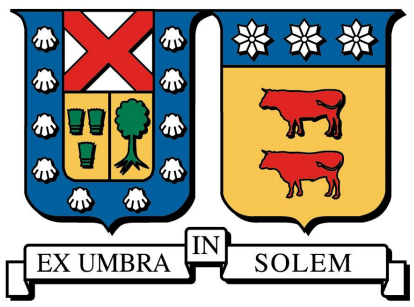
Universidad Técnica Federico Santa María
Departamento de Informática
Valparaíso, Chile



Impact evaluation of experience on efficacy of an architectural design decision-making technique

Juan Pablo Brito Carvajal

Tesis para optar al Grado de
Magíster en Ciencias de la Ingeniería Informática

Profesor Guía: Prof. Dr. Hernán Astudillo

3 de mayo de 2022

Acknowledgments

I want to express my gratitude to the Navy and to the UTFSM, especially to those who contributed to carry out this thesis. To Pabla Valdebenito, who guided me in all administrative aspects; TOESKA group, especially Pablo Cruz, Gastón Márquez and Felipe Osses, who were a great contribution and guide to my work; and to my teachers, especially Dr. Marcello Visconti and my advisor, Dr. Hernán Astudillo, who was a fundamental support to get here. In a separate chapter, I acknowledge the great family sacrifice, especially that of my daughter Isidora, who had to suffer the distance from her father for a long time.

Abstract

The software engineering literature describes several techniques for making software architecture design decisions, among which we find TaSPeR (Tactics Selecion Poker), a technique used for the selection of architectural tactics based on Planing Poker. However, there is a lack of knowledge about the impact of the experience of those who use techniques based on the selection of tactics, on their efficacy. This article addresses this research gap through an experimental study on the impact of the experience of team members who make software architecture design decisions through the selection of architectural tactics, on the efficacy of TaSPeR technique for the correct selection of these tactics. A set of 24 IT professional developers was divided into two "expert" teams and two "novice" teams (of six people each); each team had to solve two scenarios, one using TaSPeR and one using an ad-hoc technique; the scenarios were crossed to mitigate possible learning effects. TaSPeR efficacy metrics were defined as the variation when using the technique (compared to not using it) of its performance in terms of precision, recall and accuracy (selected tactics versus a ground truth); the impact of the teams' experience on the efficacy of the technique was evaluated by comparing the values of the efficacy metrics obtained by the novice teams with those obtained by the experts ones. Initial results suggest that TaSPeR improves the efficacy of novice teams but hurts that of expert ones. This result is quite unexpected and begs replication with even larger populations of IT professionals (no easy task). If the results are confirmed, the question that will rise is: if consensus techniques are so good to estimate, why would they hurt design decision-making by expert teams?

Resumen

La literatura sobre ingeniería de software describe una serie de técnicas para la toma de decisiones de diseño de arquitectura, entre las cuales encontramos TaSPeR ("Tactics Selection Poker"), técnica utilizada para la selección de tácticas arquitecturales basada en "Planing Poker". Sin embargo, falta conocimiento sobre el impacto de la experiencia de quienes utilizan técnicas basadas en la selección de tácticas, sobre su eficacia. Este artículo aborda esta brecha de investigación a través de un estudio experimental sobre el impacto de la experiencia de los miembros de equipos que toman decisiones de diseño de arquitectura de software mediante la selección de tácticas arquitecturales, sobre la eficacia de la técnica TaSPeR para la correcta selección de estas tácticas. Un conjunto de 24 desarrolladores, profesionales de TI, se dividieron en dos equipos de "expertos" y dos de "novatos" (de seis personas cada uno); cada equipo tenía que resolver dos escenarios, uno usando TaSPeR y otro usando una técnica ad-hoc; los escenarios fueron cruzado para mitigar posibles efectos de aprendizaje. Se definieron como métricas de eficacia de TaSPeR la variación al usar la técnica (respecto a no usarla) de sus rendimientos en términos de "precision", "recall" y "accuracy" (tácticas seleccionadas versus una "ground truth"); el impacto de la experiencia de los equipos en la eficacia de la técnica se evaluó comparando los valores de las métricas de eficacia obtenidos por los equipos "novatos" con los obtenidos por los "expertos". Los resultados iniciales sugieren que TaSPeR mejora la eficacia de los equipos "novatos", pero la perjudica en el caso de los "expertos". Este resultado es bastante inesperado y requiere replicación con poblaciones aún más grandes de profesionales de TI (no es una tarea fácil). Si se confirman los resultados, la pregunta que surgirá es: si las técnicas de consenso son tan buenas para estimar, ¿por qué afectarían a la toma de decisiones de equipos expertos?

Contents

List of Tables

List of Figures

Chapter 1

Introduction

THE purpose of this chapter is to contextualize the topic that will be treated in this thesis and explain what this work will be about. Section 1.1 explores the main concepts on which this thesis is based; Section 1.2 defines the research problem; Section 1.3 presents the proposed solution; Section 1.4 sets our research goals; Section 1.5 pose the research questions and explains its rationale; Section 1.6 formulates our hypotheses; Section 1.7 addresses the contribution of this research; Section 1.8 refers to the published work on which this thesis is based; and Section 1.9 details its general structure.

1.1 Background

1.1.1 Design Decisions

A design is an artifact description that is detailed enough for use in implementing that artifact, while a design process is one that aims at devising an appropriate design for that artefact. A design can be considered as appropriate if the artifact described would satisfy requirements while not being unacceptable in other ways. Two major categories of artifacts are physical artifacts, such as buildings, cities and computer hardware, and cognitive artifacts, such as notation systems and software [21], what is that of our interest.

Designing software involves numerous cognitive skills such as mental modeling, mental simulation, problem structuring and decision making [52], the latter being the one on which this thesis will focus.

1.1.2 Software Architecture and architectural decisions

The software architecture of a system is the set of structures needed to reason about the system, which comprise software elements, relations among them, and properties

of both [8]. It is the result of a set of architectural decisions [28], which are crucial to the success of a software-intensive project [22].

Architecting can be viewed as a consensus decision making process that not only seeks the agreement of most stakeholders, but also resolves or mitigates the objections of the minority to achieve the most agreeable solution [23].

1.1.3 Architectural design decision-making techniques

Architectural decisions are a subset of design decisions that are hard to make and costly to change [53]. That is why software architects need a reliable and rigorous process for selecting architectural alternatives and ensuring that the decisions made mitigate risks and maximize profit, for which the literature describes several techniques to choose and analyze architectural alternatives so-called architectural design decision-making techniques. In general, a decision-making technique describes a systematic way of choosing an alternative among a finite number of given ones.[22]

1.1.4 Software Requirements

Software requirements are a specification of what should be implemented. They are a description of how the system should behave, or of a system property or attribute. Also, they may be a constraint in the development process of the system [50].Thus, software requirements encompass three categories: functional requirements, quality attributes requirements and constraints. The first state what the system must do, and how it must behave or react to run time stimuli; the second are qualifications of the functional requirements or of the overall product; and the third are design decisions that have been already made [8].

1.1.5 Quality Attributes Requirements and Architectural Tactics

A quality attribute (QA) is a measurable or testable property of a system that is used to indicate how well the system satisfies the needs of its stakeholders. Quality attribute requirements specify the responses of the system that realize the goals of the business. There are techniques an architect can use to achieve the required quality attributes, called architectural tactics. A tactic is a design decision that influences the achievement of a quality attribute response[8].

Architectural tactics were originally proposed by Bass et al. [8]. Later, Ryoo et al. [40] and then Fernández et al. [24] refined security tactics.

1.1.6 Performance and efficacy definitions

In this work we will use the concept of performance to refer to design decision-making teams and the concept of efficacy to refer to design decision-making techniques.

The performance of a design decision-making team will have to do with how well these teams make decisions in a given scenario that arises, regardless of whether or not they use a technique to make them. In the particular case of teams that make design decisions based on the selection of architectural tactics, their performance will have to do with how correct the selection of these tactics was for a given scenario.

The efficacy of a design decision-making technique, on the other hand, will refer to how useful the technique is for making better decisions, which will be determined by verifying how much the performance of teams improves (or decreases, if it turns out to be inefficacious) in decision-making when using the technique with compared to that obtained without using it.

1.2 Problem definition

As previously stated, it is very difficult to make architectural design decisions and very expensive to modify them later, which generates the need to have techniques for making software architecture design decisions that allow to carry out reliable and rigorous processes for select architectural alternatives.

On the other hand, there is evidence to suggest that experienced designers have the perspective and knowledge to find the right problems and options, and thus be able to achieve with mere experience what inexperienced designers achieve through an established process [45, 15] that provide guidelines, such as that provided by decision-making techniques.

Taking the above into account, it could be conjectured that design decision-making techniques efficacy and, more specifically, software architecture design decision-making techniques efficacy could be greater for those who have less experience and lower or even null for those who have enough experience and, thus, could not need to use these guidelines. Although there are some works in the literature that explore and provide evidence in favor of this conjecture, as far as we know, none of them do so in particular on techniques for making design decisions through the selection of architectural tactics, which leads us to pose the following research problem:

Main research problem

There is a lack of knowledge about the impact of professional experience on the efficacy of techniques for making design decisions through the selection of architectural tactics.

1.3 Proposed solution

Proposed solution

Design and execution of an experimental study that allows evaluating the impact of the experience of members of teams of IT professionals who make software architecture design decisions through a technique based on the selection of architectural tactics, on the efficacy of this technique .

We tackled the research problem described above through the design and execution of an experimental study that will help fill this research gap. This study made it possible to evaluate the impact of the experience of software architecture design decision-making teams members on the efficacy of a technique that would fulfill this purpose (architectural design decision-making). The selected technique was TaSPeR (Tactics Selection Poker), which is a card game-based and consensus-building architectural decision making technique (based on Planning Poker) presented in Osses et al. [38]. We chose this technique because it had recently been proposed by our research group, called Toeska, which allowed us, in addition to contributing to the coverage of the detected research gap, also obtaining additional knowledge about this new technique.

TaSPeR technique allows development team members to identify, argue for, and choose among architectural security tactics according to objectives and priorities. In their work, Osses et al. assess TaSPeR fitness to be used in the selection of security tactics through several empirical tasks and activities. However, the impact of the experience of those who make up the teams on the efficacy of this technique is not explored, being the groups that were formed for the validation of this technique balanced in terms of the experience of its members.

The proposed experimental design was materialized in a study in which IT professionals belonging to a Universidad Técnica Federico Santa María master program participated. It considers the evaluation of the performance achieved by teams with and without the use of TaSPeR in terms of the same three metrics used by Osses et al. for the validation of the technique: precision, recall and accuracy with respect to a ground truth defined by a group of experts for scenarios that were raised. These metrics will be defined and explained later in this work.

1.4 Research goals

General research objective

Contribute to determining the impact of experience on the efficacy of software architecture design decision-making techniques based on the selection of architectural tactics.

The above materialized through an experimental study in which the impact of experience on the efficacy of software architecture design decision-making techniques based on the selection of architectural tactics was evaluated in particular in the TaSPeR technique.

When we talk about determining the impact of experience on the efficacy of this kind of techniques, we refer to evaluating whether the decision-making teams members experience influences, either positively or negatively, how much this technique improves decision making, without considering whether there is any impact on the amount of resources that should be used for it, such as time or other; that is, the impact on the efficiency of the technique will not be evaluated.

For the fulfillment of the general objective, two specific objectives were set:

1. Determine the impact of IT professionals' experience on the efficacy of TaSPeR for software architecture design decision-making.

2. Contribute to the state of the art with the characterization of the factors that could explain the results observed in the experimental study.

1.5 Research questions

General research question

How does the experience of IT professionals using a software architecture design decision-making technique through the selection of security tactics affect the efficacy of that technique?

This research question seeks to investigate the impact, whether positive, negative or neutral, that the experience of IT proffesionals could have when they make architectural design decisions using a technique devised for that purpose, compared to what would happen if they made the same decisions in a natural way.

Given the general research question of this thesis, two research sub-questions will be posed regarding the phenomenon studied.

The first one seeks to collect evidence of the aforementioned impact, through an experimental study in which the TaSPeR technique is used by students of a Master's

Degree in IT, for the solution of specific architectural problems that are raised. To measure this impact, the three metrics mentioned in Section 1.3 are used.

RQ1

How does the experience of IT professionals using the TaSPeR technique affect the efficacy of the technique for precision, recall and accuracy in selecting software architecture security tactics?

Rationale:

The precision in the selection of tactics gives the proportion of correctly selected tactics out of all those that were selected. This metric is not affected by the tactics that were omitted (false negatives), since it only measures the level of success among the selected tactics, without taking into account those that were not.

The recall in the selection of tactics, instead, gives the proportion of tactics that were selected out of all those that should have been selected. Unlike precision, recall is negatively affected (decreased) by omissions (false negatives), but not by incorrectly selected tactics (false positives), since it only takes into account the level of success among the tactics that should be selected, without considering those that should not be.

Accuracy gives the proportion of decisions on whether or not to select a tactic that were correct from the total of existing tactics. Therefore, it is negatively affected (decreased) by both incorrectly selected tactics (false positives) and omissions (false negatives), providing a value that integrates both what it does not take into account precision and what it does not take into account the recall.

That is, each metric allows the evaluation of the efficacy of TaSPeR from a different perspective, with the precision affected by false positives, recall by false negatives and accuracy by both at the same time. In this way, between the three they complement each other to have a complete vision of how correct the design decisions were made regarding the selection or not of tactics according to a ground truth.

RQ2

What could be possible explanations for the results obtained in the experimental study carried out to evaluate how the experience of IT professionals using the TaSPeR technique affects the efficacy of the use of the technique?

Rationale:

Once the evidence provided by the experiment is obtained, it is important to propose possible explanations for the observed behavior, for which the existing literature must be reviewed, looking for theories that could explain the phenomena that are observed.

1.6 Hypotheses

General hypothesis

The TaSPeR technique is more efficacious in selecting software architecture security tactics when used by less experienced IT professionals than when used by more experienced IT professionals.

From this general hypothesis, three sub-hypotheses were derived, each of them related to one of the three metrics used for the evaluation of the general hypothesis through experimentation.

Sub-hypothesis 1

The efficacy of the TaSPeR technique for precision in selecting software architecture security tactics is greater when used by less experienced IT professionals than when used by more experienced IT professionals.

Sub-hypothesis 2

The efficacy of the TaSPeR technique for recall in selecting software architecture security tactics is greater when used by less experienced IT professionals than when used by more experienced IT professionals.

Sub-hypothesis 3

The efficacy of the TaSPeR technique for accuracy in selecting software architecture security tactics is greater when used by less experienced IT professionals than when used by more experienced IT professionals.

1.7 Contribution

The contribution of this work is two-fold. On the one hand, it allowed collecting evidence on the impact that the experience of those who use the TaSPeR technique could have on its efficacy, helping to reduce the research gap that motivated the realization of this work; on the other hand, it proposes an experimental design that seems to be appropriate for its purpose and that can be replicated in other software architecture design decision-making techniques to assess whether effects similar to those evidenced in TaSPeR occur in them and, in this way, verify whether they are generalizable, which will allow to continue reducing the research gap detected in the future.

1.8 Published work

Based on the work carried out in this thesis, the following research articles have been published:

- Directly related to this thesis:
  Juan P. Brito, Felipe Beroíza, Gastón Márquez, Marcello Visconti, Hernán Astudillo: Evaluating the Impact of Experience in Architectural Design Decision-Making Techniques: An Experimental Design. JCC 2019: 1-8 Concepción, Chile.

Figure 1.1: Jornadas Chilenas de Computación 2019.

- Related to software architecture and technical debt:
  Boris Pérez, Juan P. Brito, Hernán Astudillo, Darío Correal, Nicolli Rios, Rodrigo Oliveira Spínola, Manoel Mendonça, Carolyn Seaman: Familiarity, Causes and Reactions of Software Practitioners to the Presence of Technical Debt: A Replicated Study in the Chilean Software Industry. JCC 2019: 1-7 Concepción, Chile.

1.9 General structure

This thesis is made up of six chapters, which are structured as follows (see figure 1.2: Chapter 1 explores the main concepts on which this thesis is based, and presents the research problem, research goal, research questions and contributions of this work; Chapter 2 explores state of the art; Chapter 3 presents a proposal that contributes

to the solution of the problem raised, reducing the research gap detected; Chapter 4 describes the experimental study carried out to answer research question 1 and presents its results; In Chapter 5, the results obtained are discussed and possible explanations for them are proposed; Chapter 6 presents the conclusions and proposes future work to continue delving into this line of research.

