using the tool will improve my work Using the tool/dashboard makes my work easier	Perceived usefulness

E User Acceptance graphs

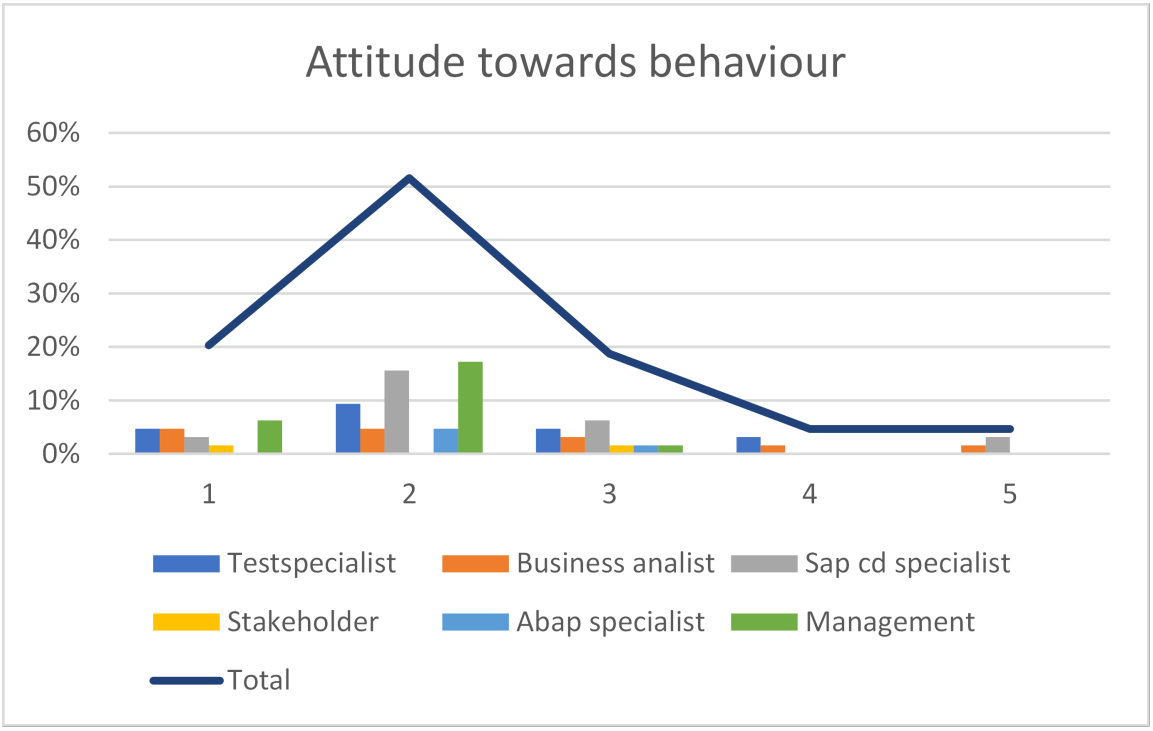

Figure E.2: Attitude towards behaviour operational dashboard per function

