UNIVERSITÀ
DEGLI STUDI
DI PADOVA

UNIVERSITA' DEGLI STUDI DI PADOVA

Dipartimento di Ingegneria Industriale DII

Dipartimento di tecnica e gestione dei sistemi industriali DTG

Corso di Laurea Magistrale in Ingegneria Meccanica

Tesi di Laurea

Innovation in Agile and Waterfall project management: a qualitative analysis.

Relatore

Ch. ma Prof.ssa Daria Battini

Laureando

Nicola Battoia 1156430

Correlatore estero presso TUM München

Ch. mo Prof. Rainer Kolisch

Persone in supporto TUM: *Tobias Lieberum, Dr. Sebastian Schiffels*

Anno Accademico 2018/2019

Abstract

Different project management approaches produce different outcomes from the product development phase. Nevertheless, it is not clear which approach should be used to reach innovation in NPD. I develop an individual laboratory experiment, that simulates the main characteristics of traditional and Agile paradigms in design, planning and execution phases. The purpose is to analyze if people behave differently and how this affects performances. Differences are researched in average and peak performance. The results show that the Agile approach is more efficient and less risky to reach a higher average result. Under specific conditions, it allows also to reach a peak performance in the project. The traditional approach produces a slower pace of work, with delays and late completion of requirements. One driver of poor performance is low levels of motivation and challenge. The Agile approach demonstrates to foster these incentives, while the traditional approach has not a clear impact on them.

Table of contents

Riassunto Esteso

Questo lavoro di tesi è stato svolto all'interno di un flusso Erasmus presso la Tecnische Universität München (TUM) School of Management, durante il semestre invernale 2018/2019. L'argomento trattato si colloca nell'ambito della gestione dell'innovazione prodotto. In particolare, l'interesse è quello di effettuare un'analisi comparativa di due tecniche di project management (PM), che viene specificata nel raggiungimento di risultati innovativi. Per realizzare questa comparazione, ho sviluppato un esperimento comportamentale individuale, che simula le caratteristiche principali di Waterfall e Agile project management nelle fasi di sviluppo prodotto. Lo scopo è analizzare se le persone si comportano in modo differente e come questo influenzi i risultati.

La ricerca in letteratura viene sviluppata su tre livelli. Nel primo, i campi dell'innovazione e dello sviluppo prodotto vengono analizzati separatamente. La necessità è quella di delineare le differenze tra innovazione radicale e incrementale, e tra traditional e Agile PM. Il secondo livello riguarda come i differenti approcci producano differenti risultati. Il terzo presenta gli aspetti rilevanti del campo delle Behavioral Operations (studi comportamentali applicati all' Operations management). Gli argomenti di innovazione e PM sono stati trattati separatamente sotto molti aspetti. La ricerca e la pratica delle aziende hanno sviluppato numerose teorie sui tipi di innovazione possibili e come raggiungerli. La branca di ricerca sul project management ha iniziato a svilupparsi per rispondere alle necessità di gestione di progetti complessi. Negli ultimi anni questo settore ha suscitato sempre più interesse, anche sospinto dalle tecniche di Modern Project Management (MPM). Una distinzione rilevante per questa ricerca è tra le tecniche tradizionali e quelle moderne di project management. Con traditional PM (TPM) vengono indicate le tecniche sviluppate negli anni '60 sulla base dell'approccio scientifico. In quegli anni il livello di complessità e costi richiedeva un approccio rigido alla pianificazione e gestione. L'approccio chiamato Waterfall veniva largamente impiegato e viene identificato dai ricercatori come l'approccio tradizionale per antonomasia. Quello moderno invece viene sviluppato per reagire alle nuove richieste del mercato nel ventunesimo secolo. Le nuove condizioni richiedono alle aziende di ridurre i cicli di vita per lo sviluppo prodotto, aumentare la responsività e la consapevolezza riguardo i bisogni dei clienti. Comunemente viene fatto corrispondere

l'inizio del MPM con la pubblicazione del Manifesto Agile nel 2001. All'interno dei principi che delineano l'approccio Agile, diverse pratiche per lo sviluppo prodotto sono state sviluppate. In questa tesi viene considerato l'approccio Scrum, delineato nelle sue caratteristiche dalla Scrum Guide™.