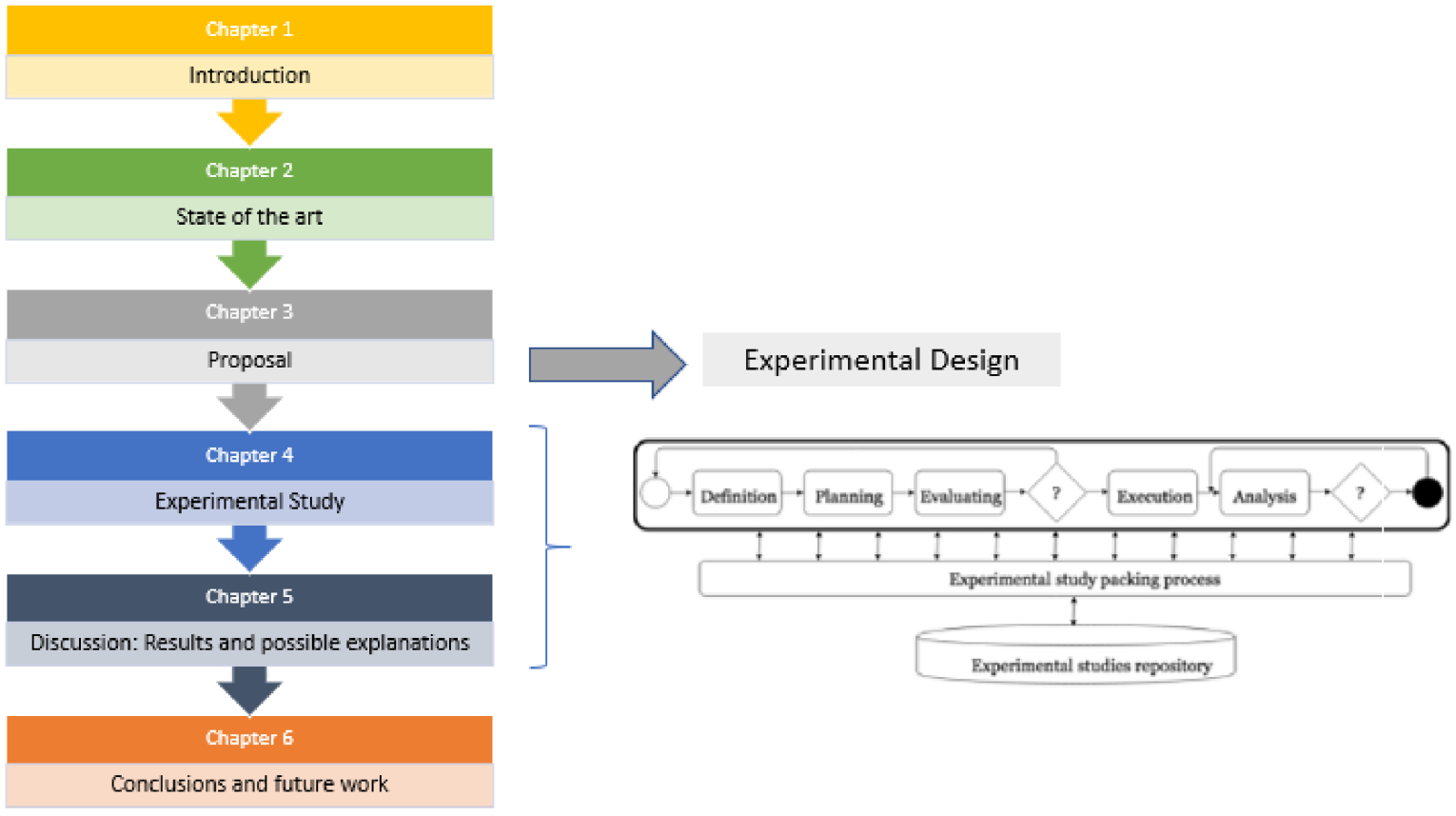


Figure 1.2: Thesis Structure.

Chapter 2

State of the art

In this chapter, a review of the state of the art will be carried out, first, of the main existing software architecture design decision-making techniques and, then, of the works that, like this, have focused some way in the performance analysis of people related to the IT area (whether students or professionals) using software architecture design decision-making techniques. Section 2.1 shows some of the major papers presenting software architecture design decision-making techniques; section 2.2 presents some works which analyzes experience impact in various software engineering techniques; and finally, section 2.3 summarizes this chapter.

2.1 Architectural design decision-making techniques

The software architecture design process focuses on deriving an architecture from software requirements. It includes the application of a decision-making technique, which focuses on the issue of selecting among a number of alternatives for achieving the system's desired quality attributes. There are several techniques for choosing among architecture alternatives [22] that have been proposed over time, of which we present now some of the most important.

In 1999, Chung et al. [13] outlined an approach that formulates architectural properties such as modifiability and performance as "softgoals" which are incrementally refined. This approach uses tradeoffs and architectural decisions are traced to stakeholders and their dependency relationships.

In 2000, Kazman et al. [30] presented ATAM (Architecture Tradeoff Analysis Method), a technique for analyzing software architectures that had been developed and refined in practice over the past three years that not only reveals how well an architecture satisfies particular quality goals, but it also provides insight into how those quality goals interact with each other.

In 2001, Kazman et al. [29] proposed an architecture-centric approach to the economic modeling of software design decision making called CBAM (Cost Benefit Analysis Method), in which costs and benefits are traded off with system quality

attributes.

In 2003, Moore et al. [35] described the practical difficulties and experiences in applying the method to a large real-world system, on the basis of which developed a new version of CBAM, called CBAM 2. Unlike the former, the latter is a iterative procedure where the understanding of the system's benefits is refined via the elicitation of progressively more information.The same year, Svahnberg et al. [43] propose a quality-driven support method for identifying software architecture candidates. It allows to aid in the understanding of different architecture candidates for a software system and creates a support framework, using a multi-criteria decision method, supporting comparison of different software architecture candidates for a specific software quality attribute and vice versa, and then uses this support framework to reach a consensus on the benefits and liabilities of the different software architecture candidates and to increase the confidence in the resulting architecture decision.

In 2005, several proposals stand out. Andrews et al. [6] presented a framework for evaluating design choices with respect to meeting competing requirements. Specifically, they develop a model to estimate the performance of a UML design subject to changing levels of security and fault-tolerance. This analysis provides a way to identify design solutions that are unfeasible and then apply multi-criteria decision-making techniques to evaluate the remaining feasible alternatives. Al-Naeem et al. [3] proposed a quantitative quality-driven approach that attempts to find the best possible fit between conflicting stakeholders' quality goals, competing architectural concerns, and project constraints which uses optimization techniques to recommend the optimal candidate architecture. Gilb et al. [26] presented the Impact Estimation Method, whose purpose was to help answer the question of how design ideas impact all a system's critical performance attributes (such as usability and reliability) and all its resource budgets (such as the financial cost and staff headcount) for implementation and operational running. It can be used for a wide variety of project purposes, with the comparison of alternative design ideas to determine which is the best one among its most important uses.

In 2006, Dabous et al. [17] proposed a framework for evaluating alternative architectures and its application to financial business processes. In their work, they adopt the view that for a particular problem context, the architectural design process can be considered as a series of choices regarding the application of a number of architectural design strategies. The problem context they described is common to a category of e-business applications that arise from the e-finance domain. Given a formal representation of this context, they identify and formalise a number of applicable design strategies and show the resulting architectures.

In 2007, Wallin et al. [48] presented a method for making decisions on integration strategy for in-vehicle automotive systems, which is a combination of the Architectural Tradeoff Analysis Method, ATAM, and the Analytical Hierarchy Process, AHP.

In 2008, Stoll et al. [42] proposed the Influencing Factors (IF) method, which

guides the architect through stakeholders' concerns to architectural decisions in line with current business goals.

In 2010, Vijayalakshmi et al. [47] developed a quantitative evaluation method based on MCDA (Multicriteria Decision Analysis) methods and Multicriteria fuzzy decision making technique, that also models the variation in preference according to changes in the architecture structure, which avoids the necessity to repeat the entire evaluation process.

In 2011, Gilson et al. [27] introduced a tracing mechanism for architectural rationale, design decisions and design alternatives as part of a transformation oriented design method for distributed systems that intertwines architecturally-significant requirement modelling and architecture modelling.

In 2012, van Heesch et al. [46] introduced the decision forces viewpoint as an extension to their framework for documenting architecture decisions. It was validated in a multiple-case study, demonstrating its ability to support inexperienced software engineers during decision making process, by providing a structure that triggers them to consider multiple architectural decision alternatives and systematically compare them in the context of all important forces.

In 2014, Chavarriaga et al. [11] presented a process that exploits models of architectural tactics and their corresponding design alternatives to determine which options to use in order to implement a combination of tactics when developing and deploying applications in the cloud. In the same year, Pedraza-Garc'ia et al. [39] presented a methodological approach to address and specify the quality attribute of security in architecture design applying security tactics.

In 2015, Chavarriaga et al. [12] presented an approach based on feature models to help manage trade-offs. It is based on the specification of relationships between architectural tactics and design alternatives that describe, for each tactic, which combination of designs can be used or must not be used. In the same year, Kim et al. [31] proposed a quantitative approach to choosing security architectural tactics using architectural tactic knowledge base, in that a cost of an architectural tactic is estimated.

In 2016, Schriek et al. [41] proposed a simple card game to help novice designers use design reasoning for which several common reasoning techniques were chosen to be represented by the card game (problem structuring, option generation, constraint analysis, risk analysis, trade-off analysis, and assumption analysis).

In 2017, Lopes et al. [32] proposed a combination of concepts from two architecture definition methods, CEADA (Collaborative Evaluation of Enterprise Architecture Design Alternatives) [36] and CBAM 2 (Cost Benefit Analysis Method) [35] into a single approach that can be used in agile projects and addresses the most critical concerns of group decision-making.

In 2018, Alashqar et al. [4] presented a framework based on fuzzy measures using Choquet Integral approach which takes into account the impact of architectural tactics on quality attributes, the preferences of quality attributes and the interac-

tions between them. In the same year, Osses et al. [38] proposed TaSPeR, a card game-based technique and consensus-building technique (based on Planning Poker) that allows development team members to identify, argue for, and choose among architectural security tactics according to objectives and priorities.

Boer et al [20] presented a card game to teach how to make design decisions for real software projects, taking into account the concerns of stakeholders and, where appropriate, understand how design decisions are influenced when there are unexpected changes in the environment.

In general, it can be seen that there is a large number of works that propose techniques that seek to support decision-making in software architecture design, with a marked trend in the last decade to works that base the above on architectural tactics and gamification, among which the one by Osses et al. [38] incorporates both.

2.2 Analyzes of experience impact in techniques

There are some studies that analyze the main drivers in software architecture [19] and have concluded that it is mainly the experience and the intuition of the architects that guides the design decision making, rather than a specific tool or method. The above speaks about the importance of knowledge management regarding design decision-making, and this concern is only confirmed with other studies [7] that seek methods and tools to capture knowledge.

In the software engineering literature there are several studies that, like ours, address the impact of the experience of those who use a technique, on the technique.

In 2003, Carver et al. [9] conducted an experimental study to evaluate the impact of background and experience on software inspections. One of his hypotheses was that inspectors who had more experience with the inspection process would find more defects than those who have less experience. This study examined observation of an inspection as a method for gaining inspection experience. The results of showed that based on qualitative data, the subjects believed that the observation was beneficial both to their understanding of the process as well as to their effectiveness in using the process. The quantitative data showed this improvement in only some cases. Furthermore, it appeared that for less experienced inspectors to gain any benefit from observing an inspection, that inspection must be focused on a requirements document from a domain in which the inspector has high domain knowledge. Like ours, this study evaluates through an experiment the impact of experience on the efficacy of a technique that provides guidelines, in this case for software inspections; however, it only compares the performance obtained, in terms of the number of defects found, when used by experts versus when used by novices, without taking into account the performance obtained without using the technique and how they vary when using it, unlike our work. It could be that without using the technique and the guidelines it provides, the performance of novices would be lower than that of experts, implying that the technique is more efficient when used by novices than

by experts. Additionally, unlike our work, the technique on which experience impact is evaluated supports software inspections, not software architecture design decision making.

In 2008, Carver et al. [10] made a large-scale controlled inspection experiment with over 70 professionals that focused on the relationship between an inspector's background and his effectiveness during a requirements inspection. The results showed that inspectors with degrees in non-computer-related majors (i.e., engineering, math, science, or business) found significantly more defects during a requirements inspection than inspectors with degrees in computer-related majors (i.e., computer science, software engineering, electrical engineering, or management information systems). In addition, those with experience writing requirements were significantly more effective than those without such experience. In other dimensions that were also analyzed, they observed that level of education (Masters, PhD), prior industrial experience, or other job-related experiences did not significantly impact the effectiveness of an inspector. Unlike ours, this study evaluates through an experiment the impact mainly of background and not of experience (although it also does so secondarily) on the effectiveness (similar to efficacy) of a technique, also in this case for software inspections; however, as in the previous case, it only compares the performance obtained, in terms of the number of defects found, when used by experts versus when used by novices, without taking into account the performance obtained without using the technique and how they vary when using it, unlike our work. As in the previous work, it could be that without using the technique and the guidelines it provides, the performance of those with a lower level of education, previous industrial experience, or other work-related experience (dimensions that did not significantly impact the effectiveness of an inspector when they used the technique) , would be different from those with a higher level of education, level of education, previous industrial experience or other work-related experiences, which would imply that there would be an impact on the effectiveness of the technique in the face of differences in these dimensions. As in the previous work, additionally, unlike our work, the technique on which the impact of background and the other dimensions is evaluated, supports software inspections, not software architecture design decisions.

Also in 2008, Tang et al. [45] carried out an experiment which goal was to investigate if there was any quality improvement to software design when design reasoning was applied. The experiment involved twenty designers, whom were asked to design a user interface and their designs were scored and compared. The results showed that the test group that was equipped with design reasoning produced in average a higher quality design than the control group. However, the authors noted that only those with less than 5 years of experience improved their design by using design reasoning, while for those with more than 5 years of experience no major change was noticed. As in our work, this also evaluates the impact of the experience of those who use a software architecture design decision-making technique (a design reasoning aproach), on the efficacy of the technique, measured as the improvement

in the score obtained by using design reasoning. However, the above is analyzed only superficially, since it is not the main objective of the study, but a secondary one, unlike our work, where it is the main objective. In addition, in this study, the impact of experience on the efficacy of a design decision-making technique based on design reasoning is evaluated, not on the selection of tactics through gamification, as in ours.

In 2018, Tang et al. [44] carried out an experiment to test if a reminder card approach based on design reasoning could improve designers' reasoning and improve the amount of design rationale they could find. One of the dimensions that was analyzed in this experiment was the difference between the results of the students and those of the professionals who participated in this experiment. As a result, they observed that the reminder card approach improved reasoning carried out by both students and professionals, and generally enable them to provide more design rationale; that is, they did not observe differences between students and professionals. Unlike the other Tang et al. [45] study mentioned above where they found that reminders can help novice but not professionals, this study showed that it could also benefit professionals as well as novice. The analysis of what this work addresses with respect to ours is similar to that analyzed in the previous paragraph, although in this case it also coincides with ours in the use of gamification, but neither does it address the problem of our work as the main objective or with respect to software architecture design decision making techniques based on the selection of architectural tactics. In addition, the performance metric based on which the efficacy of the technique was evaluated in this case was the quality of design reasoning and the quantity of design rationale.

In 2016, Schriek et al. [41] conducted an experimental study to test on groups of students the card game they were developed to help design reasoning, resulting in noticeable differences between the control and test groups. Those who used the cards produced better design arguments: the groups with the card game on average perform 75% more reasoning than the control groups. Although this study coincides with ours in that a software architecture design decision-making technique based on gamification through cards was used, it differs in that the efficacy of the proposed technique was only evaluated in novices, without contrasting with that obtained by experts. In addition, the performance metric on which the technique efficacy metric was based was the quality of the design arguments and the amount of reasoning used.

To the best of our knowledge, there is no work in the literature that addresses the impact of experience on those using software architecture design decision-making techniques based on the selection of tactics, on the efficacy of the technique. The experimental design that is proposed and the experimental study that is carried out in this work contributes to fill this research gap.

2.3 Summary

In general, it can be seen that there is a large number of works that propose techniques that seek to support decision-making in software architecture design, with a marked trend in the last decade to works that base the above on architectural tactics and gamification, among which the one by Osses et al. [38] incorporates both.

There are some studies that analyze the main drivers in software architecture [19] and have concluded that it is mainly the experience and the intuition of the architects that guides the design decision making, rather than a specific tool or method. In the software engineering literature there are several studies that, like ours, address the impact of the experience of those who use a technique, on the technique, but to the best of our knowledge, there is no work that addresses the impact of experience on those using software architecture design decision-making techniques based on the selection of tactics, on the efficacy of the technique. The experimental design that is proposed and the experimental study that is carried out in this work contributes to fill this research gap.

Chapter 3

Proposal

This chapter aims to present the proposed experimental design, previously explaining the background on which it is based. Section 3.1 introduces to the proposal of this work proposal; section 3.2 provides the TaSPeR technique background; section 3.3 presents the experimental design; and finally, section 3.4 summarizes this chapter.

3.1 Introduction

The Toeska research group explored the line of investigation related to architectural design decision-making through the completion of a series of works mainly related to architectural tactics [24, 25, 37, 33, 34], among which is the TaSPeR proposal [38], a new architectural design decision-making technique based on the consensual selection of architectural tactics . However, for 2 of the 3 metrics defined to evaluate the efficacy of the technique, the experimental study carried out for its validation failed to reject the null hypothesis formulated and in some cases, it even yielded results contrary to those expected.

The work teams used for decision-making by consensus in this experiment were balanced in terms of their level of experience; on the other hand, a series of previous works shed light on the effects that experience could have on the performance of architectural design decision-making techniques. This made us wonder if, in particular, in the case of TaSPeR, the experience of those who used it could influence its efficacy. The foregoing led us to the need to design and execute an experimental study that would allow us to evaluate the impact of experience on the efficacy of architectural design decision making using TaSPeR. The proposal of this thesis is the design and execution of this experimental study.

3.2 TaSPeR technique background

TaSPeR is a technique that supports tactics selection (specifically security tactics) to make robust design decisions from the project beginning, and during its evolution. Builded on Planning Poker [14], in TaSPeR participants share preferences among architectural tactics. Each participant prioritizes tactics from a standard set to satisfy the project goals, and shares them with other participants to converge on the most appropriate security tactics. More details about this technique can be found in [38].