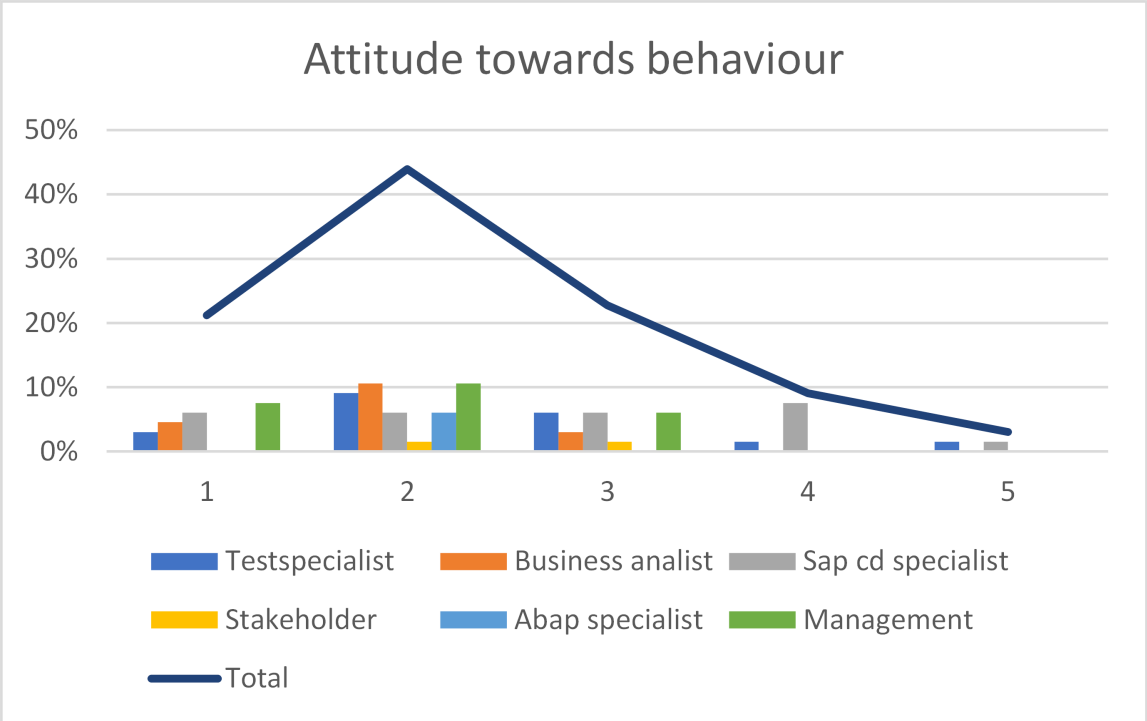

Figure E.3: Attitude towards behaviour well-being dashboard per function

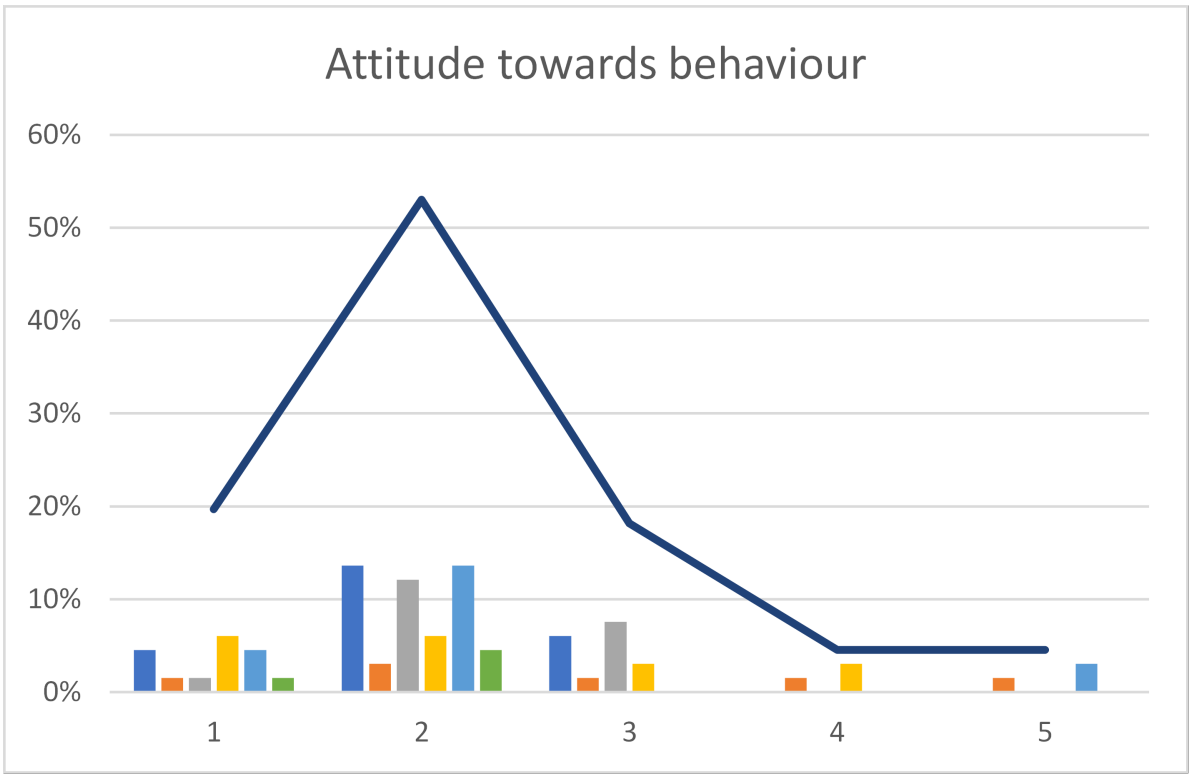

Figure E.4: Attitude towards behaviour operational dashboard per team

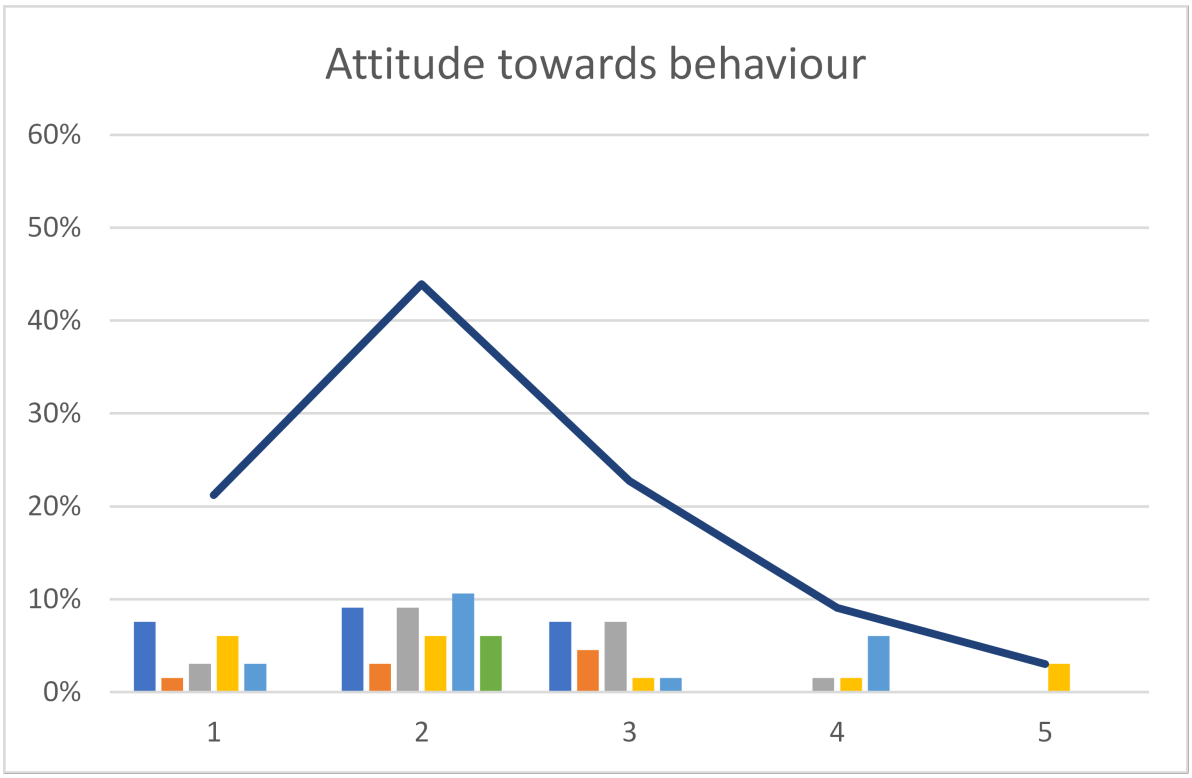