Le basi per definire le caratteristiche dell'esperimento sono ricercate nel campo Behavioral Operations. Vista la dimensione empirica, è necessaria la teoria preesistente in altri settori per una corretta comprensione. Per comprendere i comportamenti individuali vengono utilizzati i concetti della psicologia cognitiva. L'interesse specifico è focalizzato sui processi di decisione individuale. I principali effetti, valutati come rilevanti per questo esperimento, sono:

(1) Anchoring and insufficient adjustment heuristics, e planning fallacy;

(2) Procrastinazione e Parkinson's law;

(3) Il macro-fenomeno della Behavioral hill;

(4) Effetti del multitasking e della disponibilità di informazioni.

Applicando questi biases e modelli euristici, insieme ai risultati della ricerca sperimentale sulle attività di sviluppo prodotto, verranno definite le ipotesi dell'esperimento. L'impostazione dell'esperimento prevede che i partecipanti costruiscano delle strutture usando dei mattoncini LEGO®, secondo delle specifiche richieste presentate come richieste di clienti da soddisfare. I partecipanti vengono assegnati casualmente a uno dei due trattamenti presenti, che simulano traditional e Agile PM. Le differenze tra i due approcci vengono declinate in alcune caratteristiche specifiche:

(a) approccio alla pianificazione;

(b) relazione tra la pianificazione e l'esecuzione nella gestione del tempo;

(c) documentazione richiesta in supporto al processo;

(d) flessibilità ai cambiamenti durante lo sviluppo.

Riguardo la pianificazione, l'approccio tradizionale è caratterizzato da una scelta preventiva di tutte le specifiche. Lo scopo è quello di prevedere ogni inconveniente e risolverlo in anticipo. Le metodologie Agile invece pianificano solo l'iterazione successiva e reagiscono ai cambiamenti. La pianificazione e l'esecuzione sono nettamente separate nell'approccio tradizionale, mentre in quello Agile si alternano come conseguenza naturale del processo iterativo. Una documentazione estensiva è fondamentale nel TPM. Al contrario, le metodologie Agile non la riconoscono come un'attività che aggiunge valore. Riguardo ai cambiamenti, questi non influenzano il progetto nel suo complesso per l'Agile PM, perché possono essere implementati nell'iterazione successiva. Al contrario, nel TPM i cambiamenti implicano spesso dover riprogrammare l'intero progetto.

Le caratteristiche generali dell'esperimento sono un tempo totale fisso e un set di regole da rispettare, che simulano le differenze sopra descritte. Ai partecipanti è richiesto di costruire tre strutture principali, chiamate items, caratterizzate con sette sottostrutture in totale, chiamate sub-items. L'ordine di presentazione è lo stesso per tutti i partecipanti. Ogni sub-item costruito da ciascuno dei venti partecipanti viene valutato. La valutazione è eseguita tramite un sondaggio online, in cui 120 partecipanti hanno espresso un voto da 0 a 100 sulle foto dei sub-items costruiti. I risultati di questo sondaggio costituiscono la base per la valutazione delle performance dei due approcci, e vengono utilizzati per testare le ipotesi dell'esperimento. Le prime due ipotesi considerano gli approcci separatamente, comparando i risultati ottenuti nel set di sub-items:

H1. Con l'approccio traditional, i partecipanti raggiungono un risultato significativamente più alto nel primo sub-item. Si riscontra una differenza significativa tra il risultato del primo sub-item rispetto agli altri. I risultati mostrano un trend decrescente dal primo sub-item ai seguenti.

H2. Con l'approccio Agile, i partecipanti raggiungono un risultato medio in tutti i sub-items. Non si riscontra una differenza significativa tra nessun sub-item e gli altri.

La terza ipotesi è comparativa ed è divisa in due parti:

H3.1 Con l'approccio Agile, i partecipanti raggiungono un risultato medio più alto in confronto all'approccio traditional.

H3.2 Con l'approccio traditional, i partecipanti raggiungono un risultato significativamente più alto nel primo sub-item in confronto all'approccio Agile.