3.2.1 Security tactics used

Tactics taxonomy used in TaSPeR is based in the Fernández et al. [24] security tactics. It was defined by experts belonging to the Toeska Research Group and consists of 17 tactics grouped into four categories: detect attacks, stop or mitigate attacks, react to attacks, and recover from attacks, according to the following detail (see figure 3.1):

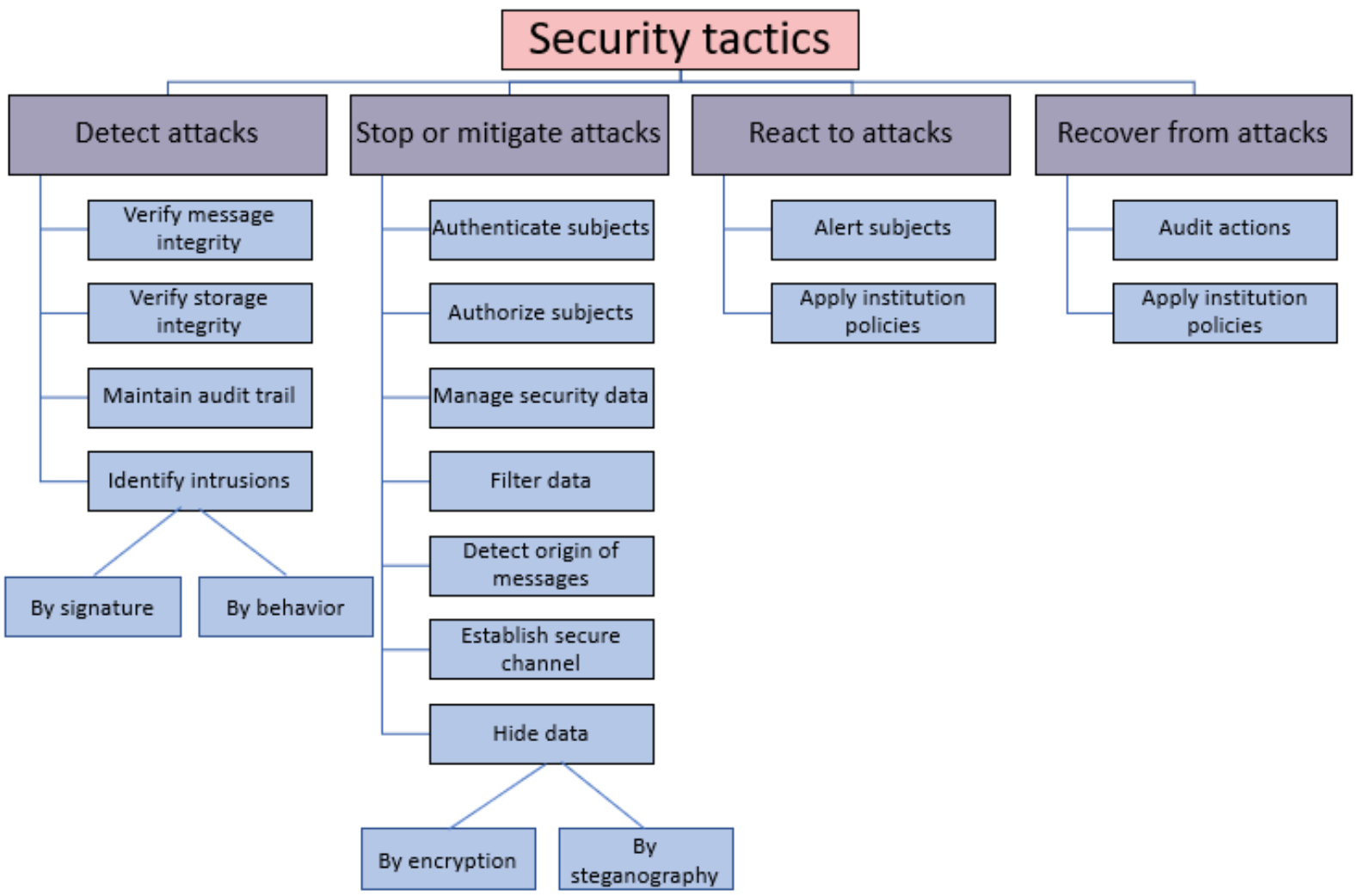


Figure 3.1: Security Tactics used in TaSPeR technique [24, 38]

1. Detect attacks tactics:

- Verify message integrity: This tactic employs techniques such as checksums or hash values to verify the integrity of messages, resource files, deployment files, and configuration files [8].
- Verify storage integrity: This tactic was added by Fernández et al. [24] to Bass detect attacks tactics in order to define measures to make sure that databases have not been modified.
- Mantain audit trail: This tactics was considered in Bass et al. [8] exclusively as a recover from attacks tactic consisting in keep a record of user and system actions and their effects to help trace the actions of, and to identify, an attacker. In their review, Fernández et. al considered this tactic as a system function that can be used not only to recover from an attack, but also to detect it.
- Identify intrusions: This tactic was not considered by Bass. It was added by Fernández et al. [24], being divided into "by signature" and "by behavior", the 2 standard ways to apply intrusion detection.

2. Stop or mitigate attacks tactics are the following:

   - Authenticate subjects: Bass proposed the tactic "authenticate actors", meaning by authentication to ensure that an actor (a user or a remote computer) is actually who or what it purports to be. Passwords, one-time passwords, digital certificates, and biometric identification provide a means for authentication [8]. Fernández et al. [24] changed the word “actors” to “subjects”, since according to standard security terminology, subject is an active entity that can request resources and includes humans and executing processes.
   - Authorize subjects: Bass proposed the tactic "authorize actors", meaning by authorization to ensure that an authenticated actor has the rights to access and modify either data or services [8]. As in the previous tactic, Fernández et al. [24] changed the word "actors" to "subjects".
   - Manage security data: This tactic was not considered by Bass. It was added by Fernández et. al [24]. Includes the management of keys for cryptography, the secure storage of authorization rules, and other ways to handle security information.
   - Filter data: This tactic was not considered by either Bass or Fernandez. It was incorporated into the set of tactics to be used in the TaSPeR technique by the experts of the Toeska group during its definition process and aims to avoid attacks based on abnormal inputs or from untrusted sources. The above is achieved implementing content filters on the data received through user requests to the system.

- Detect origin of messages: This tactic was not included by Bass. It was added by Fernández et. al [24] for having been considered a safety aspect not contemplated in Bass tactics.
- Establish secure channel: This tactic was not considered by Bass either. It was added by Fernández et. al [24] because they felt it was needed to provide secure communications in a distributed system.
- Hide data: Bass proposed the tactic "encrypt data". Fernández et al. [24] changed it to "hide data" with two varieties: “use cryptography” and “use steganography”.

3. React to attacks tactics are the following:

   - Alert subjects: Bass proposed the tactic "inform actors", based on the fact that the set of relevant actors ("subjects" according to the Fernández et al. [24] work), such as operators, other personal or cooperating systems, must be notified when the system has detected an attack, as some action might be required from some of them [8]. Fernández et al. [24] changed the name of this tactic to "alert subjects".
   - Apply institution policies: This tactic was not considered by Bass. It was added by Fernández et al. [24] considering that the specific functions depend on institution policies, should be performed by the system, and it does not make sense to define general functions. Policies are high-level guidelines defining how an institution conducts its activities in its business, professional, economic, social, and legal environment. Institution security policies include laws, rules, and practices that regulate how an institution uses, manages and protect resources.

4. Recover from attacks

   - Audit actions: The name of the tactic "mantain audit trail" introduced by Bass, was changed in this category by Fernández et al. [24] to "audit actions".
   - Apply institution policies: As in "react to attacks" category, this tactic was also included in this category by Fernández et al. [24].

3.2.2 The technique

As in Planning Poker, cards are also used in TaSPeR. Each of this cards has the following fields (see figure 3.2):

- Number: Identify the tactic number.
- Quality attribute: Quality attribute requirement related to the tactic

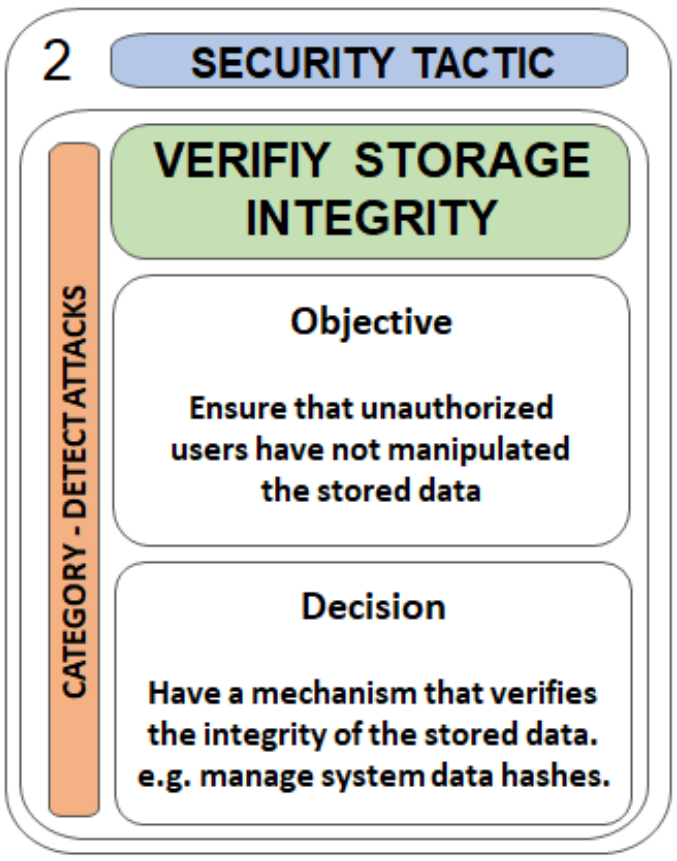


Figure 3.2: TaSPeR Card.

- Name: Indicates the tactic name.
- Objective: Describes the incentive to use the corresponding tactic.
- Decision: Gives a description of the response associated with the corresponding tactic.
- Category: Identifies the taxonomy category of the tactic.

TaSPeR is carried out in 3 steps:

1. **Discussion**
   - One of the subjects assumes as the moderator and takes control of the meeting.
   - The moderator is the one who guides to achieve the consensual decision of security tactics.
   - The first moderator task is to present the context and scenario to be discussed, and distribute the security tactics cards.
   - Each participant receives a deck of security tactics cards.
   - The context, goals and scenario are discussed among all participants.
2. **Tactics choice**

- Each participant privately selects the most appropriate card(s) for the situation and goals.
- The cards and their choices are revealed by all participants.

3. **Consensus**

- Each participant argues their choice.
- The moderator records the rationales.
- If one or more security tactics are selected by all the participants, they become the selected tactics.
- If there is no consensus, participants can discuss immediately and try to make a new common choice on the spot; if this is difficult, the moderator can start everything again from the second step.

To evaluate the use of TaSPeR in the selection of security tactics, Osses et al. [38] defined a context, some scenarios and, for each of these scenarios, a ground truth or set of tactics whose selection is considered optimal. Three metrics were also defined to determine to what extent the security tactics selected by the teams participating in the experiment were close to the ground truth. This metrics, widely used in data retrieval for performance evaluation, are precision (P), recall (R) and accuracy (A)[5] and are defined as follows:

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- TP (True Positives) are correctly selected tactics.
- TN (True Negatives) are correctly NOT selected tactics.
- FP (False Positives) are incorrectly selected tactics.
- FN (False Negatives) are incorrectly NOT selected tactics.

3.3 Experimental design

The study was planned following Wohlin at al.'s proposal [51]. We reused the context and some scenarios of Osses et. al [38] and their associated ground truths (details about how it was obtained can be found in [38]).

An experimental design was required to collect evidence on the effect of experience on the efficacy of the use of the TaSPeR technique for the correct selection of architectural tactics. To carry out the experiment, there were 24 IT professionals belonging to a master's program at the Universidad Técnica Federico Santa María.

3.3.1 Metrics

First, it was required to define the metrics that would be used as response variables in this experiment, which should quantify the efficacy of the use of the TaSPeR technique for the correct selection of architectural tactics. These metrics were obtained from those used in Osses et. al [38] (see section 3.2.3.) to measure teams performance in correct selection of tactics (precision, recall and accuracy) in a given scenario. The TaSPeR efficacy metrics were defined as the variation of these teams performance metrics when using the technique (precision, recall, and accuracy variation).

3.3.2 Hypotheses

Tables 3.1, 3.2 and 3.3 present the null hypothesis and the alternative hypotheses corresponding to each of the three efficacy metrics previously defined as response variables to be measured. It was necessary to formulate three families of hypotheses, one per metric, since each one in itself could give different results, so they had to be treated independently. Regarding the two alternative hypotheses for each family, the first one is due to the conjecture made that the more experienced teams should be more efficacious in the use of TaSPeR, while the second covers the possibility that the result would be contrary to expected, that is, that the more experienced teams be more efficacious in using TaSPeR.

Having formulated the null and alternatives hypotheses for each of the three defined metrics, we will now proceed to determine and explain the experimental design that will be used to test them. For this, we will begin by defining for this study some basic elements of an experimental design, which are presented in table 3.4.

3.3.3 Subjects

The experiment considered the division of the 24 participants into four teams, of six members each. The subjects were each of the four teams. Two of them were made up of the 12 IT professionals with less years of experience (teams 1 and 2) and the

Table 3.1: Precision related hypotheses

H_{10}	The efficacy of the TaSPeR technique for precision in selecting software architecture security tactics does not differ when used by more experienced or less experienced IT professionals.
H_{11}	The efficacy of the TaSPeR technique for precision in selecting software architecture security tactics is greater when used by less experienced IT professionals than when used by more experienced IT professionals.
H_{12}	The efficacy of the TaSPeR technique for precision in selecting software architecture security tactics is greater when used by more experienced IT professionals than when used by less experienced IT professionals.

other two by the 12 IT professionals with more years of experience in the area (teams 3 and 4).

Table 3.5 shows the characterization in terms of years of professional experience in the IT field, of the two teams with less experience as a whole (e-), of the two teams with more experience as a whole (e+) and of the members of each of the four teams separately. The names and years of experience of each of the participants can be found at the experimental package, which is available in `https://tinyurl.com/yyyv9o8f`.

In general, it was sought to assign the 12 professionals with less experience to teams 1 and 2, and the 12 with more experience to teams 3 and 4, trying to equate to the maximum the mean, the median and the standard deviation between the team pairs of the same level of experience. However, on the day the experiment was carried out there were unforeseen absences of participants that forced a restructuring. As part of this restructuring, team 4 (one of the two with the more experienced members) was assigned a professional with just six years of experience, which contrasted with the 9 years of experience that the more experienced professionals in among those who belonged to teams 1 and 2 (less experienced teams). This, added to another unexpected absence of a professional with 12 years of experience that was supplied by one with 25 years of experience, unbalanced a bit the similarity in the parameters that characterize the level of experience of the two teams with more experience, being generated between both a difference of 3.4 years in the average experience (19.2 years in team 3 versus 15.8 years in team 4), eight years in the median (22 years in team

Table 3.2: Recall related hypotheses

H_{20}	The efficacy of the TaSPeR technique for recall in selecting software architecture security tactics does not differ when used by more experienced or less experienced IT professionals.
H_{21}	The efficacy of the TaSPeR technique for recall in selecting software architecture security tactics is greater when used by less experienced IT professionals than by more experienced IT professionals.
H_{22}	The efficacy of the TaSPeR technique for recall in selecting software architecture security tactics is greater when used by more experienced IT professionals than by less experienced IT professionals.

3 versus 14 years in team 4) and 1.7 in the standard deviation (6.6 in team 3 versus 8.3 years in team 4).

On the other hand, the difference in average experience between the two more experienced teams seems insignificant compared to the difference in average experience of the more experienced teams with respect to the less experienced ones. For example, the average experience of team 4 (15.8 years), which is the least experienced of the two more experienced teams, is 150% higher than that of the least experienced teams (both 6.3 years).; in contrast, the average experience of team 3 (19.2 years) is only 21% greater than that of team 4 (15.8 years).

Not all participants had a degree in computer science, but all of them had been working in the IT field for several years, which was the metric used to determine their experience in order to form the most and least experienced teams.

Design

In accordance with what is presented in table 3.4, this experimental design consists of three factors, each of which in turn has two treatments or levels. That is, it is a 3-factor, 2-level experiment. These three factors are:

1. Technique used for tactics selection: One treatment consists of the use of an ad-hoc technique (T1) and the other, in the use of TaSPeR (T2) for the selection of architectural tactics.

2. IT professionals' level of experience in the field: One treatment consists in the

Table 3.3: Accuracy related hypotheses

H_{30}	The efficacy of the TaSPeR technique for accuracy in selecting software architecture security tactics does not differ when used by more experienced or less experienced IT professionals.
H_{31}	The efficacy of the TaSPeR technique for accuracy in selecting software architecture security tactics is greater when used by less experienced IT professionals than by more experienced IT professionals.
H_{32}	The efficacy of the TaSPeR technique for accuracy in selecting software architecture security tactics is greater when used by more experienced IT professionals than by less experienced IT professionals.

selection of tactics by teams whose members have more experience in the IT field (E1) and the other, in the selection of tactics by teams whose members have less experience in the IT field (E2). This experiment seeks to observe the effect it produces on the response variables, to fix this factor at one level or the other or, in other words, apply one treatment or the other.

3. Defined scenarios: One treatment consists of the selection of tactics in the first proposed scenario (S1) and the other, in the selection of tactics in the second proposed scenario (S2). This is a confounding factor that generates noise in the result of the experiment, but cannot be avoided, since, if the teams always worked on the same scenario, the learning effect would be produced. However, this effect can be blocked through experimental design, as explained later.

In order to cover all the possible combinations of factors and block the confounding factor generated by the different scenarios explained above, a full factorial experimental design was chosen, as illustrated in table 3.6 (symbols defined in table 3.4). To calculate the number of possible combinations of this 3-factor, 2-level experiment, equation $C = t^k$ is used, where t is the number of treatments or levels (two), k the number of factors (three) and C the number of possible combinations. In this way, we obtain that $C = 2^3 = 8$; that is, there are eight possible combinations.

To avoid the learning effect as a confounding factor, we chose to divide the experiment into two phases: a first phase in which an ad-hoc technique was used, without explaining yet what TaSPeR consisted of and a second phase in which it

Table 3.4: Basic experimental elements defined for this study

Concept	**In this experiment**	
Context	Four teams, each consisting of six IT professionals, selecting security tactics for two different scenarios, first using an ad-hoc technique and then using TaSPeR.	
Object of study	TaSPeR architectural tactics selection technique.	
Subjects	Two teams made up of IT professionals with less experience and other two made up of IT professionals with more experience in the field.	
Factors	Technique used for tactics selection.	$T1$: Without TaSPeR (Ad-hoc).
		$T2$: With TaSPeR.
	IT professionals' experience level in the field.	$E1$: Less experience.
		$E2$: More experience.
	Defined scenarios.	$S1$: Scenario 1.
		$S2$: Scenario 2.
Response variables	$\triangle P$: Precision variation with respect to the ground truth when applying TaSPeR.	
	$\triangle R$: Recall variation with respect to the ground truth when applying TaSPeR.	
	$\triangle A$: Accuracy variation with respect to the ground truth when applying TaSPeR.	

Table 3.5: Team members experience characterization

	Experience level		Team number			
	$e-$	$e+$	1 $(e-)$	2 $(e-)$	3 $(e+)$	4 $(e+)$
Min	3	6	3	4	10	6
Max	9	30	9	9	25	30
Mean	6.3	17.5	6.3	6.3	19.2	15.8
Median	6	17	6.5	6	22	14
Mode	6	20,24,14	-	6	24	14
σ	1.9	7.4	2.2	1.9	6.6	8.3

was used TaSPeR, previous induction of the technique. This induction considered the delivery of the cards and instructions for the use of TaSPeR, a presentation in which the technique was explained and the training of its use applied to a scenario, in order to minimize the effect of the learning curve.