Figure E.5: Attitude towards behaviour well-being dashboard per team

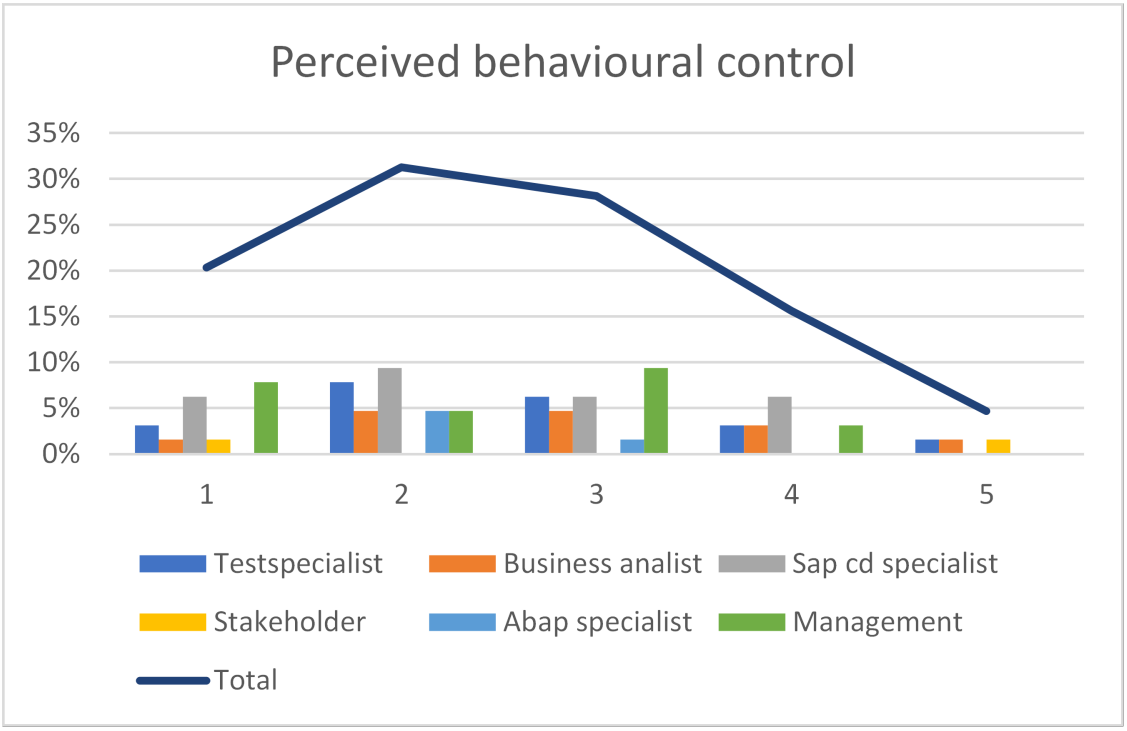

Figure E.6: Perceived behavioural control operational dashboard per function

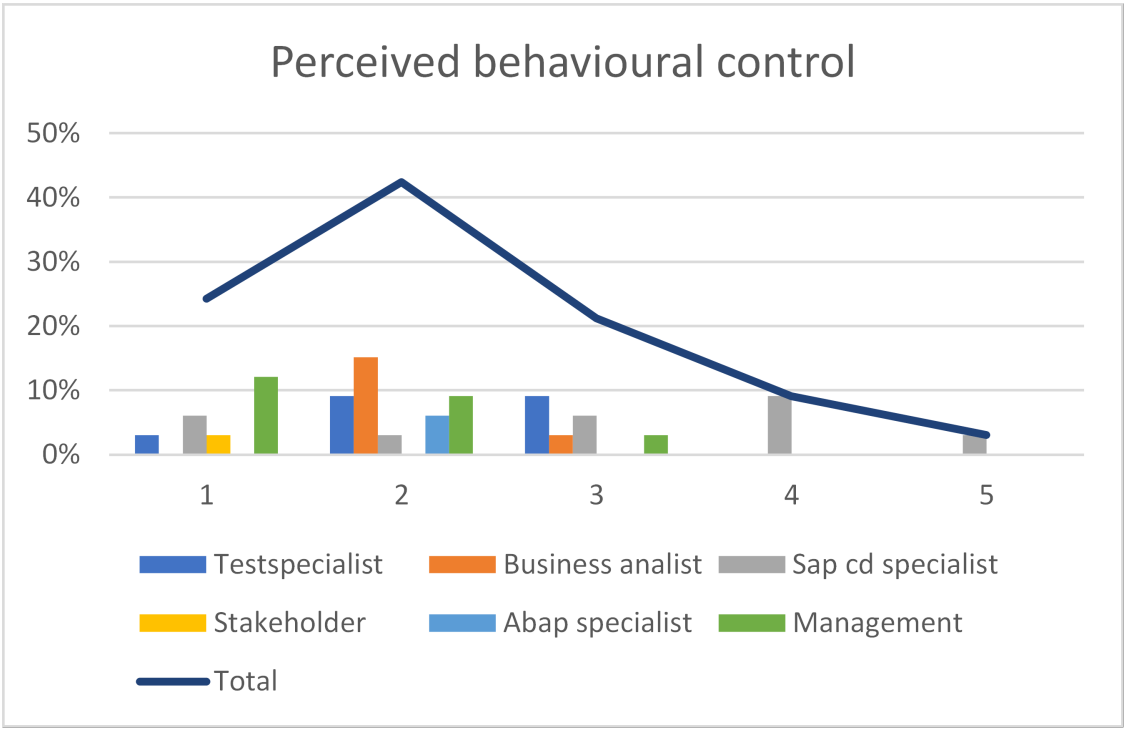