In aggiunta ai test statistici sui risultati, l'analisi è supportata dalle osservazioni qualitative sul comportamento dei partecipanti. Il risultato più significativo dell'esperimento è che con l'approccio Agile i partecipanti hanno raggiunto un picco di performance, che è risultato più significativo rispetto al trattamento traditional. È sorprendente in quanto in contraddizione con il risultato atteso. La spiegazione può essere trovata nel modo in cui i partecipanti hanno organizzato il loro lavoro. Nel trattamento Agile hanno finito per concentrarsi su una richiesta specifica. Le ragioni sono state o per incrementare un risultato o per raggiungere un livello sufficiente in qualcosa di incompleto. In aggiunta, i risultati confermano l'ipotesi H2, perché i partecipanti hanno raggiunto un risultato medio significativamente più alto. L'esperimento conferma anche la ricerca che sostiene come gli approcci incrementali e a spirale riducano il rischio di fallimento in un progetto. Si è visto come i partecipanti nell'approccio traditional hanno fallito la consegna di un numero maggiore di sub-item.

Le implicazioni nell'ambito dell'innovazione sono che l'approccio Agile può essere utilizzato per raggiungere più velocemente e in sicurezza un risultato medio nel progetto, che implica normalmente un risultato più economico. Inoltre, con la corretta calibrazione di scadenze intermedie e distribuzione del lavoro, il metodo Agile può spingere il raggiungimento di un incremento significativo. Per quanto riguarda il TPM, questo non ha dimostrato di influenzare i partecipanti a focalizzarsi su una specifica richiesta, elemento questo spesso riconosciuto essere uno dei difetti della metodologia, che porta alla formazione di ritardi. Il risultato dell'esperimento è influenzato dalle sue specifiche caratteristiche, e i comportamenti attesi potrebbero avere un impatto molto maggiore in progetti reali. Una spiegazione del perché non si è verificato un picco di risultato nel trattamento traditional può risiedere nel fatto che i partecipanti non erano abbastanza motivati e hanno lavorato con un livello di stress troppo basso. Questo è avvenuto perché nelle caratteristiche di questo approccio non ci sono degli elementi che aiutino a motivare i partecipanti. Significa che, in applicazioni reali, delle sorgenti di motivazione dovrebbero essere auspicabilmente aggiunte con fattori esterni.

In aggiunta all'esperimento, un'altra prospettiva sulla comparazione dei due approcci è data dall'applicazione di un Learning game. Lo scopo principale è quello di insegnare in corsi universitari di project management le differenze degli approcci. Tuttavia, consente anche di valutare più di un aspetto interessante su come le persone si comportano all'interno di un gruppo, mentre svolgono queste specifiche attività. I partecipanti vengono suddivisi in due team, che simulano i due approcci TPM e Agile PM. L'impostazione è simile a quella dell'esperimento come tipo di richieste, utilizzo dei LEGO®, e regole per distinguere i due approcci. Un'aggiunta influente è la figura dei clienti, interpretati da alcuni tra i partecipanti, che interagiscono con i diversi team.

I due approcci hanno dei risultati differenti nei seguenti aspetti:

(a) Management style e comportamenti nella fase di progettazione;

(b) Impatto del Project manager e modo di lavorare nella fase di esecuzione;

(c) Livello di stress percepito durante il gioco;

(d) Impatto del rapporto con i clienti.

In seguito all'esperienza maturata con tre sessioni del Learning game, vengono suggerite alcune implementazioni per la sua realizzazione. Lo scopo è quello di caratterizzare maggiormente le differenze tra i due approcci, e viene perseguito aggiungendo delle specifiche per garantire un maggior rispetto delle caratteristiche chiave. Inoltre, l'interesse è quello di ottenere dei risultati comparabili tra i due team, e delineare un modo per capire le differenze dovute agli approcci. Con questi accorgimenti, il Learning game può venir applicato in maniera più sistematica ed anche costituire un'interessante base per studi sul comportamento dei gruppi in relazione allo sviluppo prodotto.