This design allows the blocking of the effect of different scenarios, since, for example, if one of the teams with less experience used an ad-hoc technique in scenario 1 and TaSPeR in scenario 2, the other team with less experience do the inverse.

Response variables identified in table 3.4 ($\triangle$P, $\triangle$R and $\triangle$A) quantify the variation in teams performance when using TaSPeR to select tactics with respect to that achieved without the use of this technique (i.e., the efficacy of the technique). Table 3.7 (symbols defined in table 3.4) formally define the aforementioned response variables for the more and less experienced teams. For example, less experienced teams precision variation ($\triangle\ P_{E1}$) is defined as their precision using TaSPeR ($P_{E1\,T2}$) minus their precision without using this technique ($P_{E1\,T1}$).

Based on the response variables described above, table 3.8 formally define hypotheses formulated for this study. The null hypotheses (H_{10}, H_{20} and H_{30}) states for each response variable, that its value is similar for more experienced and less ex-

Table 3.6: 3-factor and 2-level complete factorial experimental design

		Combination number	Factors			Subjects
			Technique	Experience	Scenario	
Phase	I	1	$T1$	$E1$	$S1$	Team 1
		2	$T1$	$E2$	$S1$	Team 3
		3	$T1$	$E1$	$S2$	Team 2
		4	$T1$	$E2$	$S2$	Team 4
	II	5	$T2$	$E1$	$S2$	Team 1
		6	$T2$	$E2$	$S2$	Team 3
		7	$T2$	$E1$	$S1$	Team 2
		8	$T2$	$E2$	$S1$	Team 4

Table 3.7: Response variables formal definitions

	Response variable	Team experience	Formal definition
Variation	Precision variation	Less	$\triangle P_{E1} = P_{E1\,T2} - P_{E1\,T1}$
		More	$\triangle P_{E2} = P_{E2\,T2} - P_{E2\,T1}$
	Recall variation	Less	$\triangle R_{E1} = R_{E1\,T2} - R_{E1\,T1}$
		More	$\triangle R_{E2} = R_{E2\,T2} - R_{E2\,T1}$
	Accuracy variation	Less	$\triangle A_{E1} = A_{E1\,T2} - A_{E1\,T1}$
		More	$\triangle A_{E2} = A_{E2\,T2} - A_{E2\,T1}$

perienced teams (i.e., the efficacy of the technique is similar in both cases); the first alternative hypotheses (H_{11}, H_{21} and H_{31}) states for each response variable value, that is higher for less experienced teams than for more experienced ones (i.e., the technique is more efficacious for less experienced teams than for more experienced ones or, if the technique turns out to be detrimental, less inefficacious for less experienced teams than for more experienced ones); and the second alternative hypoteses (H_{12}, H_{22} and H_{32}) states for each response variable value, that is higher for more experienced teams than for less experienced ones. (i.e., the technique is more efficacious for more experienced teams than for less experienced ones or, if the technique turns out to be detrimental, less inefficacious for more experienced teams than for less experienced ones).

For each response variable (ΔP, ΔR and ΔA), two samples are contrasted, both of size two: that of more experienced teams and that of less experienced ones. For each sample, one observation is obtained using an ad hoc technique in scenario 1 and TaSPeR in scenario 2 and the other observation, vice versa. For each observation, the value obtained for its performance metric (precision, recall or accuracy, as the case may be) without using TaSPeR is subtracted from the value obtained using the technique, this difference being the value of the response variable measured by the observation. In this way, the observations that make up each sample correspond to the variations when using TaSPeR in the values of the performance metrics

Table 3.8: Hypotheses formalization

Metrics	H_0
Precision	$H_{10} :\triangle P_{E1} =\triangle P_{E2}$
Recall	$H_{20} :\triangle R_{E1} =\triangle R_{E2}$
Accuracy	$H_{30} :\triangle A_{E1} =\triangle A_{E2}$

Metrics	H_1
Precision	$H_{11} :\triangle P_{E1} >\triangle P_{E2}$
Recall	$H_{21} :\triangle R_{E1} >\triangle R_{E2}$
Accuracy	$H_{31} :\triangle A_{E1} >\triangle A_{E2}$

Metrics	H_2
Precision	$H_{12} :\triangle P_{E1} <\triangle P_{E2}$
Recall	$H_{22} :\triangle R_{E1} <\triangle R_{E2}$
Accuracy	$H_{32} :\triangle A_{E1} <\triangle A_{E2}$

mentioned above.

3.3.4 Sample sizes

Samples sizes of two are not very adequate to carry out statistical tests, since they lack enough statistical power. Before opting for this experimental design and for these sample sizes, we made an analysis of alternatives with which we could achieve sample sizes large enough to obtain statistical significance if our conjecture was correct and the effect size was enough to be detected by a statistical test. This dependeded in part on whether or not the population distribution was normal, since the statistical power of a parametric test is much higher. To verify the normality of the distributions, it was necessary to perform a a Shapiro Wilk test, but this required a sample size of at least three. A homoscedasticity test would also have been necessary to verify the similarity of the variances of both samples.

We identified two ways to obtain sample sizes greater than two: increasing the number of teams and increasing the number of scenarios to be solved by each team. If both measures were applied, the sample sizes could be increased even further. We will analyze both options.

Regarding the option of increasing the sample sizes by increasing the number of teams, this implied that each team would have fewer members. One option was to divide the 24 participants into six teams of four members each, of which three teams would be more experienced and three less experienced; however, this meant that two less experienced teams would solve the scenarios in one order and just one in the reverse order, which would generate an imbalance in blocking the effect of the order of the scenarios. The same would happen with the most experienced teams.

The alternative to avoid this was to form eight teams of three members each, of which four teams would be more experienced and four less experienced. This would allow two less experienced teams to solve the scenarios in one order and the other two in the reverse order, balancing the blocking effect of the order of the scenarios. The same would happen with the most experienced teams. This alternative allowed increasing from sample sizes of two to sample sizes of four. However, this option was rejected as it was considered that teams made up of only three people did not allow a technique based on consensus to be adequately carried out, so it was privileged to work with larger teams to favor further discussion in the search for consensus involving TaSPeR.

Regarding the option of increasing the sample sizes by increasing the number of scenarios resolved by each team, the increase from two to four scenarios to be resolved by each of them was analyzed, so that they began by solving two scenarios without TaSPeR and then they ended by solving the other two scenarios with TaSPeR. Like the previous one, this alternative allowed increasing sample sizes of two to sample sizes of four. Furthermore, if it had been possible to implement both alternatives at the same time, it could have been increased from sample sizes of two to sample sizes of eight, which would have led to a much more adequate statistical power. However, this increased the duration of the experiment by approximately 1 hour, which exceeded the time available to do so. The option of eliminating some of the complementary activities of the experiment, such as inductions or the resolution of the training scenario, was also analyzed, in order to have the time necessary to solve these two additional scenarios. However, it was considered that all these activities were necessary for the validity of the experiment, so it was rejected to eliminate them as well.

Therefore, the results obtained in this study are analyzed without obtaining conclusions that have a statistical value, for which new studies will be necessary to confirm them. which should replicate this experiment to add new observations and, if possible, achieve larger sample sizes.

3.3.5 Post-experiment survey

At the end of the experiment, a brief post-experiment survey was considered to obtain some information about the profile of the participants and their impressions.

3.3.6 Pilot study

Before carrying out the experiment, a pilot version was made to test the experimental design. In this pilot study, seven students from scientific postgraduate courses and one undergraduate student from Universidad Técnica Federico Santa María participated. Given the number of people, it was carried out with a partial factorial design, which considered only one team with less experience and one with more

experience, each consisting of four people. This design did not block the effect of different scenarios.

From this pilot study we obtained some lessons that allowed us to improve the actual experimental study, being some of them:

- It was better to translate the cards into Spanish, since not everyone is fluent in English.
- Making the cards of each tactical category of a different color facilitates its use.
- The concepts of signature in detection of attacks and steganography, are not known by all, so they must be previously explained.
- Some aspects of the context and scenarios should be better specified.
- It was necessary to explain in detail to the participants what the activity consisted of and to clarify some aspects to avoid ambiguities, such as the fact that when moving to a new scenario, those in which they had previously worked should not be considered.

3.4 Summary

Software requirements are a specification of what should be implemented. Quality attribute requirements specify the responses of the system that realize the goals of the business.There are techniques an architect can use to achieve the required quality attributes, called architectural tactics. TaSPeR is a technique that supports tactics selection to make robust design decisions.

The proposal of this work is an experimental design which evaluate the impact of experience in the TaSPeR technique efficacy. Experimental design considered three metrics to be used as response variables, which quantify the efficacy of TaSPeR: precision, recall and accuracy variation when using the technique (ΔP, ΔR and ΔA, respectively). It was necessary to evaluate the impact of the experience on these variables. For each of them, a null and two alternatives hypotheses were proposed. The null hypotheses states for each response variable, that its value is similar for more experienced and less experienced teams (i.e., the efficacy of the technique is similar in both cases); the first alternative hypotheses states for each response variable value, that is higher for less experienced teams than for more experienced ones (i.e., the technique is more efficacious for less experienced teams than for more experienced ones); and the second alternative hypotheses states for each response variable value, that is higher for more experienced teams than for less experienced ones (i.e., the technique is more efficacious for more experienced teams than for less experienced ones).

Experimental design is of three factors, each of two levels: the technique used for tactics selection (an ad-hoc technique or TaSPeR), the IT professionals' level

of experience in the field (less or more experience) and the scenario (scenario 1 or scenario 2). In order to cover all the possible combinations of factors, a full factorial experimental design was chosen. The samples obtained were of size 2, which does not allow obtaining statistical significance in the results. Alternatives that allowed larger sample sizes were explored, but these had disadvantages that made their implementation not recommended.

Before materializing this study, we carried out a pilot study from which we learned some lessons that allowed us to improve the design and execution of this experiment.

Chapter 4

Experimental study

This chapter aims to present the experimental study results and its threats to validity. Section 4.1 introduces to the experimental study; section 4.2 reports the results from different points of view and answers the first research question; section 4.3 presents some threats to validity to those results; and finally, section 4.4

4.1 Introduction

The experimental study was carried out on July 6th, 2019, at the Vitacura headquarters of the Universidad Técnica Federico Santa María and the entire activity lasted two and a half hours.

Interesting results and surprising findings were obtained from its execution, which are presented and analyzed in this chapter. In addition, the validity threats related to the results obtained are exposed.

4.2 Results

Tactics selection results within each scenario, as well as values of performance and efficacy obtained for each performance metric and how they varied after using TaSPeR are shown in table 4.1. They can be found at the experimental package, which is available in `https://tinyurl.com/yyyv9o8f`.

This table lists the 17 security tactics used in TaSPeR and, for each scenario, indicates with an “X" symbol which of them are part of the ground truth (GT) and which teams selected them. It does the latter by indicating for each selection the result obtained in terms of the confusion matrix (TP: True positive; FP: False positive; TN: True negative; FN: False negative) and highlighting this result in bold for the selected tactics. The numbers that head the columns correspond to those of the teams and for each team, it is specified in parentheses whether it made the selection of tactics using an ad hoc technique (T1) or TaSPeR (T2). In addition,

Table 4.1: Tactics Selection Results

		GT	Scenario 1				GT	Scenario 2			
			Teams Experience level					Teams Experience level			
			Less		More			Less		More	
			1 (T1)	2 (T2)	3 (T1)	4 (T2)		2 (T1)	1 (T2)	4 (T1)	3 (T2)
Detect Attacks											
1	Verify Message Integrity	**X**	FN	FN	FN	FN		**FP**	TN	TN	TN
2	Verify Storage Integrity		TN	TN	TN	TN	**X**	**TP**	**TP**	**TP**	**TP**
3	Maintain Audit Trail		**FP**	TN	**FP**	TN		**FP**	**FP**	TN	TN
4	Identify Intrusion by Signature		TN	TN	TN	TN		TN	**FP**	TN	TN
5	Identify Intrusion by Behavior		**FP**	TN	TN	TN		TN	**FP**	TN	TN
Stop or Mitigate Attacks											
6	Authenticate subjects	**X**	**TP**	**TP**	**TP**	**TP**		**FP**	TN	**FP**	**FP**
7	Authorize Subjects	**X**	**TP**	**TP**	**TP**	**TP**		**FP**	TN	**FP**	**FP**
8	Manage Security Information	**X**	**TP**	**TP**	**TP**	FN	**X**	**TP**	**TP**	**TP**	**TP**
9	Filter Data		**FP**	**FP**	TN	TN		**FP**	TN	TN	TN
10	Verify Origin of Message	**X**	FN	**TP**	**TP**	FN		**FP**	TN	TN	TN
11	Establish Secure Channel	**X**	FN	FN	FN	FN		**FP**	TN	TN	TN
12	Hide Data by Encription	**X**	FN	**TP**	FN	FN	**X**	**TP**	**TP**	**TP**	**TP**
13	Hide Data by Steganography		TN	TN	TN	TN		TN	TN	TN	**FP**
React to Attacks											
14	Alerts Subjects		TN	**FP**	**FP**	**FP**		**FP**	**FP**	TN	**FP**
15	Apply Institutions Policies		**FP**	**FP**	TN	TN		**FP**	**FP**	TN	**FP**
Recover from Attacks											
16	Audit Actions		**FP**	**FP**	**FP**	**FP**		**FP**	**FP**	TN	**FP**
17	Apply Institutions Policies		**FP**	**FP**	TN	TN		**FP**	**FP**	**FP**	**FP**
TN	Correctly NOT selected tactics	10	4	5	7	8	14	3	7	11	7
TP	Correctly selected tactics	7	3	5	4	2	3	3	3	3	3
FN	Incorrectly NOT selected tactics		4	2	3	5		0	0	0	0
FP	Incorrectly selected tactics		6	5	3	2		11	7	3	7
P	By team		0.333	0.500	0,.571	0.500		0.214	0,.300	0,.500	0,.300
	Variation using TaSPeR		0,167		-0,071			0,086		-0,200	
R	By team		0,429	0,714	0,571	0,286		1,000	1,000	1,000	1,000
	Variation using TaSPeR		0,286		-0,286			0,000		0,000	
A	By team		0,412	0,588	0,647	0,588		0,353	0,588	0,824	0,588
	Variation using TaSPeR		0,176		-0,059			0,235		-0,235	

for each scenario, the results of the less experienced teams are grouped in the two leftmost columns and the results of the more experienced teams in the two furthest to the right.

Further down in the same table, the four options of the confusion matrix are presented, indicating what each of them refers to in terms of selection of tactics. In addition, for each scenario, the number of tactics that were counted within each of the four options of the confusion matrix is given, both for the ground truth, and for each of the four teams.

4.2.1 Analysis by scenario

At the end of table 4.1, the values for each performance metric (P: precision; R: recall; A: Accuracy) are given, for each team, selecting tactics within each scenario.

In addition, for each metric and within each scenario, the difference in performance between each pair of teams with the same level of experience is given, a value obtained by subtracting in each case the performance of the team that used TaSPeR (right column of each pair of teams with the same level of experience), to the performance of the one that used an ad hoc technique (left column of each pair of teams with the same level of experience). That is, these differences correspond to the variations in performance when using TaSPeR within each pair of teams with the same level of experience that selected tactics in the same scenario, which coincides with the definition of efficacy of a software architecture design decision-making technique given in section 1.1.6 and with the efficacy metric defined for this experiment in section 3.3.1. We just have to make the caveat that we are comparing the performance of different teams, beyond the fact that by sharing a similar level of experience, we can assume that they are equivalent, although strictly speaking they are not. The foregoing constitutes a threat to the validity of the conclusions that we could draw from these results, in addition to the fact that they are merely anecdotal as they do not have adequate sample sizes to obtain statistical significance from them.

With this experimental design, the four teams worked in both scenarios, but in each scenario, of the two teams with the same level of experience, one of them worked using TaSPeR and the other did not. For example, in scenario 1, both less experienced teams worked: team 1 and team 2. Of these, team 1 worked without using TaSPeR and team 2 using the technique. The same happened for each pair of teams with the same level of experience that worked in the same scenario.

When reviewing the results, it was observed that for each pair of teams with less experience that worked in the same scenario, the one that did it using TaSPeR obtained a better performance than the one that did it without using the technique. For example, in scenario 1, of the two least experienced teams, team 1, which was the one that did not use TaSPeR in that scenario, obtained a precision of 0.333, while team 2, which was the one that did use it, obtained a precision of 0.500. On the contrary, for each pair of teams with more experience that worked in the same scenario, the one that did it using TaSPeR obtained a worse result than the one

that did it without using the technique. For example, in scenario 2, of the two more experienced teams, team 4, which was the one that did not use TaSPeR, obtained a recall of 0.824, while team 3, which was the one that did use it, obtained a recall of 0.588.

The above was fulfilled in both scenarios and for the three performance metrics, except in the case of the recall of scenario 2, where a particular situation occurred, since the four teams obtained a recall of 1.0, which means that all of them selected 100% of the tactics that were correct according to the ground truth. This would be explained because the ground truth of this scenario had only three correct tactics, which, in view of the results, seem to be easy to infer.

4.2.2 Scenario equivalence analysis

We should make a comparison of the performance obtained by each team without using TaSPeR with that obtained by the same team using the technique. However, this analysis implies the assumption that both scenarios are equivalent in terms of their difficulty from the perspective of the different performance metrics established. This is a threat to conclusion validity that must be evaluated.

In a first analysis, it could be affirmed that a scenario with a greater number of tactics considered correct according to its ground truth should favor precision, since this implies fewer tactics considered incorrect, which are the ones that could result in false positive and negatively affect this metric. On the contrary, a scenario with fewer tactics considered correct according to its ground truth should favor recall, since this implies fewer tactics considered correct, which are those that could result in false negatives and negatively affect this metric. According to the above, scenario 1 should be more prone than scenario 2 to high precision and low recall, since its ground truth considers seven tactics as correct, against three of scenario 2. On the contrary, scenario 2 should be more prone than scenario 1 to high recall and low precision.