Figure E.7: Perceived behavioural control well-being dashboard per function

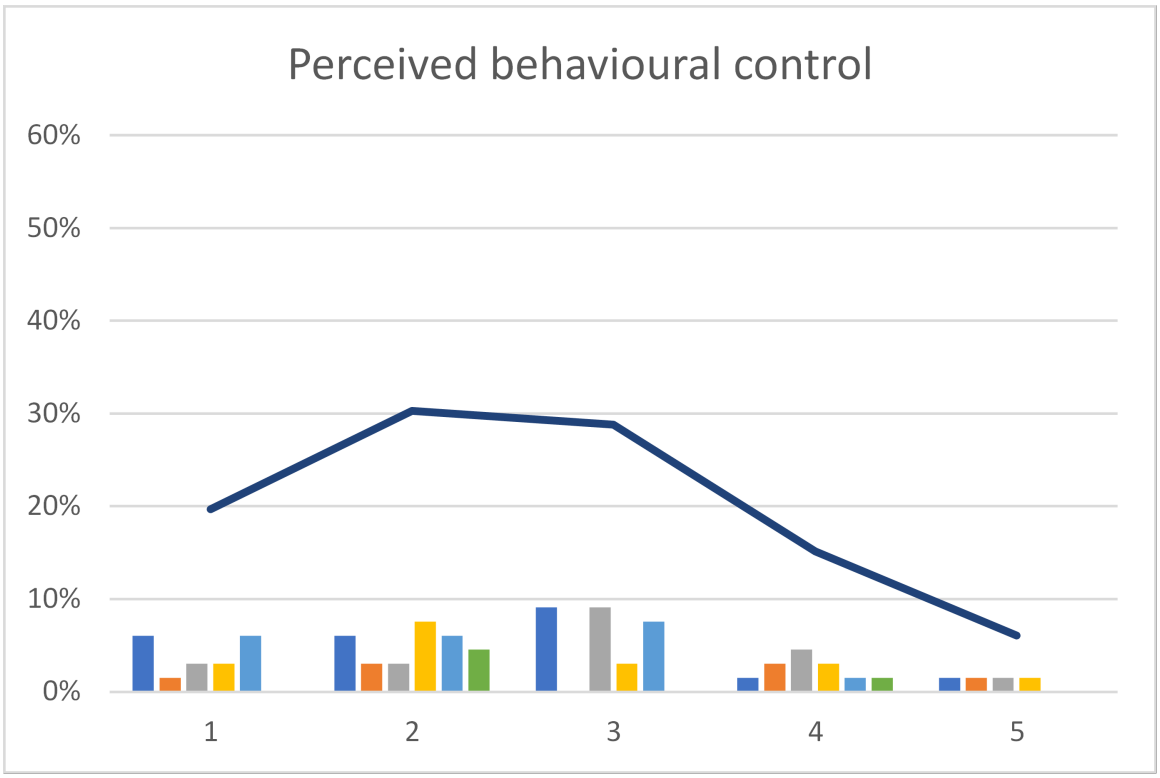

Figure E.8: Perceived behavioural control operational dashboard per team

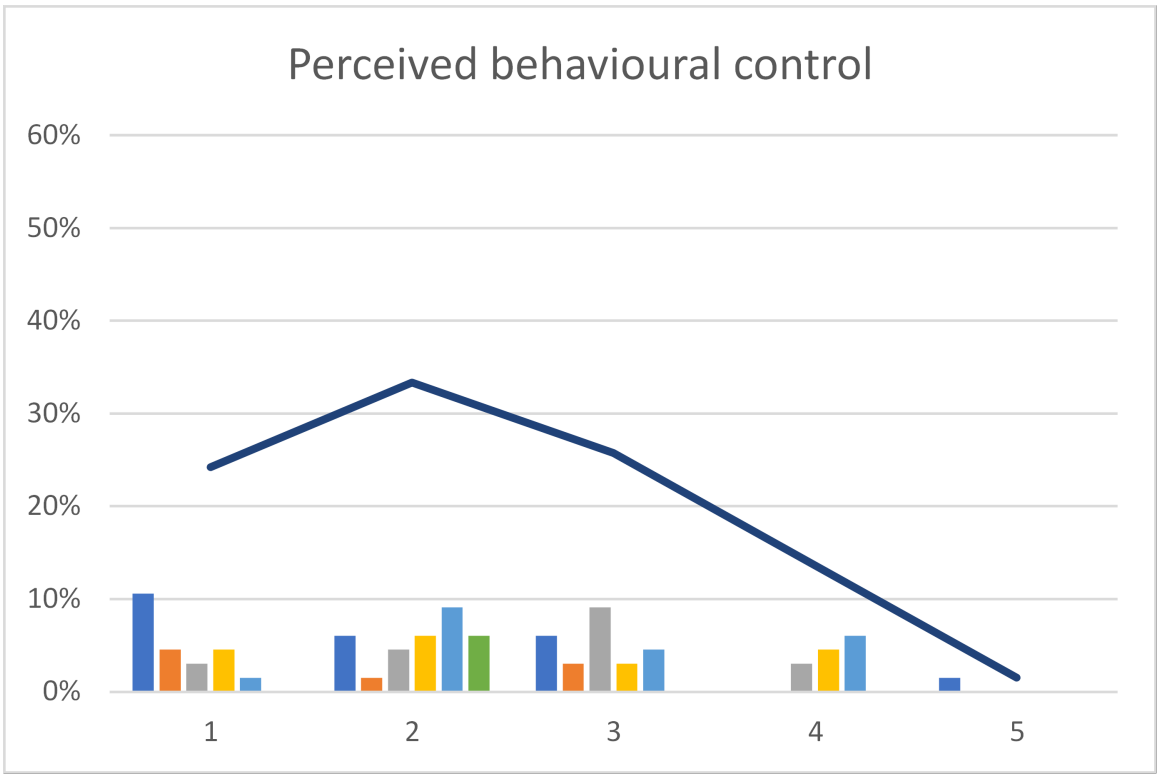