You can plan for a hundred years,
But you don't know what will happen the next moment.
Indian yogi – Neem Karoli Baba

1 Introduction

The theoretical research on innovation is vast and has interested scholars for many years (Dewar and Dutton, 1986). The matter can be analyzed with many lenses and researchers have discussed the different types of innovation and how to reach them. The project management field of research is also extensive, academic texts and organizations have developed a large set of practices on how to approach project management. During the years opposing paradigms have been proposed, with, in general, the same purpose to reach a more efficient project management approach. Nevertheless, the correlations between innovation and project management have received little attention in the academic literature (Shenhar, 2001). In particular, the comparative research on innovative results obtained with different project management approaches is lacking. Given the increasing popularity of Agile project management, the interest in a comparison with traditional approaches is rising among both scholars and practitioners. Contributing to the comparative research, this thesis work is focused on the product development phase. Within the vast set of procedures that constitute the product development, three main ones are considered. Specifically, the design definition from customer requirements, the planning to realize the project and the plan execution. The distinction between product development approaches is not always clear. Due to the fact that they result mainly from the practitioners' world, there is not a unifying theory and differences often collide. To make the comparison meaningful, in this work two approaches are identified as representative and compared. The comparison is clarified in the following aspects, where differences between approaches are eloquent. (a) approach to planning; (b) relation between planning and execution in how the time is managed; (c) required documentation to support the process; and (d) flexibility to changes during development. Regarding the planning techniques, the traditional approach is characterized by heavy-upfront planning of the overall project. The aim is to predict and prevent every inconvenience that might occur and solve it in advance. This approach is strongly contested by the Agile methodologies, that plan only the next iteration and react to changes. The two planning paradigms are the anticipatory and adaptive project management styles. The planning and execution are clearly separated in the traditional approach, while in the Agile the two phases alternate each other as a natural consequence of the iterative process. Extensive documentation is key to the traditional approach. Instead, the Agile methodologies see it as a not valuable activity. The last point is flexibility. In the Agile framework, changes don't affect the project in its overall and they can be implemented in the following working iteration. In the traditional framework changes often imply rescheduling the whole project.

The comparison between the two frameworks is detailed using the behavioral operations approach to the matter. As common in this field of research, I develop an individual experiment to study two questions, unanswered in the literature: (1) How people behave in different approaches to PM? (2) Do people behaviors affect the results? And how?

The experiment is a real-effort physical task with a time limit and a set of rules to simulate the differences in the two approaches. Participants are asked to build a certain number of structures with LEGO® bricks, they are presented with three main structures that have certain features required. In the following, I call those features sub-items, and the main structures, items. Given the literature review in both product development and behavioral operations fields, I argue that the different approaches to product development may correlate to individual behaviors with consequences for the project performances. On these performances, I formulate the following hypotheses. The first two consider the frameworks separately, comparing the results obtained among the set of sub-items. The third one is a comparative hypothesis between the frameworks, split into two parts.

H1. With the traditional framework, participants reach a significantly higher result in the first sub-item. There is a significant difference between the result of the first sub-item and the others. The results show a decreasing trend from the first sub-item to the followings.

H2. With the Agile framework, participants reach an average result in all sub-items. There is not a significant difference between any sub-item and the others.

H3.1 With the Agile framework, participants reach an average result significantly higher than with the traditional framework.

H3.2 With the traditional framework, participants reach a significantly higher result in the first sub-item compared to the Agile framework

These hypotheses are tested using as data the sub-items evaluation made via an online survey. In addition to the test results, the analysis is supported by the qualitative considerations on peoples' behaviors. This study is the first attempt I am aware of to analyze the effects of different product development frameworks on performance and behaviors. My contributions fall into three categories. First, the experimental results support the research that asserts the Agile approach is less risky in product development and arguably more efficient. At the same time, results contradict the hypothesis regarding the traditional framework. There is no evidence that the framework fosters the development of an outstanding result or brings workers to focus on a single task. In addition, the results open to the possibility that found differences are due to the calibration between requirements and Sprints in the Agile framework. Second, the used approach to the matter of product development could be utilized to deepen the understanding of different approaches implications. The behavioral Operations studies have proved to be a consistent base to study this type of comparisons.

Third, the implementation of a learning game is presented in addition to the experiment. While its main purpose is to teach students the approaches to project management, it contributes to understanding the implications of group behaviors. This experience adds texture to the use of these specific practices as useful teaching tools. If applied in a systematic way in universities, they could constitute also a valid research base for studies on group behaviors in correlation with product development.

One does not begin with answers.
One begins by asking, "What are our questions?"
- Peter Drucker

2 Research

The questions that motivate this thesis work are the following.

1) Why innovation? And which type of innovation?
2) How to manage innovation?
3) Different approaches to project management (PM) produce different innovative outcomes?
4) How people behave in different approaches to PM?
5) Do people behaviors affect the results? And How?