Additionally, we made a comparison of the results obtained in this experiment to verify if the above was fulfilled empirically. Table 4.2 and figures 4.1, 4.2 and 4.3 allows this comparison to be carried out.

Table 4.2 has the same structure explained above for table 4.1 regarding the grouping of teams by scenario and by level of experience. As in that table, the numbers in the columns correspond to those of the teams and after each one of them it is indicated in parentheses if the team that selected tactics in that scenario did so using an ad hoc technique (T1) or TaSPeR (T2). As in the table above, the performance metrics are denoted as P (precision), R (recall) and A (acuracy). For each of these metrics, the following is delivered:

- The performance of each team in selecting tactics in each scenario.
- The average performance obtained by each pair of teams with the same level of experience when selecting techniques in the same scenario.

Table 4.2: Performance Comparison between Scenarios

		Scenario 1				Scenario 2			
		Teams Experience Level							
		Less		More		Less		More	
		1 (T1)	2 (T2)	3 (T1)	4 (T2)	2 (T1)	1 (T2)	4 (T1)	3 (T2)
P	By Team	0.333	0.500	0.571	0.500	0.214	0.300	0.500	0.300
	Experience Level Media	0.417		0.536		0.257		0.400	
	Scenario Media	0.476				0.329			
	Scenarios Difference	0.148 (>)							
R	By Team	0.429	0.714	0.571	0.286	1.000	1.000	1.000	1.000
	Experience Level Media	0.571		0.429		1.000		1.000	
	Scenario Media	0.500				1.000			
	Scenarios Difference	0.500 (<)							
A	By Team	0.412	0.588	0.647	0.588	0.353	0.588	0.824	0.588
	Experience Level Media	0.500		0.618		0.471		0.706	
	Scenario Media	0.559				0.588			
	Scenarios Difference	0.029 (<)							

- The average of the performance obtained by the four teams in each scenario.
- The difference between the average obtained in scenario 1 and that obtained in scenario 2, specifying between parentheses which of them was greater.

Figures 4.1, 4.2 and 4.3 allow comparing for each metric the performance obtained by each team in both scenarios.

It can be seen that scenario 1 turned out to be more prone to precision, both in average values (see table 4.2) and comparing the performances obtained by each team (see figure 4.1). In the case of recall, it can be seen that scenario 2 turned out to be clearly more prone to precision, both in average values (see table 4.2) and comparing the performances obtained by each team (see figure 4.2); however, the particular situation of scenario 2 mentioned above, in which the four teams selected the three tactics that were correct for this scenario according to the ground truth, causing all of them to obtain a perfect recall (1.0), implied that did not record differences between performance without and with TaSPeR for this metric in this scenario. This resulted in scenario 2 not providing information for the recall analysis. Regarding accuracy, both scenarios had a more balanced behavior in terms of this metric, both at the level of averages (see table 4.2), as compared by team (see figure 4.3). In this way, it can be concluded that the precision tended to be greater in scenario 1, the recall was clearly greater in scenario 2 (in fact, it had a perfect behavior) and the accuracy proved to be quite balanced between both scenarios, so it seems to be the most appropriate metric to make an analysis.

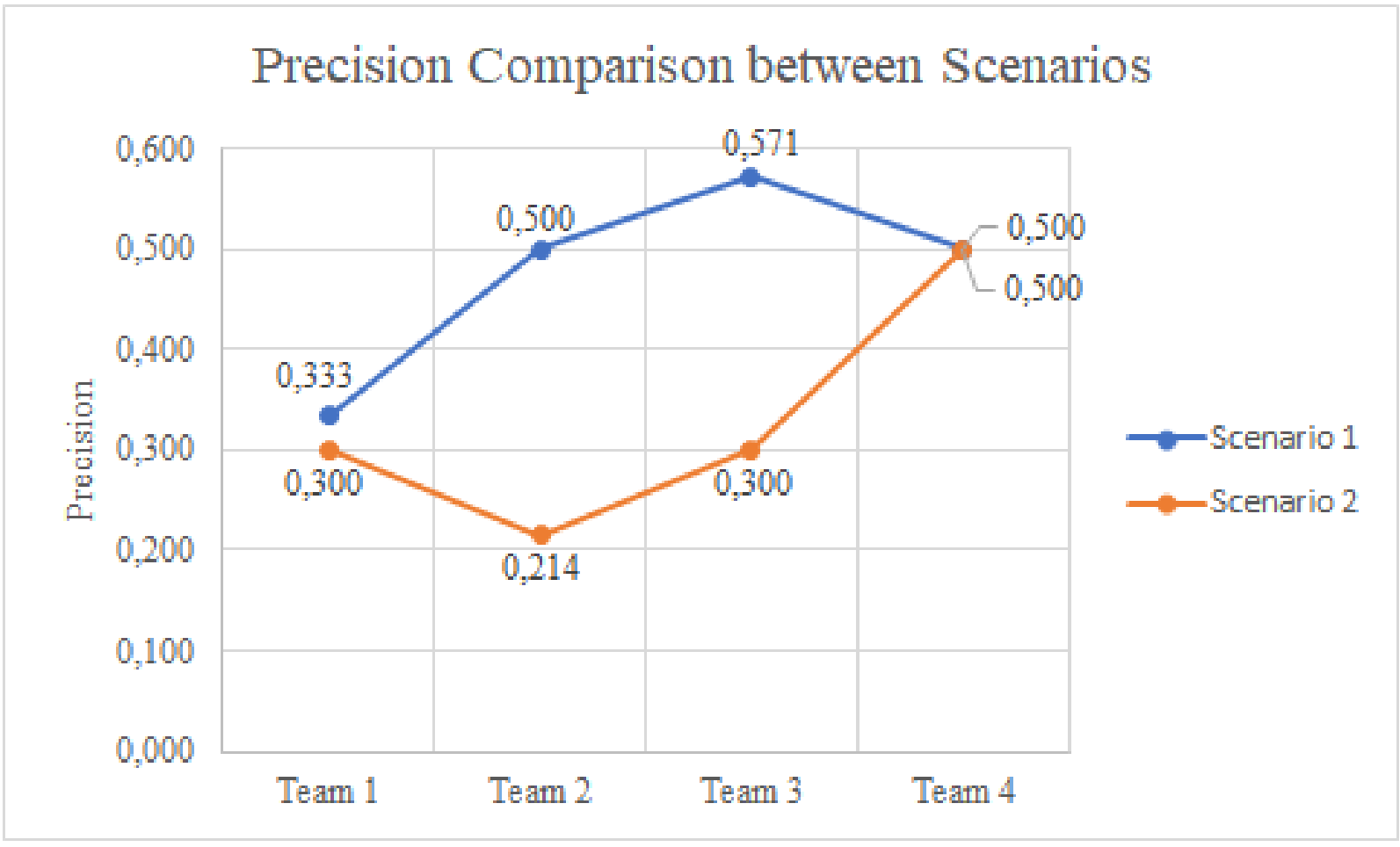


Figure 4.1: Precision comparison between scenarios.

4.2.3 Analysis by team

In table 4.1, for teams with the same level of experience that had selected tactics within the same scenario, the difference in performance for each metric was calculated between the one that had done it using an ad hoc technique and the one that had done it using TaSPeR. With this, the variation in performance is obtained when using TaSPeR, in the selection of techniques in the same scenario, although comparing different teams classified within the same level of experience. In table 4.3, this same variation is obtained, but this time comparing for each team the performances of the same team when using the technique or not, although this time selecting tactics in different scenarios. The latter is obtained by calculating for each team the difference between their performance in the correct selection of tactics using an ad hoc technique and their performance using TaSPeR. As the experimental design considered that the same team did not select tactics both times on the same scenario so that the results were not affected by the learning effect, this analysis is affected by the difference between the scenarios, especially considering the analysis in this regard carried out previously, in which it was concluded that this effect seemed to be important for the case of the precision and recall metrics.

Considering the above, table 4.3 groups on the left side teams that selected tactics using an ad hoc technique in scenario 1 and using TaSPeR in scenario 2 (team 1 of inexperienced and team 3 of experienced) and on the right side, those that did so in reverse, that is, who selected tactics using an ad hoc technique in scenario 2 and

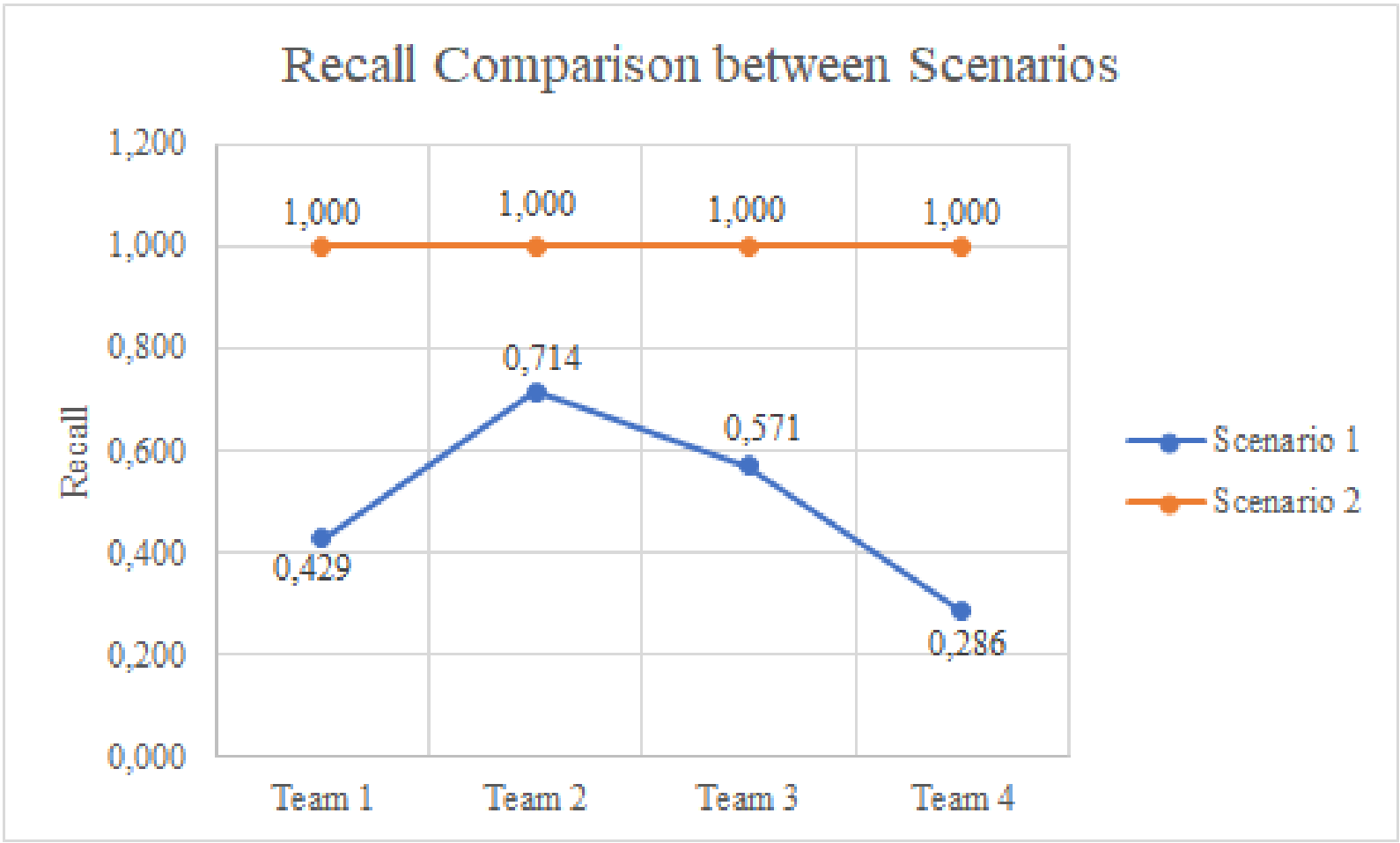


Figure 4.2: Recall comparison between scenarios.

Table 4.3: Comparison of performance variation of each team when using TaSPeR

		Teams from Scenario 1 to 2				Teams from Scenario 2 to 1			
		Team 1 (E1)		Team 3 (E2)		Team 2 (E1)		Team 4 (E2)	
		T1 (S1)	T2 (S2)	T1 (S1)	T2 (S2)	T1 (S2)	T2 (S1)	T1 (S2)	T2 (S1)
P	Without/With TaSPeR	0.333	0.300	0.571	0.300	0.214	0.500	0.500	0.500
	Variation Using TaSPeR	-0.033		-0.271		0.286		0.000	
R	Without/With TaSPeR	0.429	1.000	0.571	1.000	1.000	0.714	1.000	0.286
	Variation Using TaSPeR	0.571		0.429		-0.286		-0.714	
A	Without/With TaSPeR	0.412	0.588	0.647	0.588	0.353	0.588	0.824	0.588
	Variation Using TaSPeR	0.176		-0.059		0.235		-0.235	

using TaSPeR in scenario 1 (team 2 of inexperienced and team 4 of experienced). We will refer to these team groupings as teams "S1 to S2" and "S2 to S1", respectively. This way of grouping the teams allows to compare the variation when using TaSPeR of the performance in the correct selection of tactics, i.e., the efficacy shown by the technique, among the team with the lower level of experience that worked the scenarios in a certain order and the team with the higher level of experience that solved the scenarios in the same order, so the differences between the efficacy shown by TaSPeR when used by the inexperienced team and the experienced team is not affected by the difference between the scenarios. Within each group of teams ("S1 to S2" and "S2 to S1"), the less experienced team (E1) is located on the left and the more experienced team (E2) on the right. For each team, its performance for each metric when selecting tactics using an ad hoc technique (T1) is given in the

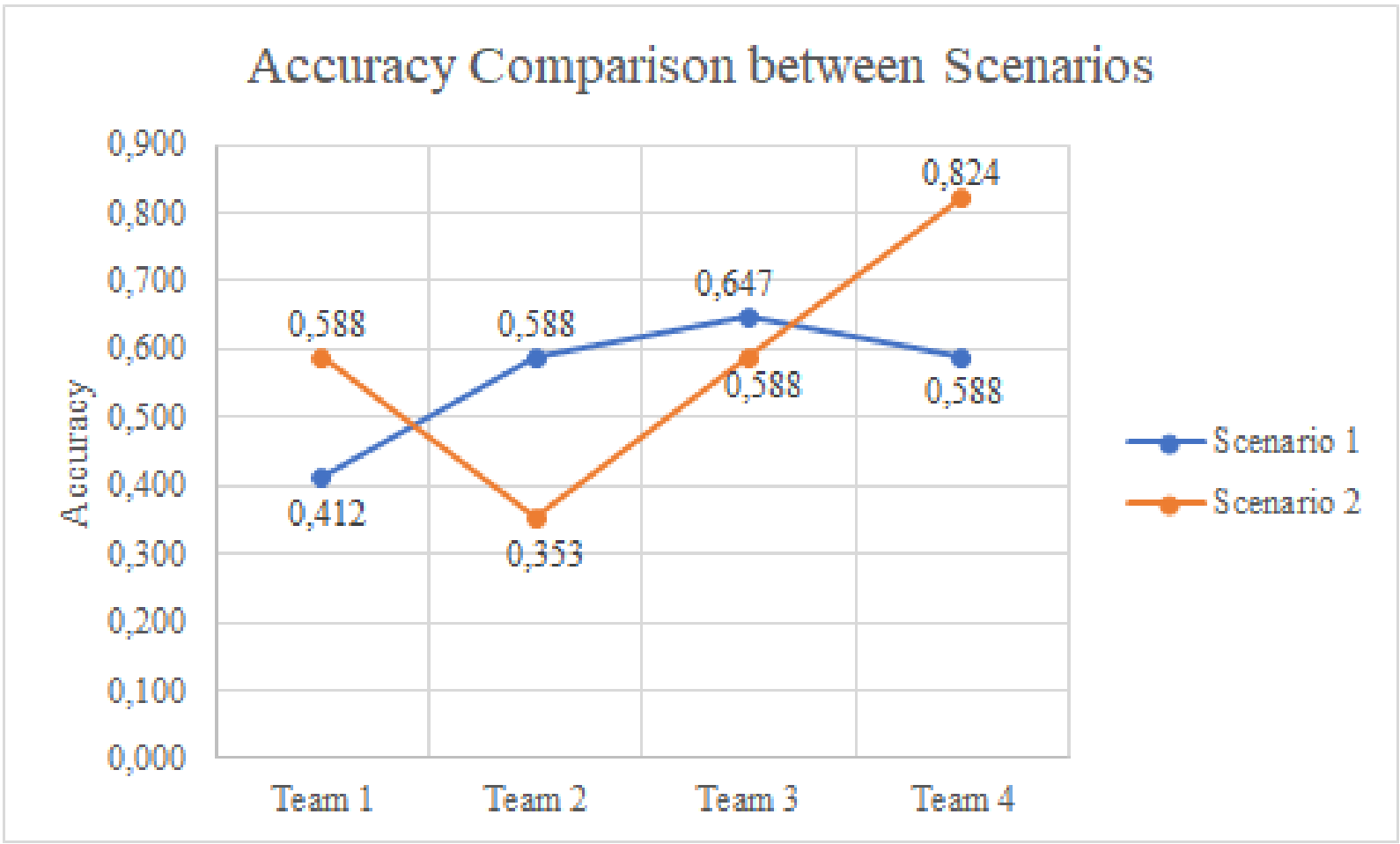


Figure 4.3: Accuracy comparison between scenarios.

left column and its performance for each metric when selecting tactics using TaSPeR (T2) is given in the right column. For each case, it is indicated in parentheses after T1 or T2 if the tactics were selected in scenario 1 (S1) or in scenario 2 (S2).

Finally, for each team and for each performance metric when selecting tactics, the subtraction of its value when using TaSPeR minus its value when using an ad hoc technique is calculated; in other words, its variation when using TaSPeR or the efficacy shown by the technique when used by each team. The way these values are arranged in the table allows to easily compare for each metric, the efficacy shown by TaSPeR when used by each inexperienced team that worked the scenarios in an order, with the efficacy shown by TaSPeR when used by the expert team that worked the scenarios in the same order.