Figure E.9: Perceived behavioural control well-being dashboard per team

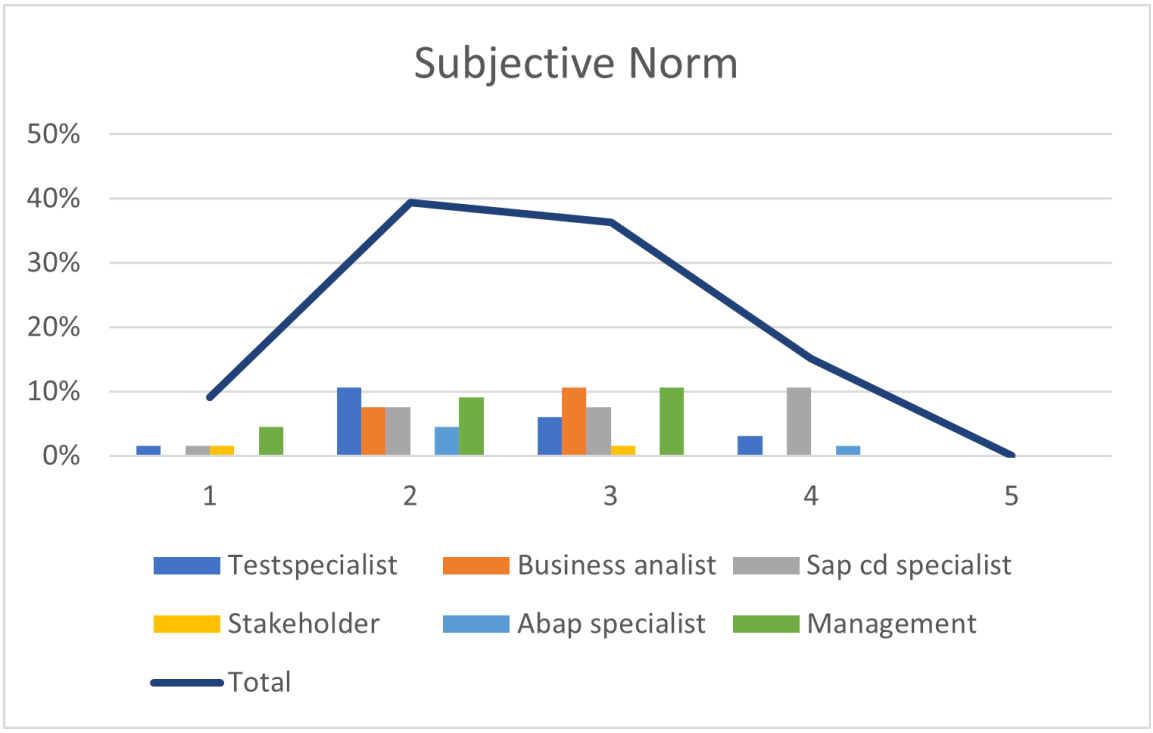

Figure E.10: Subjective norm operational dashboard per function

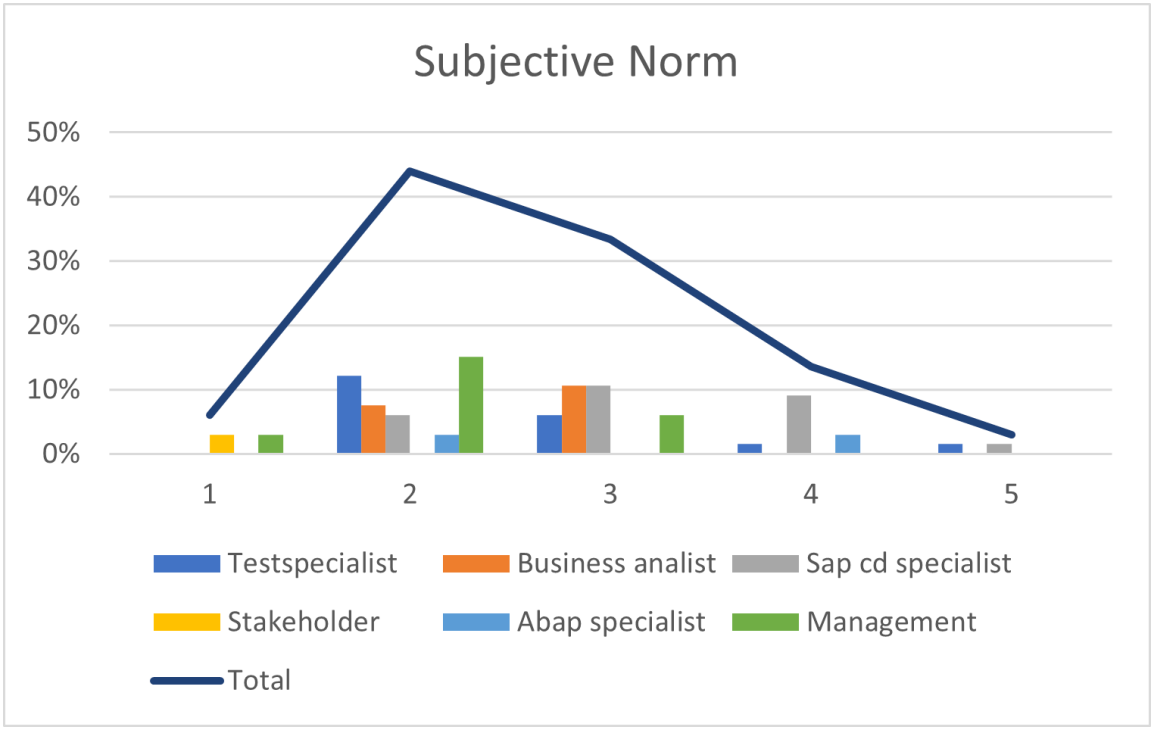

Figure E.11: Subjective norm well-being dashboard per function