The first two questions are very general, but they help to collocate this work correctly in the research field. The third question is the one not entirely answered in the literature. The purpose of this work is to add a contribution to it by answering the last two questions.

This chapter presents the literature review on those aspects, pointing out where more knowledge would be needed. Section 2.1 introduces the innovation issue and how companies manage it at different levels. Section 2.2 presents different approaches to PM. Section 2.3 compares the approaches under the innovation lens. Section 2.4 presents the relevant concepts from the Behavioral Operations (BeOps) field and formulates the hypotheses for the PM approaches.

2.1 Innovation Complexity

Innovation is one of the critical factors that determine the health and success of a business organization. All organizational levels have to understand and deal with innovation's relevance. Underestimate its power can lead to dramatic results, even for well-established companies. In the infamous case of General Motor (GM), a myopic attitude to innovation was one of the long term causes that brought GM to lose its leading position in the car industry. Indeed, Sloan (1990) wrote about his years at GM that "…it was not necessary to lead in technical design or run the risk of untried experiments [provided that] our cars were at least equal in design to the best of our competitors in a grade." Womack et al. (1990) argue how this approach – common between western mass producer at that time – produced an organization of the firm that wasn't able to reach startling innovations in any case. The authors explained that GM understood this mistake only in the '90s and still lost its leading position years later. From this approach to the matter, many steps ahead were made, and extensive researches have been done.

When discussing innovation, it's necessary to clarify which type of innovation is considered. Regarding the results achieved by the innovation and the degree of newness that implies, the widely accepted distinction is between radical and incremental innovation. Dewar and Dutton (1986) proposed a neat distinction, arguing that radical innovation can be reached only with high technical levels and is easier reached by large firms. Instead, the authors argued that every type of firm could achieve incremental innovation. Current research still utilizes the distinction between radical and incremental, but the perspective has changed. The focus has shifted from mere technology to satisfying the customer. New branches of research on innovation are due to the changing environment that companies have to face. Brown and Katz (2011) explain how the twentieth-century market was driven by the companies, that created new products and consumers passively consumed them. Instead, in today market the focus is shifted to clients and their needs. Different approaches have been suggested by practitioners and academics to deal with the current situation. Kim and Mauborgne (2014) focus on discovering the hidden value of a product or a service, looking in the chain of actors involved or in the boundaries between sectors. They suggest a set of techniques to discover blue oceans, that means a sector where the competition is not relevant anymore. Ulwick (2016) highlights the importance to understand the real customers' needs, that should be done with a deep analysis of their "jobs to be done". There have been some seminal works on the different types of innovation sources, and the need to understand that producers' model innovation is not the only one anymore, an "increasingly important model is the open user innovation" (Von Hippel, 2010). Innovation cannot come anymore only from internal research, "in a hugely interconnected world, isolationism stifles innovation" (Chesbrough et al., 2006). Within this fertile environment, the Agile approach inserted itself, with its founding principles in focusing on the customer and working - valuable products, that are developed in iterations and frequently delivered (Beck et al., 2001).

After a breath on the types of innovation, the following point is about how to manage and incorporate innovation in an organization successfully. As previously mentioned, all levels need to implement innovation. This request is due to an integrated nature of the matter, that implies a high complexity in its management.

This thesis work, with its experimental setting, is focused on the product development phase. Before concentrating on this phase, it's needed to collocate it among the set of activities done to manage innovation. Biazzo et al. (2016) developed a summarizing model to gain an overall vision of the innovation processes: The innovation pyramid model. It is based on a three levels system of activities: absorb, explore and create. The first - absorb level incorporates all the activities aiming to get knowledge from the external environment. The second – explore level refers to looking for innovation opportunities, not only Research and technological experimentation but also the world of Open Innovation. The third – create level is the one where a new product idea is transformed into a marketable product ready to be profitably produced. The product development activities are done at the third level and are highly dependent on the lower levels of this pyramid model.

2.2 Product Development

The previous section collocates product development activities in innovation management correctly. In the following the focus is on product development (PD), and specifically new product development (NPD). The product development in its complexity requires a mix of activities that are controlled among three decision areas: program management, project portfolio management, and project management.