Figures 4.4, 4.5 and 4.6 allow to graphically visualize the results presented in table 4.3, by metric (precision, recall and accuracy, respectively). In each graph, the use or not of TaSPeR is presented on the “X" axis and, on the “Y" axis, the performance obtained in each case for each metric. The points corresponding to each team are joined with a line of a different color. The legend of each graph indicates the team that each color represents and in parentheses, if it is a less experienced (E1) or more experienced (E2) team. In addition, the lines of the teams corresponding to the group “S1 to S2" are continuous and those of the teams corresponding to group “S2 to S1" are segmented.

When analyzing the results obtained, it can be seen that when comparing the

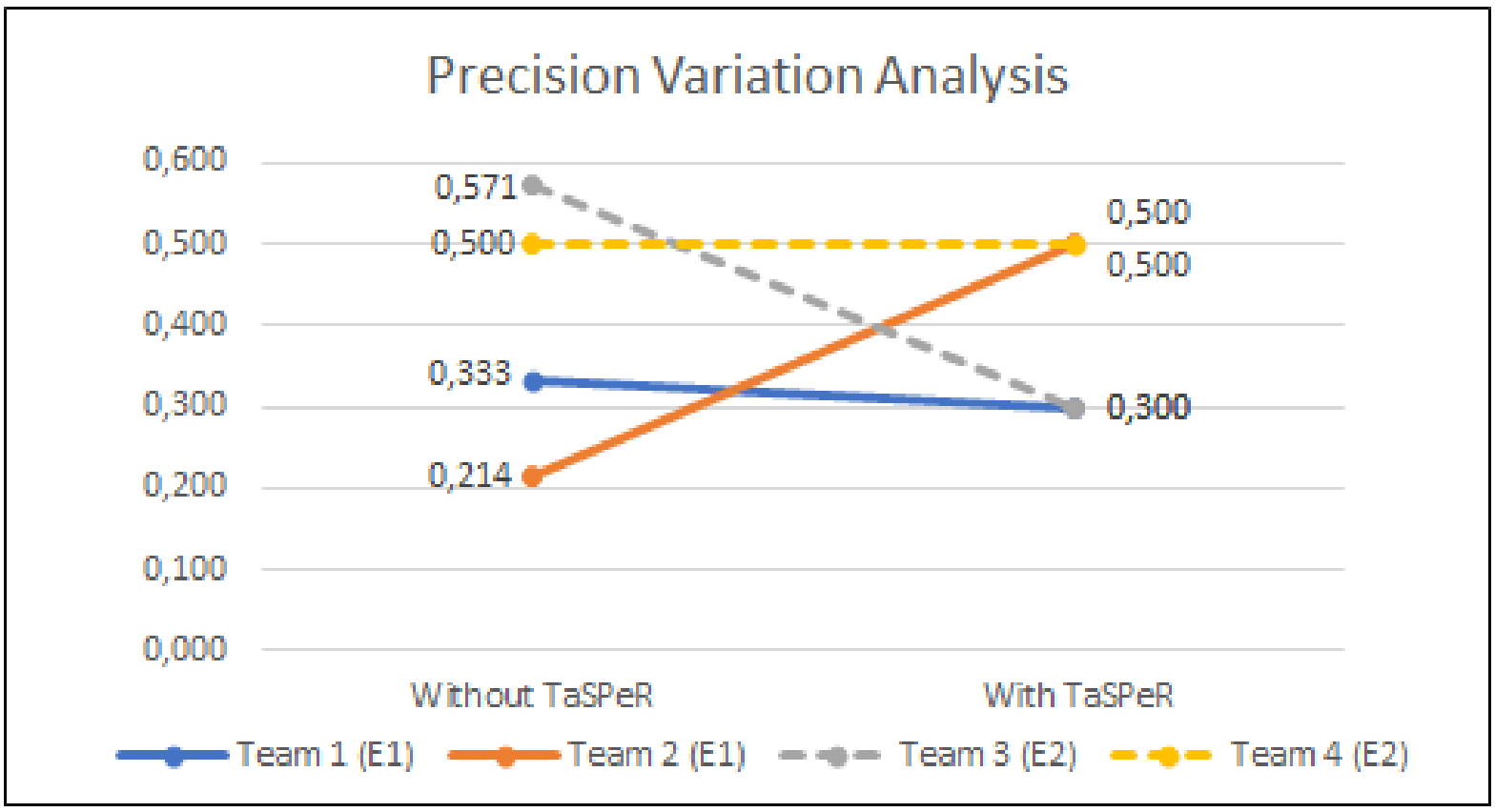


Figure 4.4: Precision variation analysis. Variations when using TaSPeR in teams S1 to S2 are represented with continuous lines and those corresponding to teams S2 to S1 are represented with segmented lines.

variation when using TaSPeR of the performance in the correct selection of tactics of the less experienced teams with respect to the more experienced ones, that is, when comparing the efficacy of the technique when used for one or the other, for the three metrics, it is true that the technique turns out to be more efficacious or at least less inefficacious when it is used by less experienced teams than when it is used by more experienced ones. We will analyze them one by one below. Figures 4.1, 4.2 and 4.3 will allow to graphically visualize each one of the situations that will be described.

In the case of the “S1 to S2" teams, which are team 1 of the less experienced teams and team 3 of the more experienced ones, for each metric, the following can be observed:

- Precision: Under this metric, the technique turned out to be inefficacious for both teams, which could be explained by a scenario 2 prone to lower precision than scenario 1. However, TaSPeR turned out to be less inefficacious for the less experienced team (variation when using TaSPeR of -0.033) than for the more experienced one (variation when using TaSPeR of -0.271).

- Recall: Under this metric, the technique turned out to be efficacious for both teams, which could be influenced by a scenario 2 prone to higher recall than scenario 1. TaSPeR turned out to be more efficacious for the less experienced team (variation when using TaSPeR of 0.571) than for the more experienced one (variation when using TaSPeR of 0.429).

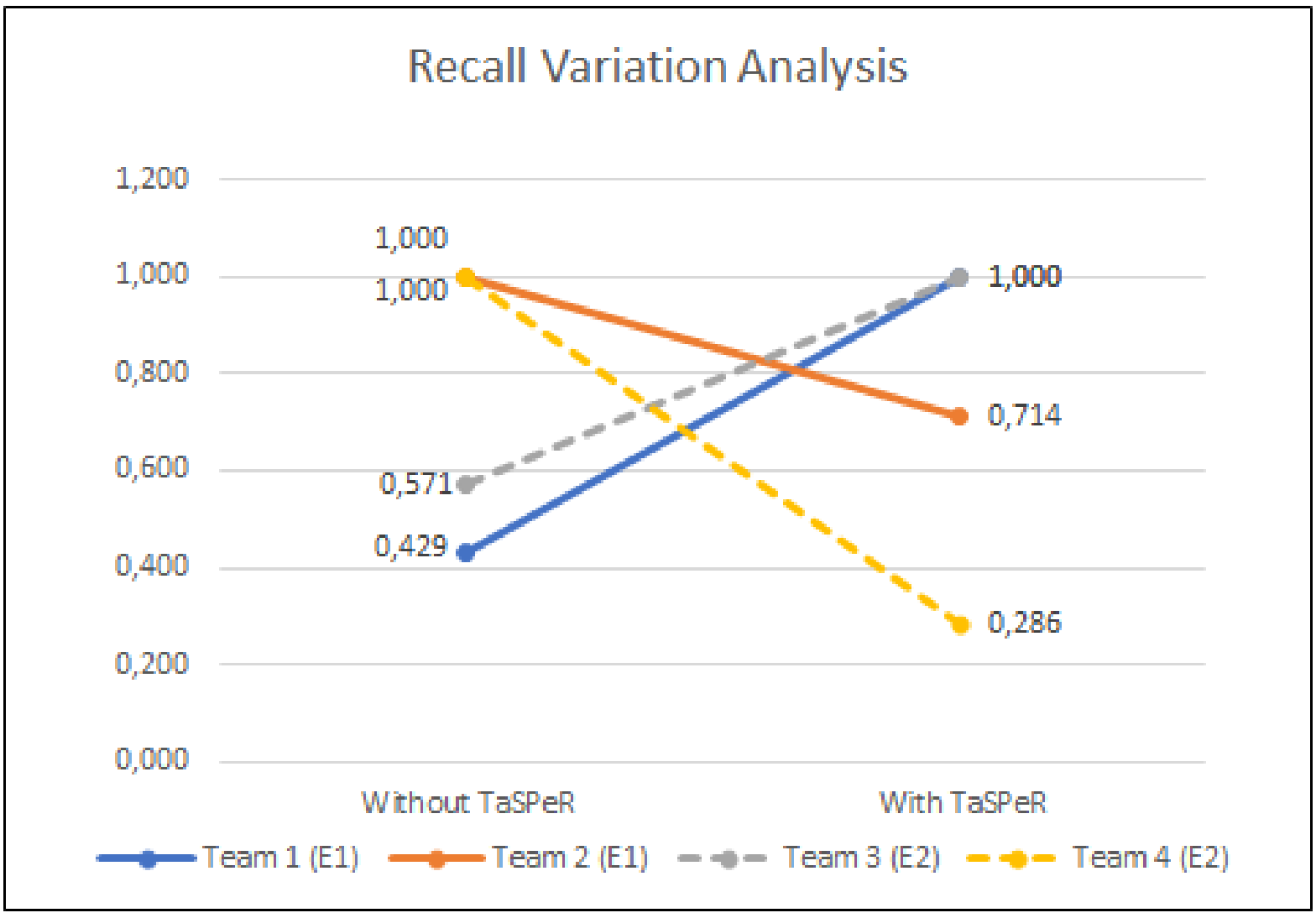


Figure 4.5: Recall variation analysis. Variations when using TaSPeR in teams S1 to S2 are represented with continuous lines and those corresponding to teams S2 to S1 are represented with segmented lines.

- Accuracy: Under this metric, which is not affected by the effect of the scenarios, the technique turned out to be efficacious for team 1 (variation when using TaSPeR of 0.176) and inefficacious for team 3 (variation when using TaSPeR of -0.059). That is, TaSPeR was efficacious for the less experienced team and inefficacious for the more experienced one.

In the case of the "S2 to S1" teams, which are team 2 of the less experienced teams and team 4 of the more experienced ones, for each metric, the following can be observed:

- Precision: Under this metric, the technique turned out to be efficacious for team 2 (variation when using TaSPeR of 0.286) and neutral for team 4 (variation when using TaSPeR of 0.000), which could be influenced by a scenario 1 prone to higher precision than scenario 2. That is, TaSPeR turned out to be efficacious for the less experienced team and neutral for the more experienced one.

- Recall: Under this metric, the technique turned out to be inefficacious for both teams, which could be explained by a scenario 1 prone to lower recall

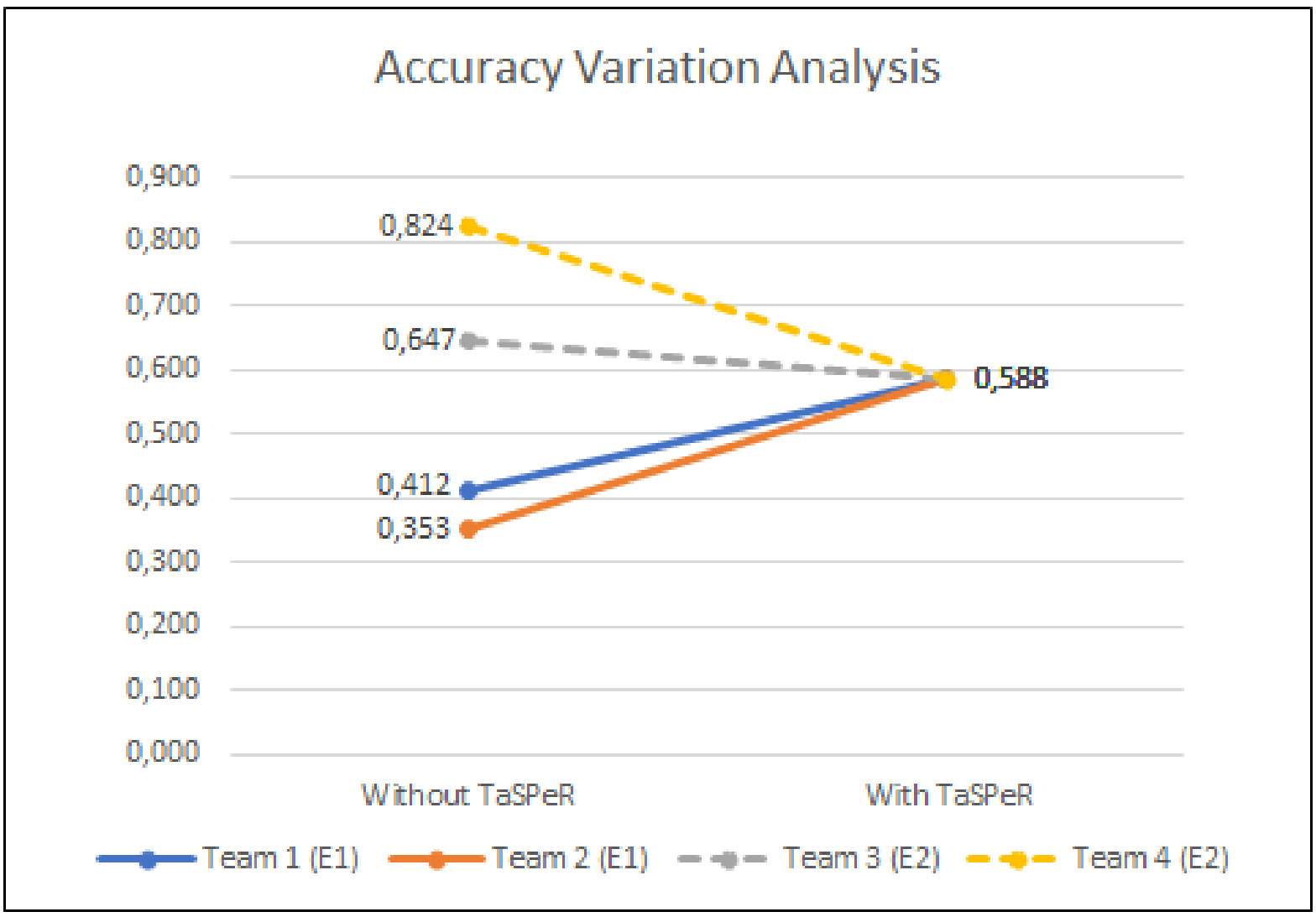


Figure 4.6: Accuracy variation analysis. Variations when using TaSPeR in teams S1 to S2 are represented with continuous lines and those corresponding to teams S2 to S1 are represented with segmented lines.

than scenario 2. However, TaSPeR turned out to be less inefficacious for the less experienced team (variation when using TaSPeR of -0.286) than for the more experienced one (variation when using TaSPeR of -0.714).

- Accuracy: Under this metric, which is not affected by the effect of the scenarios, the technique turned out to be efficacious for team 1 (variation when using TaSPeR of 0.235) and inefficacious for team 3 (variation when using TaSPeR of -0.235). That is, TaSPeR was efficacious for the less experienced team and inefficacious for the more experienced one.

Summarizing the results obtained with respect to the precision metric, when analyzing in table 4.3 the precision variation of each team when using TaSPeR and in the graphic of figure 4.4 the slopes of the lines that represent them, it can be seen that the variation in precision when using TaSPeR was negative for both teams S1 to S2 (teams 1 and 3), and positive (team 2) and neutral (team 4) for teams S2 to S1 (teams 2 and 4), which would be explained because scenario 1 is prone to greater precision than scenario 2. Despite the distortion produced by this lack of equivalence of scenarios in terms of precision, it can be seen that among the teams in which the

variation of this metric when using TaSPeR was equivalent, it was less harmful for the less experienced team (team 1) that for the more experienced one (team 3) in the case of teams S1 to S2 and, in the case of teams S2 to S1, there was no variation in precision when using TaSPeR by the more experienced team (team 4), in contrast to the positive variation recorded by the less experienced team (team 2). That is, in both cases the results in terms of precision when using TaSPeR were better (or less bad) for less experienced teams than for more experienced ones, which is consistent with our conjecture of that this technique should have better results when used for less experienced teams than when used for more experienced ones.

A similar analysis can be done for the recall (see table 4.3 for variation when using TaSPeR and graphic of figure 4.5 for the slopes of the lines that represent them). For this metric, both teams S1 to S2 had a positive variation when using TaSPeR and both teams S2 to S1 had a negative variation when not using it, consistent with a scenario 2 more prone to a greater recall than scenario 1. In this case, it is possible appreciate that, among the S1 to S2 teams, the less experienced team (team 1) obtained a greater recall increase when using TaSPeR than the less experienced one (team 3), while among the S2 to S1 teams, the more experienced team (team 4) obtained a greater recall decrease by using the technique than the less experienced one (team 2). That is, as in the case of precision, it was also in this metric in both cases that the results when using TaSPeR were better (or less bad) for less experienced teams than for more experienced ones, also consistent with our conjecture.

Finally, in the case of accuracy (see table 4.3 for variation when using TaSPeR and graphic of figure 4.6 for the slopes of the lines that represent them), a more direct analysis can be made, since scenarios 1 and 2 appear to be relatively equivalent in terms of this metric. It can be clearly seen that when using TaSPeR the accuracy of both less experienced teams (teams 1 and 2) increased, while that of the less experienced ones (teams 3 and 4) decreased, which is consistent with the previous analysis by scenarios, in which this same result was given for all three metrics.

4.2.4 Statistical Analysis

As mentioned in the previous chapter, considering that the samples obtained are size 2, the statistical tests lack sufficient power to detect significant differences between both populations (more and less experienced teams), even if there were any. Furthermore, for small sample sizes like this, it is advisable to do a non-parametric test, since the distribution tests (normality and homoscedasticity) also lack the necessary power to deliver significant results. Nor do we have information from previous studies that would allow us to assume that these samples come from populations with normal distribution and homogeneous variances, conditions necessary to carry out a parametric test. Nonparametric tests provide less statistical power than parametric ones, which makes the situation even worse.