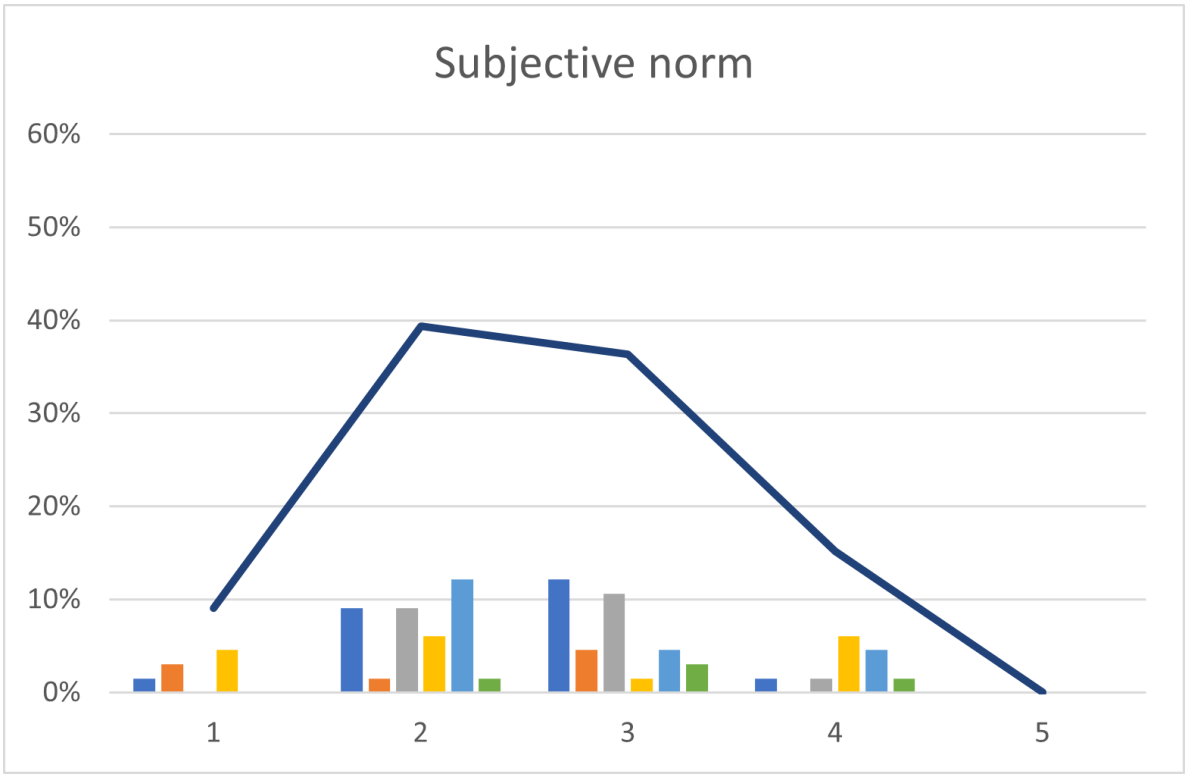

Figure E.12: Subjective norm operational dashboard per team

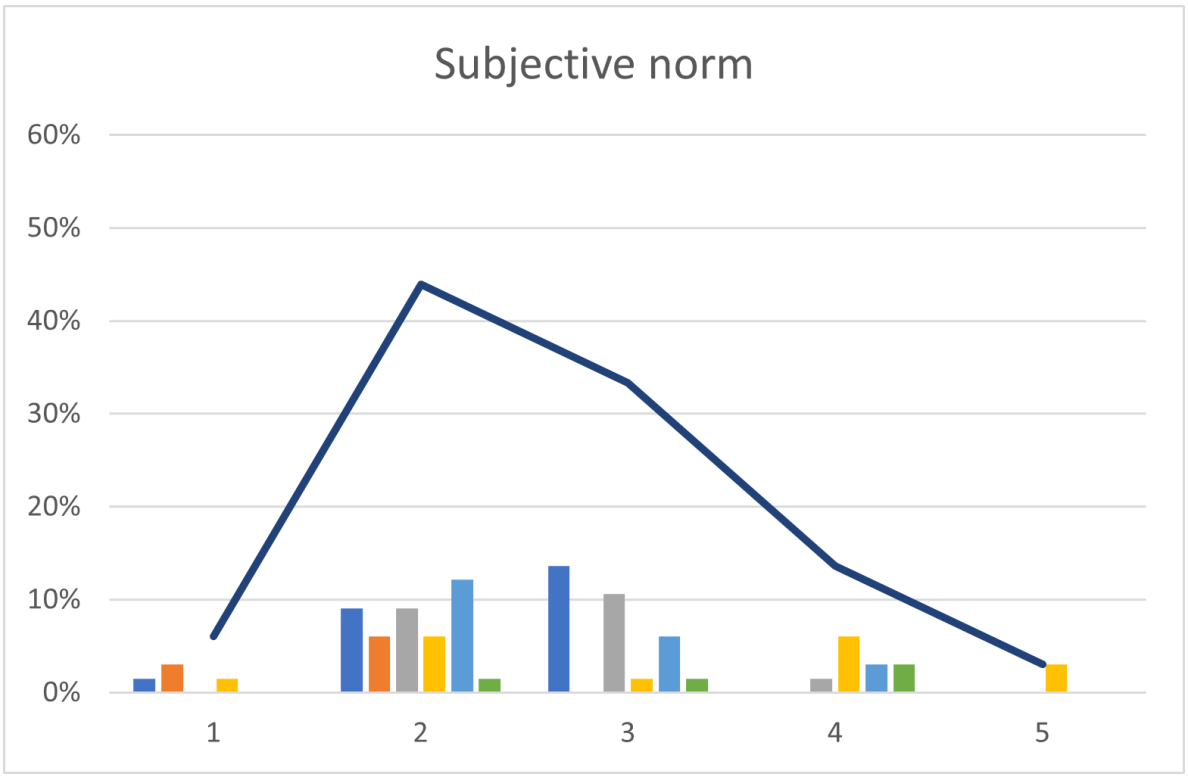

Figure E.13: Subjective norm well-being dashboard per team

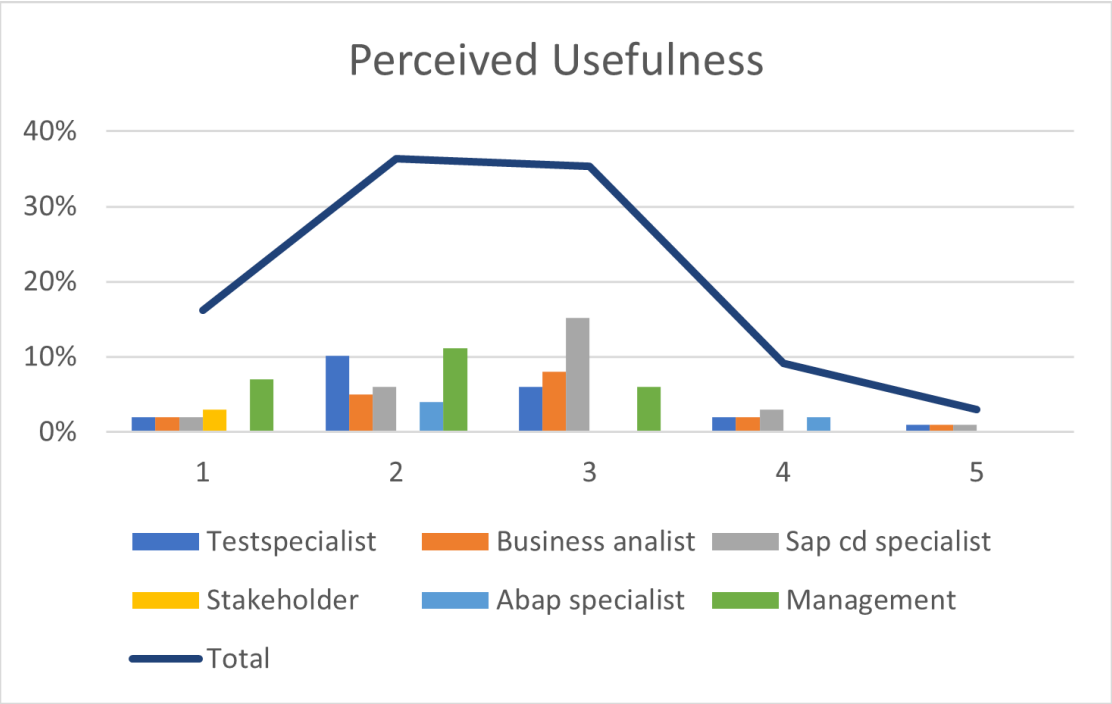