About project management the literature and practitioners' knowledge are extensive. Fundamental books treat the argument in detail. The PMBOK® Guide (2008) is "the standard for managing most projects most of the time across many types of industries." It defines project management as "the application of knowledge, skills, tools, and techniques to project activities to meet the project requirements." The PMI (Project Management Institute) identifies five process groups for the PM activities: (1) Initiating; (2) Planning; (3) Executing; (4) Monitoring and controlling; and (5) Closing. Within this broad field of research, different approaches to project management have been developed and applied. The interest here is to present and compare two of them. traditional project management (TPM) and modern project management (MPM) identify the opposing paradigms. In particular, for the TPM the waterfall method is considered, and for the MPM the Scrum application of Agile project management is utilized. Spalek (2016) gives a literature review on comparing traditional and modern approaches to PM, that dates back to the 1960s the scientific approach to project management. In those initial years, the characteristics of projects in terms of complexity and costs brought to rigid planning and control of the project (Kerzner, 2013, Wyrozebski and Spalek, 2014). Feng and Sedano (2011) demonstrate that the waterfall approach was widely applied for managing projects. This long-time application brought researchers to universally recognized that approach as the traditional project management (Hebert and Deckro, 2011, Pellegrinelli, 2011). Modern project management approaches were developed to react to the twenty-first century changing market (Curlee, 2008, Shenhar, 2001). The new conditions require to companies to reduce life cycles for product development, gain augmented responsiveness and be aware of customers' requests (Kach et al., 2012, Liberatore and Pollack-Johnson, 2013, Relich, 2015).

2.2.1 Traditional – Waterfall Project Management

Salgado and Dekkers (2018) in their work of comparing different approaches to PD, identify three main articles that define the waterfall approach (Bassler et al., 2011, Bullinger et al., 2003, Joore and Brezet, 2015). The name "waterfall" comes from the seminal paper of Royce (1987), and it's due to a diagram presented by the author. Royce was reporting his experience in how to manage the development of a large computer program for delivery to a customer. The "waterfall" diagram presents the implementation steps needed, with the primary purpose to underline the differences with small program development. The steps are shown in a one-way dependent sequence that pictures a waterfall. Royce highlighted clearly that a mere implementation of that models is risky and invites failure because each phase produces effects not confined only to the following step. In the following of the paper, Royce presents five additional steps to eliminate most of the development risk. Those implementations characterize the correct application of the waterfall approach. The primary purpose is to uncover in advance all the possible problems and solve them before the test phase. The practices to reach that goal are: do the design before beginning analysis and coding, produce complete documentation and build a pilot model. Even with these rigorous steps in the early phases of the project, the author recognizes the uncertainty level present in the process, and to deal with it suggests complete testing and collaboration with the customer.

The traditional approach constitutes the basis of the rational paradigm to PM, that is characterized by a clear separation between the planning and the execution phases. To achieve that division the project needs to be planned and decided in advance, that requires a focus on rigid procedures and techniques utilized. Some well-known tools have been developed to support those procedures, like the Work breakdown structure, Gantt charts, critical path method, and the program evaluation and review technique – PERT (Eppinger 2001, Kim and de la Garza 2005, Makhloof et al. 2014, Zang et al. 2013). This rational approach is grounded in a vision of how people work that Sterman (2000) called Open-loop thinking, where the steps from the identification of a problem to its implementation are linear. Sterman argues that real complex systems evolve in a very different way, due to the correlations between the various elements the process is never linear. Petersen et al. (2009) produce a case study to analyze the commonly recognized issues of waterfall development. The main arguments are related to the responsiveness to change, the generation of a lot of rework and the uncertainty on final quality due to the late testing phase. Petersen summarized the literature and identified nine issues in waterfall development. Between those arguments, there is a lack of customer feedback and increasing lead-time due to the need to approved large artifacts at each gate of the project.