The non-parametric tests that we could use in this case are the Mann-Whitney U

test, if they are independent samples, and the Wilcoxon signed-rank test, if the data are paired. When speaking of samples with paired data, it is usually understood that each pair of data was taken on the same subject, one before and the other after applying a certain treatment. However, as noted in Walpole et al.[49], another illustration of pairing involves choosing n pairs of subjects, with each pair having a similar characteristic such as IQ (intelligient quotient), age, or breed. As we have previously analyzed, for the precision and recall metrics, the order in which the scenarios were worked affects the values obtained for these metrics; so that the above does not affect the results of statistical significance of the difference between both samples, we considered as paired samples those obtained from pairs of teams (subjects) that worked the scenarios in the same order (teams “S1 to S2" and “S2 to S1 "), considering that this constitutes a similar characteristic that each pair of teams has, which according to Walpole et al. [49], allows them to be considered as paired. Thus, we used the Wilcoxon signed-rank test for the precision and recall metrics and the Mann-Whitney U test for the accuracy metric.

Under the precision metric approach, both for the “S1 to S2" teams (teams 1 and 3) and for the “S2 to S1" ones (teams 2 and 4), the efficacy of TaSPeR when used by the less experienced team it was higher than that obtained when used by the less experienced one. Specifically, within the “S1 to S2" teams, the efficacy of the technique when used by the less experienced team (team 1) was -0.033, 0.238 higher than that of the more experienced one (team 3), which was -0.271 (see table 4.3); within the “S2 to S1" teams, the efficacy of TaSPeR when used by the less experienced team (team 2) was 0.286, 0.286 higher than that of the more experienced one (team 4), which was 0 (see table 4.3). Although in both observations the same result was obtained for this metric (greater efficacy of the technique when used by less experienced team than by more experienced one), when applying the Wilcoxon signed-rank test for a sample of size two, a p-value of 0.5 is obtained, which does not obtain a statistically significant difference between the efficacy at use the technique of the less experienced teams and that of the more experienced ones, so it cannot be concluded that under this metric the technique turned out to be more efficacious for the less experienced teams than for the more experienced ones. To do so, it would require at least four other observations (six in total) in which the same result was repeated in each one of them (greater efficacy of the technique when used by less experienced team than by more experienced one), which would give a p-value of 0.032. If in any of these observations a greater efficacy of the technique was not obtained when used by less experienced teams than by more experienced ones, a greater number of observations would be required to be able to conclude statistical significance in the results. Additionally, if the distribution of both populations (efficacy when using TaSPeR of the less experienced teams and efficacy when using TaSPeR of the more experienced ones) were normal and their variances similar, which we have no way of knowing, when applying a paired t-test to the obtained samples, the p-value would be 0.058, with which statistical significance would not be obtained either.

Something similar happens under the recall metric approach, in that also for both the “S1 to S2" teams (teams 1 and 3) and for the “S2 to S1" ones (teams 2 and 4), the efficacy of TaSPeR when used by the less experienced team it was higher than that obtained when used by the less experienced one. Specifically, within the “S1 to S2" teams, the efficacy of the technique when used by the less experienced team (team 1) was 0.571, 0.143 higher than that of the more experienced one (team 3), which was 0.429 (see table 4.3); within the “S2 to S1" teams, the efficacy of TaSPeR when used by the less experienced team (team 2) was -0.286, 0.429 higher than that of the more experienced one (team 4), which was -0.714 (see table 4.3). As in the case of the precision approach, although in both observations the same result was obtained for this metric (greater efficacy of the technique when used by less experienced team than by more experienced one), when applying the Wilcoxon signed-rank test for a sample of size two, a p-value of 0.5 is obtained, which does not obtain a statistically significant difference between the efficacy at use the technique of the less experienced teams and that of the more experienced ones, so it cannot be concluded that under this metric the technique turned out to be more efficacious for the less experienced teams than for the more experienced ones. To do so, it would require at least four other observations (six in total) in which the same result was repeated in each one of them (greater efficacy of the technique when used by less experienced team than by more experienced one), which would give a p-value of 0.032. If in any of these observations a greater efficacy of the technique was not obtained when used by less experienced teams than by more experienced ones, a greater number of observations would be required to be able to conclude statistical significance in the results. Additionally, if the distribution of both populations (efficacy when using TaSPeR of the less experienced teams and efficacy when using TaSPeR of the more experienced ones) were normal and their variances similar, which we have no way of knowing, when applying a paired t-test to the obtained samples, the p-value would be 0.295, with which statistical significance would not be obtained either.

In the case of the accuracy metric approach, as noted above, the Mann-Whitney U test was used. For this metric, both less experienced teams obtained a positive efficacy value (0.176 team 1 and 0.235 team 2) and both more experienced teams obtained a negative efficacy value (-0.059 team 3 and -0.235 team 4). Despite this, when applying the Mann-Whitney U test, a p-value of 0.333 is obtained, which does not obtain a statistically significant difference between the efficacy at use the technique of the less experienced teams and that of the more experienced ones, so it cannot be concluded that under this metric the technique turned out to be more efficacious for the less experienced teams than for the more experienced ones. To do so, it would require at least two other observations (four in total) in which the same result was repeated in each one of them (greater efficacy of the technique when used by less experienced team than by more experienced one), which would give a p-value of 0.029. Additionally, if the distribution of both populations (efficacy when using TaSPeR of the less experienced teams and efficacy when using TaSPeR of the more

experienced ones) were normal and their variances similar, which we have no way of knowing, when applying a t-test for independent samples to the obtained samples, the p-value would be 0.128, with which statistical significance would not be obtained either.

4.2.5 First research question answer.

Having done this analysis, we are able to answer our first research question:

RQ1

How does the experience of IT professionals using the TaSPeR technique affect the efficacy of the technique for precision, recall and accuracy in selecting software architecture security tactics?

Answer:

Although the results obtained do not have statistical significance given the size of the sample, they initially suggest that TaSPeR is more efficacious when used by teams with less experienced IT professionals than by those with more experienced IT professionals for the three established metrics. Moreover, it seems that TaSPeR favors the less experienced and disadvantages the more experienced. However, given that scenarios 1 and 2 were not equivalent in terms of the performance obtained in them by teams in the precision and recall metrics, this result is more evident only in the case of accuracy, since the performance with respect to this metric it is relatively equivalent in both scenarios. For precision and recall, it can only be seen consistently that the results are better when used by more experienced teams than for less experienced ones, but it is not clear to what extent the variations when using TaSPeR were affected as the scenarios are not equivalent for these metrics. Only under the assumption that teams with a similar level of experience are equivalent in their performance with respect to these metrics, we could say that TaSPeR actually benefited the less experienced teams and harmed the more experienced ones under these two metrics.

4.3 Threats to Validity

This section presents threats of validity for the current design. The above means, any situation related to the experiment (internal and/or external), from its design, construction, application, among others, that could affect the results and the validation of it. Threats that are not applicable to this experimental design are not mentioned

4.3.1 Construct validity

- Inadequate preoperational explication of constructs: A validation of the experiment definition, design and analysis strategy was planned and performed through a pilot, to guarantee the consistence of problem and the experiment and to improve any other thing that affect the comprehension and/or execution of the experiment.

- Social Threats - Evaluation Apprehension: It was explicitly stated that evaluation results would not be publicized identifying each participant, and that there is an academic grade involved just for participating, but this is not linked to performance or the results of the experiment; the grade is only for participation.

- Experimenter Expectancy: Experimenters did not solve doubts about requirements or the announcement of the problem. They just supervised the activity and informed the time left during the execution of the different scenarios.

4.3.2 Internal validity

- History: to avoid the occurrence of events that could change history of subjects between their two participations, both scenarios were performed one after another, to not improve their technical abilities or the knowledge about the techniques or about the problem and to maintain the same group of people performing the experiments.

- Maturation: a five minutes break was considered in the middle of each trial and the application of a training scenario after the explanation of TaSPeR.

- Testing: no feedback was provided to subjects before the end of the three trials. Subjects were not told about how their performance had to be evaluated or about the answers of the problem given..

- Selection: Subjects come from different kind of industries and have variety of experience, so to mitigate this threat. Furthermore, they were assigned to the different teams based on their experience in the IT field. Also, there were no volunteers, so no self-selection existed. Teams were made maintaining the balance between experience of the participants, balancing the average experience and a standard deviation no more than 8.3

- Mortality: Due the trials were made one after another, there were no desertions, which mitigated possible threats to internal validity. The only thing that changed, was that two of the participants were not the day of the experiment and instead there were two other participants not informed to the experimenters. They occupied the empty spaces and moved the standard deviation of the experiences in the two teams with more experienced members.

4.3.3 External validity

- Representative Subjects: the subjects belonged to different type of industries being representatives not just under a single field, for that reason, results can be generalized for similar situations, although not necessarily for the selection of tactics of other quality attributes or for the use of techniques different from this one.

4.3.4 Conclusion validity

- Reliability of the Measures: the metrics used are those used by Alvarez [5].
- Reliability on Treatment Implementations: Training considered an explanation of TaSPeR and the application of the method in a training scenario to ensure process conformance.
- Random Heterogeneity of Subjects: the participants used for the experiment were students belonging to a Universidad Técnica Federico Santa María IT master program, people who work in the industry in different positions and with different experience, which introduces heterogeneity on the subjects.

4.4 Summary

In this chapter, the results were presented and analyzed from different perspectives.

First, an analysis by scenario was carried out. Variations in performance when using TaSPeR, that is, the efficacy of the technique, were analyzed within each pair of teams with the same level of experience that selected tactics in the same scenario. As a result, it was observed that for each pair of teams with less experience that worked in the same scenario, the one that did it using TaSPeR obtained a better performance than the one that did it without using the technique and, for each pair of teams with more experience that worked in the same scenario, the one that did it using TaSPeR obtained a worse result than the one that did it without using the technique. The above was fulfilled in both scenarios and for the three performance metrics, except in the case of the recall of scenario 2, where the four teams obtained a recall of 1.0.

Afterwards, a scenario equivalence analysis was carried out, concluding that the precision tended to be greater in scenario 1, the recall was clearly greater in scenario 2 (in fact, it had a perfect behavior) and the accuracy proved to be quite balanced between both scenarios.

Subsequently, an analysis by team was carried out. Variation when using TaSPeR of the performance in the correct selection of tactics, i.e., the efficacy shown by the technique, among the team with the lower level of experience that worked the scenarios in a certain order and the team with the higher level of experience that

solved the scenarios in the same order. As a result, it can be seen that the technique turns out to be more efficacious or at least less inefficacious when it is used by less experienced teams than when it is used by more experienced ones.

When doing a statistical analysis of the results obtained, it is concluded that there is no statistical significance, which is explained by the sample sizes, as indicated in the previous chapter.

With all this information, it was possible to answer our first research question.

Finally, the validity threats of the present study were presented in terms of construct, internal, external and conclusion validity, pointing out how they were mitigated.

Chapter 5

Discussion: Results and possible explanations

This chapter aims to discuss the experimental study results and look for possible explanations in the literature. Section 5.1 introduces to the discussion; section 5.2 carry out the results discussion; section 5.3 explores in the literature some results possible explanations; section 5.4 answers the second research question; and section 5.5 summarizes this chapter.

5.1 Introduction

Having presented and analyzed the results in the previous chapter, this time a more in-depth discussion is carried out on these and possible explanations for them are sought in the existing literature.

5.2 Discussion

The results obtained do not allow us to conclude with statistical significance, as discussed in Chapter 3, but provide evidence that could be corroborated with new studies, suggesting that the experience could have an impact on the efficacy of TaSPeR in terms of precision, recall and accuracy, being greater in the case of teams with less experienced members than in those with more experienced ones and even hurting the latter. These results are consistent with our initial conjecture, according to which, since the guidelines that grant this type of technique are more useful for those who have less experience than for those who have more experience, these should be more suitable for the former, although in this case it even hurts the latter. Further, in order to generalize this particular TaSPeR result to the rest of the software architecture design decision-making techniques or in a broader scope (for example, software engineering design decision-making techniques in general), it would be necessary to

replicate this experimental study in other techniques of this type within wider or different scopes.

There was consistency in that for all three metrics, performance variations when using TaSPeR were more favorable (or less unfavorable) for less experienced teams than for more experienced ones. In the other hand, as indicated in the previous section, although results obtained seem to indicate that TaSPeR helped the less experienced teams but harmed the most experienced ones, the lack of equivalence evidenced between both scenarios in terms of precision and recall generates a confounding factor that threatens validity of this statement for these two metrics. Not so in the case of accuracy, a metric in which the scenarios turned out to be equivalent enough not to lift this threat to validity.

A lesson learned from this situation is that for this type of analysis it is convenient to avoid scenarios whose ground truths have such a small number of correct selections as was the case in scenario 2, since they tend to be less prone to omissions and make it difficult to obtain valid conclusions, especially for the recall. In addition, it should be avoided that the ground truth of the used scenarios have a very different amount of correct selections, as it seems to favor the lack of equivalence of scenarios in terms of precision and recall, since those whose ground truth considers less correct selections tend to be prone to lower precision (more false positives) and higher recall (less false negatives) than those that considers more correct selections. Indeed, the characteristics of scenario 2, whose ground truth was made up of only 3 tactics, to which it seems that it was quite evident that they should be selected . This made it a scenario too little prone to omissions, which negatively affected the purposes of this experiment. This background suggests that it is convenient to avoid scenarios of these characteristics for this type of experiments.

The fact that TaSPeR seems to be more efficacious for the selection of tactics when used by teams with less experience than by those with more experience could be because it provides guidelines for a deeper analysis, first at the individual level and then at the collective level, also promoting the participation of all team members, all of which could contribute to complement incomplete individual visions product of knowledge shortcomings of less experienced members.

On the other hand, the fact that TaSPeR seems to be detrimental to the selection of tactics by the teams with more experience is something that we did not expect. In the worst case, we would have expected the technique to have had little or no positive effect for teams with more experienced members. A possible explanation for this result could be that the guidelines provided by TaSPeR end up confusing the most experienced, leading them to make worse decisions when seeking a consensus with the rest of the team members than they would otherwise, in which they would probably prevail the opinions of the most experienced ones. In addition, it should be considered that in the study in which TaSPeR was proposed, the efficacy of the technique was not evaluated in teams with more experienced members, since teams with equivalent experience were formed.

In the case of recall, all 4 teams scored the highest possible value in scenario 2, regardless of whether or not they used TaSPeR to select tactics. As a consequence of the above, it is logical that the two teams "S1 to S2" (team 1 with less experience and team 3 with more experience) have increased their recall when using TaSPeR from values lower than 1.0 to 1.0, and that the two teams "S2 to S1" (team 2 with less experience and team 4 with more experience) have reduced their recall when using TaSPeR from 1.0 to values lower than 1.0. The difference lies in the fact that in the case of teams "S1 to S2", which are those that worked in scenario 1 without TaSPeR, the more experienced team obtained a higher performance than the less experienced (0.571 team 3 vs 0.429 team 1), while in the case of teams "S2 to S1", who are the ones who worked in scenario 1 with TaSPeR, it was the least experienced team that obtained the highest performance under this metric (0.714 team 2 vs 0.286 team 4). Furthermore, the less experienced team that used TaSPeR in scenario 1 (team 2) obtained a higher recall than both teams that did not use it in the same scenario (team 1 with less experience and team 3 with more experience) and, on the contrary, the more experienced team that used TaSPeR in scenario 1 (team 4) obtained a worse recall. than both teams that did not use it on the same scenario. In the first case, this could be explained because the discussion and search for consensus tends to confuse the most experienced, making them discard tactics for which they would otherwise have opted, which can be translated into an increase in omissions. In the second case, a possible explanation could be that, as already mentioned, the fact that TaSPeR encourages the individual analysis and discussion, favors a deeper analysis and a complement of knowledge that covers possible individual flaws of less experienced teams members, that allows to reduce the number of omissions and increase the right choices in a meaningful way. This results obtained for recall are consistent with what is observed in general, in that TaSPeR appears to benefit novices, but harm experts.

A surprising result that can be seen in the graphics of figures 4.4, 4.5 and 4.6 is that when using TaSPeR, the performances of the metrics for teams S1 to S2 and those of teams S2 to S1 perfectly converged with each other at a unique value in five of the six situations analyzed. This convergence occurred in both cases for precision and accuracy, and in one case for recall ("S2 to S1"). However, in the latter case the special effect produced in the recall of scenario 2, which turned out to be 1.0 in all cases, must be considered. Only in one of the cases this convergence did not occur (and in fact it was far from occurring), which was in the recall of teams "S2 to S1". The tendency to equalize these metrics when using this technique could be due to the effect of the application of common guidelines. It is important to clarify that the selections made by each team differ among themselves for the same scenarios, which indicates that they reach the same values through different errors and successes (see table 4.1).

Even more surprising is the specific case of accuracy (see figure 4.6), which when using TaSPeR converged to a single value in all four cases (0.588), which means

that the four teams had the same number of successful decisions (true positives and negative), which were 10 in all cases, representing 58.8 % of the total tactics. This could reflect that when using the technique, teams tend to converge to this percentage of right decisions, which should be corroborated with new studies.

To better understand some of the aspects discussed in this section, this study could be replicated by recording, in addition to the tactics that the team selects, those for which each one of its members has opted before moving on to the discussion and consensus steps, in order to be able to make a further analysis of how this consensus was reached and who were the opinions that prevailed of. A deeper analysis could also involve the registration and subsequent analysis of discussions carried out to reach such consensus.

It is important to point out that the results of this work were obtained within the framework of the selection of security tactics. We do not know if the results of this experiment can be dependent on focusing on the security attribute nor what would have happened if the study would have been carried out with tactics for other quality attributes, which is an open question that could also be addressed as future work.

Another aspect that must be taken into account is that it could be that the dynamics of the exercise with novice professionals differ from the dynamics with experienced professionals and that this has influenced the results. This dynamic could also be studied as future work.Some aspects of this experiment that are worth mentioning, because they could shed light on the above, are that the experienced ones finished before their assigned times ran out, while the novices finished just right, in a hurry. Also, in general, subsequent feedback on the technique was more positive from novices than from experienced ones.

5.3 Results possible explanations

We did a review of the existing literature on design decisions in search of possible explanations for the results observed in this experiment, obtaining some interesting findings that we present below

Ahmed et al. [2] developed an observational study to understand how novice and experienced designers approach design tasks. They found that experienced designers used particular design strategies, whereas novice designers did not. They concluded that when developing support methods for novice designers, consideration should be given to informing them of such strategies in addition to providing them with knowledge and information. TaSPeR could be helping to inform novices on these strategies, leading them to improved performance.