Figure E.14: Perceived usefulness operational dashboard per function

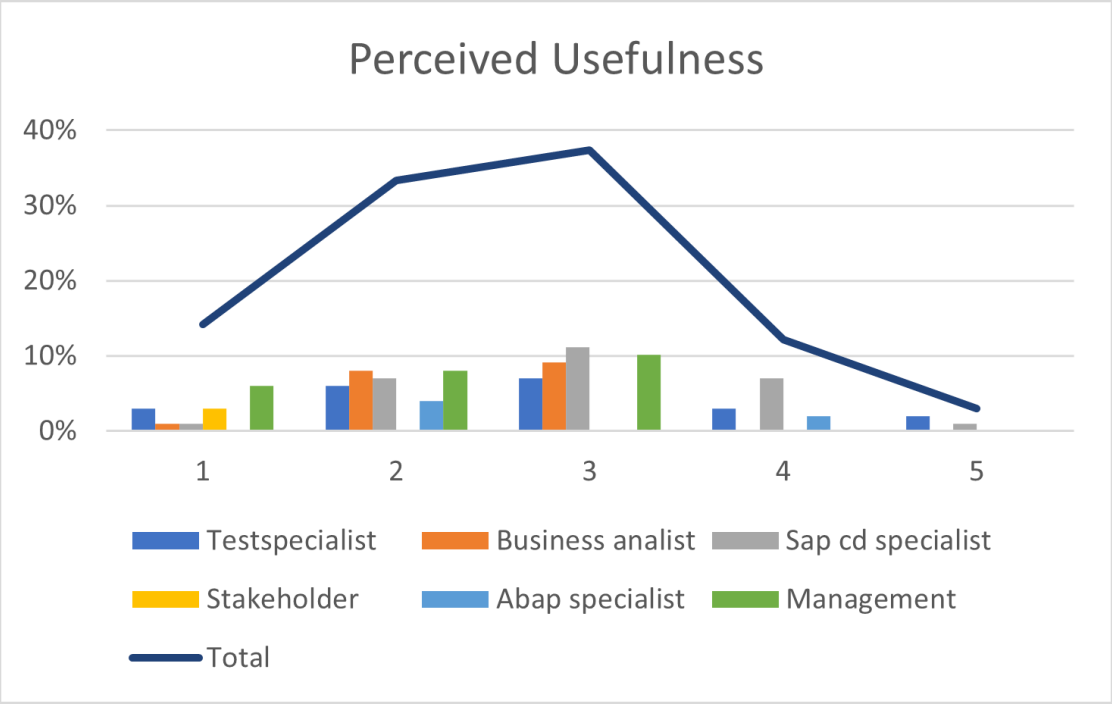

Figure E.15: Perceived usefulness well-being dashboard per function

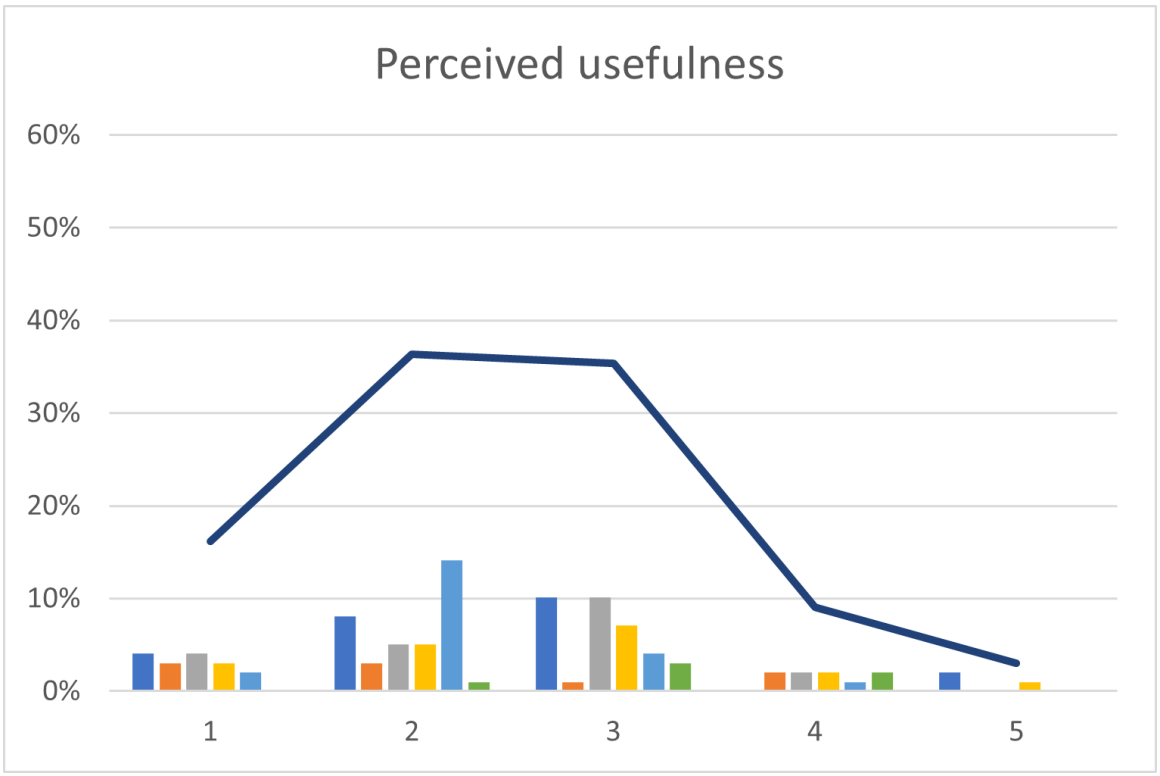