Ahmed et al. [1] carried out also a research to understand how to support designers through the provision of appropriate knowledge by an empirical study that analysed the interactions between novice and experienced engineering designers in the aerospace industry. Among the findings of this work, it is pointed out that novice

designers tend to be unaware of the design strategies employed by experienced designers. In this regard, the study suggests that supporting novice designers by simply supplying knowledge may not be enough, so they also require support in identifying what they need to know. TaSPeR provides these guides to novices, which would explain why the technique turned out to be useful for them and not for experts.

Cross et al. [16] presented a paper of the field of research in expertise in design, focused specifically on expert performance, in which they reviewed studies of design expertise refered to expert vs. novice performance, expert designer behaviour and outstanding designers. Within the conclusions of this work, they point out that generating a wide range of alternative solution concepts is an aspect of design behaviour which is recommended by theorists and educationists but appears not to be normal practice for expert designers and add that generating a very wide range of alternatives may not be a good thing: some studies have suggested that a relatively limited amount of generation of alternatives may be the most appropriate strategy. In this way, in the case of TaSPeR, the wide range of possible solutions offered by this technique could end up making the work of the experts difficult and reducing their performance in the correct selection of tactics, which could be an explanation for the detriment in performance of the experts when using the technique.

Daly et al. [18] made a study to reveal and investigate critical differences in how designers from within and outside of engineering disciplines understand what it means to design, and how those understandings are evident in their approaches to and progression through design work. Within the theoretical implications obtained from this work, they indicate that expert designers do not follow a process step by step; they make their own paths, guided by their own learned conceptions and priorities. In this sense, for an expert, TaSPeR could represent more of a difficulty than an aid in making design decisions, which would also be consistent with the results obtained.

5.4 Second research question answer

Having done this analysis, we are able to answer our second research question:

RQ2

What could be possible explanations for the results obtained in the experimental study carried out to evaluate how the experience of IT professionals using the TaSPeR technique affects the efficacy of the use of the technique?

Answer:

Some works shed light on possible explanations for the results obtained in this study. One of them concludes that when developing support methods for novice designers, consideration should be given to informing them of particular strategies

used by experienced designers, in addition to providing them with knowledge and information. Another study suggests that supporting novice designers by simply supplying knowledge may not be enough, so they also require support in identifying what they need to know. TaSPeR could be helping to inform novices on the strategies mentioned above and providing guides to support novices in identifying what they need to know, which could explain why it turns out to be more beneficial for novices than experts.

There is a study that within its conclusions points out that generating a wide range of alternative solution concepts is an aspect of design behavior which is recommended by theorists and educationists but appears not to be normal practice for expert designers and add that generating a very wide range of alternatives may not be a good thing. This could be a possible explanation about why TaSPeR turns out to be detrimental for experts performance. Another study indicates within its theoretical implications that expert designers do not follow a process step by step; they make their own paths, guided by their own learned conceptions and priorities. This could be also a a factor that explains why the step by step process provided by TaSPeR could be detrimental for experts.

5.5 Summary

In this chapter, we first discussed the results obtained in more depth. We analyze them from a statistical point of view, indicating that although they do not allow conclusions to be drawn because they are not statistically significant, future studies could allow it. Then we talk about the consistency of the results obtained, in that for all three metrics, performance variations when using TaSPeR were more favorable (or less unfavorable) for less experienced teams than for more experienced ones, although precision and recall were affected for the effect of the differences between the scenarios. Later we will talk about the lesson learned regarding the convenience of working with scenarios whose amounts of correct tactics according to the ground truth are similar. Later we made some conjectures about possible explanations for the results obtained, without yet going into what was found in the existing literature. For the recall case, we observed that in scenario 1, the less experienced team that used TaSPeR obtained a better performance than those that did not use it, while the more experienced team that used TaSPeR obtained the inverse result. We also talk about the observed trend towards the convergence of the performance of the different teams when using TaSPeR. To close the discussion, we mention some replication options of this study to deepen the analysis.

Finally, we closed this chapter presenting some possible explanations for the results obtained based in the existing literature, both for the greater efficacy of the technique when used by novices, and for why the technique turned out to be harmful

for experts. Possible explanations for why TaSPeR appears to be more beneficial for beginners than experts are that TaSPeR could help inform beginners about the strategies used by expert designers and provide guides to help them identify what they need to know. As for why using TaSPeR seems to hurt experts, this could be because generating a wide range of alternative solution concepts is an aspect of design behavior which is recommended by theorists and educationists but appears not to be normal practice for expert designer and generating a very wide range of alternatives may not be a good thing. In addition, expert designers do not usually follow a process step by step; they make their own paths, guided by their own learned conceptions and priorities, so TaSPeR could be more than a help a hindrance for them. This allowed us to answer our second research question.

Chapter 6

Conclusions and future work

THIS chapter describe in the Section 6.1 the conclusions of this Master Thesis and in Section 6.2 the future work related.

6.1 Conclusions

In this thesis, an experimental design was presented and materialized through a study to evaluate the impact of the experience of IT professionals selecting software architecture security tactics using TaSPeR, on the efficacy of the technique to achieve a correct selection of these tactics in terms of precision, recall and accuracy.

Initial results suggest that TaSPeR improves the accuracy of novice teams but hurts that of expert teams. This result is quite unexpected and begs replication with even larger populations of IT professionals (no easy task). If the results are confirmed, the question that will rise is: if consensus techniques are so good to estimate, why would they hurt design decision-making by expert teams?

The proposed experimental design seems to be adequate for its purpose, since it provided evidence to evaluate what was required, but the size of the samples that were available did not allow conclusions to be obtained with statistical significance.

An important finding is that for teams with more experienced members the use of TaSPeR turned out to be harmful for the selection of software architecture tactics, in contrast to the more experienced teams, for which the use of TaSPeR turned out to be quite beneficial. This result could be corroborated for accuracy, since for recall and precision, the lack of equivalence of the scenarios used in terms of these two metrics threatens the validity of this conclusion for them, since the degree of influence that the scenarios had on the results obtained cannot be established.

An interesting result that could be explored in greater depth in the future and that could constitute a finding is that when using TaSPeR, in 5 of the 6 cases there was a perfect convergence in the value obtained for each metric within the same scenario. Furthermore, in the case of accuracy, when using TaSPeR, in all cases the value of this metric converged to a single value: 0.588, decreasing for experts and

increasing for novices, which is quite surprising.

The existing literature on design decisions sheds light on possible explanations or factors that could influence the results obtained in this study. Possible explanations for why TaSPeR appears to be more beneficial for beginners than experts are that TaSPeR could help inform beginners about the strategies used by expert designers and provide guides to help them identify what they need to know. As for why using TaSPeR seems to hurt experts, this could be because generating a wide range of alternative solution concepts is an aspect of design behavior which is recommended by theorists and educationists but appears not to be normal practice for expert designer and generating a very wide range of alternatives may not be a good thing. In addition, expert designers do not usually follow a process step by step; they make their own paths, guided by their own learned conceptions and priorities, so TaSPeR could be more than a help a hindrance for them.

6.2 Future work

As future work, new studies could be carried out using this experimental design to evaluate the impact of experience in the efficacy of other techniques of this type, in order to obtain more evidence regarding its suitability and regarding the results obtained in this study for the technique used, evaluating if they are generalizable. Regarding the latter, this study could be replicated by reusing the experimental design, either in the selection of tactics of other quality attributes, such as in other software architecture design decision-making techniques or software engineering in general or even in other techniques used within software architecture or software engineering in general.

The study of the possible differences that could exist between the dynamics of the exercise of the expert professionals and that of the experienced ones and how this could have affected the results of this experiment could also be addressed as future work. From the above, a different way of implementing TaSPeR for experienced users could be proposed and evaluated.

Specifically for TaSPeR, a new study could be done to try to obtain statistical significance for the results obtained, as well as to contrast the tactics chosen individually by each member of a team with those selected after reaching a consensus and analyze the teams discussions that led to these consensuses object to explore the causes of the effects of the experience evidenced for this technique in this study.

Another aspect that could be addressed in future studies is the convergence of the values obtained in the different metrics when using TaSPeR and verifying if the value of 0.588 recorded in all cases is repeated for accuracy when using TaSPeR.

Bibliography

[1] S. Ahmed and K. M. Wallace. Understanding the knowledge needs of novice designers in the aerospace industry. *Design studies*, 25(2):155–173, 2004.

[2] S. Ahmed, K. M. Wallace, and L. T. Blessing. Understanding the differences between how novice and experienced designers approach design tasks. *Research in engineering design*, 14(1):1–11, 2003.

[3] T. Al-Naeem, I. Gorton, M. A. Babar, F. Rabhi, and B. Benatallah. A quality-driven systematic approach for architecting distributed software applications. In *Proceedings of the 27th international conference on Software engineering*, pages 244–253, 2005.

[4] A. M. Alashqar, H. M. El-Bakry, and A. A. Elfetouh. A framework for selecting architectural tactics using fuzzy measures. *International Journal of Software Engineering and Knowledge Engineering*, 27(03):475–498, 2017.

[5] S. A. Alvarez. An exact analytical relation among recall, precision, and classification accuracy in information retrieval. *Boston College, Boston, Technical Report BCCS-02-01*, pages 1–22, 2002.

[6] A. Andrews, E. Mancebo, P. Runeson, and R. France. A framework for design tradeoffs. *Software Quality Journal*, 13(4):377–405, 2005.

[7] M. Babar, I. Gorton, and R. Jeffery. Capturing and using software architecture knowledge for architecture-based software development. IEEE, 1 2006. Fifth International Conference on Quality Software (QSIC'05).

[8] L. Bass, P. Clements, and R. Kazman. *Software architecture in practice*. Addison-Wesley Professional, 2003.

[9] J. C. Carver. *The impact of background and experience on software inspections*. University of Maryland, College Park, 2003.

[10] J. C. Carver, N. Nagappan, and A. Page. The impact of educational background on the effectiveness of requirements inspections: An empirical study. *IEEE Transactions on Software Engineering*, 34(6):800–812, 2008.

[11] J. Chavarriaga, C. Noguera, R. Casallas, and V. Jonckers. Propagating decisions to detect and explain conflicts in a multi-step configuration process. In *International Conference on Model Driven Engineering Languages and Systems*, pages 337–352. Springer, 2014.

[12] J. Chavarriaga, C. Noguera, R. Casallas, and V. Jonckers. Managing trade-offs among architectural tactics using feature models and feature-solution graphs. In *2015 10th Computing Colombian Conference (10CCC)*, pages 124–132. IEEE, 2015.

[13] L. Chung, D. Gross, and E. Yu. Architectural design to meet stakeholder requirements. In *Working Conference on Software Architecture*, pages 545–564. Springer, 1999.

[14] M. Cohn. *Agile estimating and planning*. Pearson Education, 2005.

[15] N. Cross. Creative thinking by expert designers. *Journal of design research*, 4(2):162–173, 2004.

[16] N. Cross. Expertise in design: an overview. *Design studies*, 25(5):427–441, 2004.

[17] F. T. Dabous and F. A. Rabhi. A framework for evaluating alternative architectures and its application to financial business processes. In *Australian Software Engineering Conference (ASWEC'06)*, pages 10–pp. IEEE, 2006.

[18] S. R. Daly, R. S. Adams, and G. M. Bodner. What does it mean to design? a qualitative investigation of design professionals' experiences. *Journal of Engineering Education*, 101(2):187–219, 2012.

[19] S. DasanayakeJouni, M. MarkkulaSanja, A. AaramaaMarkku, and O. Oivo. Software architecture decision-making practices and challenges: An industrial case study. 9 2015. Conference: 24th Australasian Software Engineering Conference (ASWEC) 2015At: Adelaide, Australia.

[20] R. de Boer, P. Lago, R. Verdecchia, and P. Kruchten. Decidarch v2: An improved game to teach architecture design decision making. 3 2019. 3rd International Workshop on decision Making in Software ARCHitecture (MARCH) ; Conference date: 26-03-2019 Through 26-03-2019.

[21] A. H. Dutoit, R. McCall, I. Mistrík, and B. Paech. *Rationale management in software engineering*. Springer Science & Business Media, 2007.

[22] D. Falessi, G. Cantone, R. Kazman, and P. Kruchten. Decision-making techniques for software architecture design: A comparative survey. *ACM Computing Surveys (CSUR)*, 43(4):33, 2011.

[23] R. Farenhorst, P. Lago, and H. Van Vliet. Effective tool support for architectural knowledge sharing. In *European conference on software architecture*, pages 123–138. Springer, 2007.

[24] E. B. Fernandez, H. Astudillo, and G. Pedraza-García. Revisiting architectural tactics for security. In *European Conference on Software Architecture*, pages 55–69. Springer, 2015.

[25] D. Gatica, G. Marquez, and H. Astudillo. Systematic selection of software components through architectural tactics. is a relationship between tactics and nfrs possible? In *CIbSE*, pages 183–195, 2017.

[26] T. Gilb. *Competitive engineering: a handbook for systems engineering, requirements engineering, and software engineering using Planguage*. Elsevier, 2005.

[27] F. Gilson and V. Englebert. Rationale, decisions and alternatives traceability for architecture design. In *Proceedings of the 5th European Conference on Software Architecture: Companion Volume*, pages 1–9, 2011.

[28] A. Jansen and J. Bosch. Software architecture as a set of architectural design decisions. In *5th Working IEEE/IFIP Conference on Software Architecture (WICSA'05)*, pages 109–120. IEEE, 2005.

[29] R. Kazman, J. Asundi, and M. Klein. Quantifying the costs and benefits of architectural decisions. In *Proceedings of the 23rd International Conference on Software Engineering. ICSE 2001*, pages 297–306. IEEE, 2001.

[30] R. Kazman, M. Klein, and P. Clements. Atam: Method for architecture evaluation. Technical report, Carnegie-Mellon Univ Pittsburgh PA Software Engineering Inst, 2000.

[31] S. Kim. A quantitative and knowledge–based approach to choosing security architectural tactics. *International Journal of Ad Hoc and Ubiquitous Computing*, 18(1-2):45–53, 2015.

[32] S. V. F. Lopes and P. T. Aquino. Architectural design group decision-making in agile projects. In *2017 IEEE International Conference on Software Architecture Workshops (ICSAW)*, pages 210–215. IEEE, 2017.

[33] G. Márquez and H. Astudillo. Selecting components assemblies from non-functional requirements through tactics and scenarios. In *2016 35th International Conference of the Chilean Computer Science Society (SCCC)*, pages 1–11. IEEE, 2016.

[34] G. Márquez and H. Astudillo. Selection of software components from business objectives scenarios through architectural tactics. In *2017 IEEE/ACM 39th*

International Conference on Software Engineering Companion (ICSE-C), pages 441–444. IEEE, 2017.

[35] M. Moore, R. Kaman, M. Klein, and J. Asundi. Quantifying the value of architecture design decisions: lessons from the field. In *25th International Conference on Software Engineering, 2003. Proceedings.*, pages 557–562. IEEE, 2003.

[36] A. Nakakawa, P. v. Bommel, and H. Proper. Requirements for collaborative decision making in enterprise architecture. 2009.

[37] R. Noel, G. Pedraza-García, H. Astudillo, and E. B. Fernández. An exploratory comparison of security patterns and tactics to harden systems. In *CIbSE*, pages 378–391, 2014.

[38] F. Osses, G. Márquez, M. M. Villegas, C. Orellana, M. Visconti, and H. Astudillo. Security tactics selection poker (tasper): a card game to select security tactics to satisfy security requirements. In *Proceedings of the 12th European Conference on Software Architecture: Companion Proceedings*, page 54. ACM, 2018.

[39] G. Pedraza-Garcia, H. Astudillo, and D. Correal. A methodological approach to apply security tactics in software architecture design. In *2014 IEEE Colombian Conference on Communications and Computing (COLCOM)*, pages 1–8. IEEE, 2014.

[40] J. Ryoo, B. Malone, P. A. Laplante, and P. Anand. The use of security tactics in open source software projects. *IEEE Transactions on Reliability*, 65(3):1195–1204, 2015.

[41] C. Schriek, J. M. E. van der Werf, A. Tang, and F. Bex. Software architecture design reasoning: A card game to help novice designers. In *European conference on software architecture*, pages 22–38. Springer, 2016.

[42] P. Stoll, A. Wall, and C. Norstrom. Guiding architectural decisions with the influencing factors method. In *Seventh Working IEEE/IFIP Conference on Software Architecture (WICSA 2008)*, pages 179–188. IEEE, 2008.

[43] M. Svahnberg, C. Wohlin, L. Lundberg, and M. Mattsson. A quality-driven decision-support method for identifying software architecture candidates. *International Journal of Software Engineering and Knowledge Engineering*, 13(05):547–573, 2003.

[44] A. Tang, B. Floris, C. Schriek, and J. M. E. van der Werfb. Improving software design reasoning–a reminder card approach. *Journal of Systems and Software*, 144:22–40, 2018.

[45] H. J. v. V. H. Tang A., Tran M.H. *Design Reasoning Improves Software Design Quality*. Springer, 2008.

[46] U. Van Heesch, P. Avgeriou, and R. Hilliard. Forces on architecture decisions-a viewpoint. In *2012 Joint Working IEEE/IFIP Conference on Software Architecture and European Conference on Software Architecture*, pages 101–110. IEEE, 2012.

[47] S. Vijayalakshmi, G. Zayaraz, and V. Vijayalakshmi. Multicriteria decision analysis method for evaluation of software architectures. *International Journal of Computer Applications*, 1(25):22–27, 2010.

[48] P. Wallin, J. Froberg, and J. Axelsson. Making decisions in integration of automotive software and electronics: A method based on atam and ahp. In *Fourth International Workshop on Software Engineering for Automotive Systems (SEAS'07)*, pages 5–5. IEEE, 2007.

[49] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye. *Probability and statistics for engineers and scientists*, volume 5. Macmillan New York, 1993.

[50] K. Wiegers and J. Beatty. *Software requirements*. Pearson Education, 2013.

[51] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén. *Experimentation in software engineering*. Springer Science & Business Media, 2012.

[52] C. Zannier, M. Chiasson, and F. Maurer. A model of design decision making based on empirical results of interviews with software designers. *Information and Software Technology*, 49(6):637–653, 2007.

[53] O. Zimmermann. Architectural decisions as reusable design assets. *IEEE software*, 28(1):64–69, 2010.