Figure E.16: Perceived usefulness operational dashboard per team

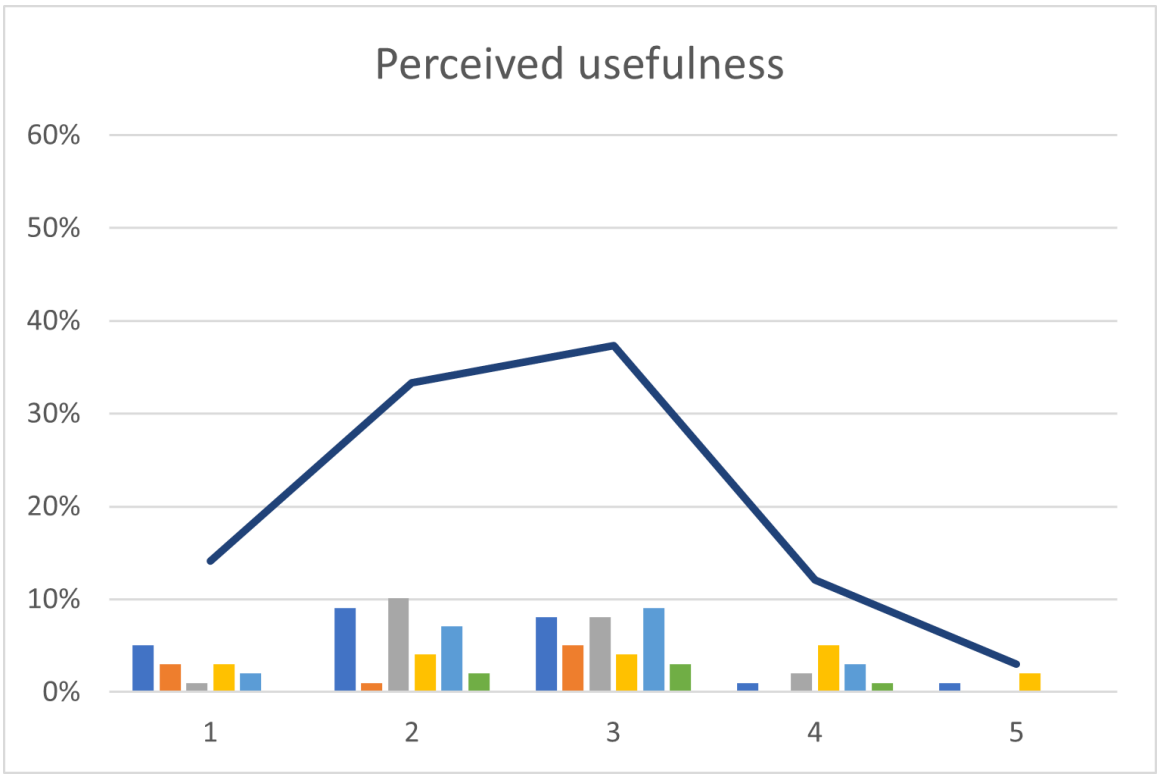

Figure E.17: Perceived usefulness well-being dashboard per team

F Form well-being dashboard

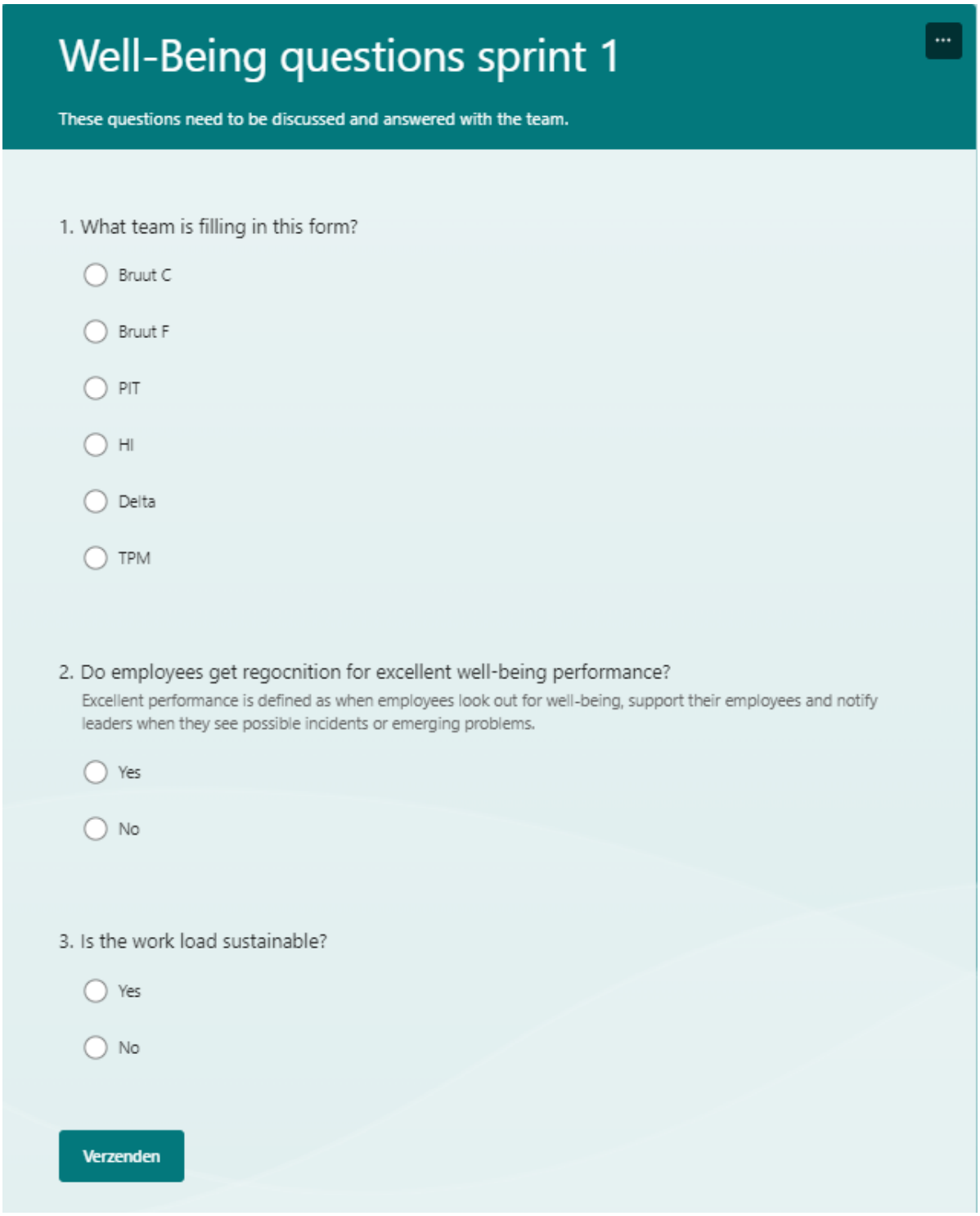

Figure F.18: Form sprint 1