2.2.2 Agile – Scrum Project Management

Erickson et al. (2005) agreed to date back the beginning of modern project management to the presentation of the Agile Manifesto in 2001. The Manifesto set the principles of what has been called the relational approach to project management, that emphasizes the importance of working products and customer satisfaction before documentation and processes. Agile practitioners introduced these core concepts in software development, but the ideas behind them date back to previous practices developed in some Japanese and US companies. Takeuchi and Nonaka (1986) discussed a method utilized by successful companies "in contrast with the traditional sequential relay race." They described the approach using rugby's scrum as a metaphor. Six characteristics for NPD processes are presented: (1) Built-in instability; (2) Self-organizing project teams; (3) Overlapping development phases; (4) Multi-learning; (5) Subtle control; and (6) Organizational transfer of learning. Implementing these principles, Sutherland and Schwaber presented at the OOPSLA conference in 1995 their Scrum approach. They developed this approach solving the problems they encountered working with waterfall management. Their main arguments against that approach are in terms of delays, rising costs, and unnecessary complexity. Sutherland and Schwaber (2017) wrote The Scrum Guide, that is the body of knowledge on the argument, it clarifies that the essence of Scrum is a small team of people and the three founding principles are transparency, inspection, and adaptation. Sutherland (2014) argues that the Scrum approach implement the natural way people do their work. This method is one of the most used in APM practices. Highsmith (2002) highlights in general that the practices introduced by APM are generative, and not prescriptive, meaning that they don’t describe every activity the team should do. The fundamental concept is to identify the practices that have an extremely high value and use them on nearly every project. Those practices create the starting point, from there on the project team will generate all that is needed for specific situations and needs.

2.3 Innovation in Product Development

Previous sections presented the two different approaches to manage NPD. The following section discusses the literature on how different approaches produce different innovative outcomes.

The comparative research on this subject is not extensive, while the relation between innovation and each approach separately has been discussed by many authors. Since "APM principles are similar to Lean Thinking principle" (Smith, 2005), I consider here also studies in the lean field. Regarding innovation at Toyota, Womack et al. (1990) reported that the activities done in the NPD are very successful in reaching incremental innovation and efficiency in product development. Nevertheless, they argued that the approach could have fallacies in reaching a disruptive innovation. Pichler and Schulze (2005) in their book review argue about different approaches to product development, underlying that APM authors “envision shifts toward greater agility to cope with growing uncertainty in markets, technologies, customer perceptions, and management direction.” In the book “Agile project management: creating innovative products” Highsmith (2009) discusses the characteristics of a reliable innovation reachable with APM. Five key objectives define a reliable innovation: (1) continuous innovation; (2) product adaptability; (3) reduced delivery schedules; (4) people and process adaptability; and (5) reliable results. The author explains that to meet today's customer requirements, a mindset that fosters reliable innovation is needed. The approach of APM is made to succeed in complex and turbulent systems. In doing so there is a high focus on the principles, that are mainly about people, iterations and working product.

The Scrum approach has the same purpose. Schwaber and Beedle (2002) identify Scrum as the answer to transform an idea in something useful, in a chaotic and complex area, without losing money and time. In the Scrum guide, Sutherland and Schwaber (2017) report that "Scrum proved to be especially effective in iterative and incremental knowledge transfer." One of the fundamental characteristics of the framework is its research to reach continuous deliver. The Scrum approach proved to be successful in fast-moving markets for numerous software development projects.

Spalek (2016) concluded from his studies that TPM methods have a broad application in companies, “However, in modern turbulent environments, they seem to be insufficient according to the new challenges organizations are facing.” He evaluated the answer many companies adopted is modern project management methods. Nevertheless, there is a discussion about the real contribution of APM in companies' success and if it's possible to run projects only with the Agile approach (Serrador and Pinto, 2015).

The research investigating the correlations between processes and radical innovation is still lacking, and the available studies focused on technology-driven radical innovation. The research of Vojak et al. (2012) analyzed this gap studying the so-called "Serial innovators," that are individuals that produce systematically radical innovation in large, mature companies. The authors discussed how the formal NPD processes (identified with the Stage-Gate® processes, that can be compared with the TPM) "may impede Serial innovators' ability to innovate effectively." The answer to which approach to NPD allows to reach radical innovation is still unclear and should be debated regarding the type of project and environment (Kuchta and Skowron, 2016). Research about how to manage radical innovation has been done at other levels than the PD, mainly to discuss strategy issues (Mcdermott and O'connor, 2002). Few theoretical studies analyzed the implications of different PD processes on innovation, Arrichiello et al. (2014) investigated how systems engineering affects innovation.

This thesis work contributes to the open question of how PM approaches affect project outcomes in the field of Behavioral operations, applying an experimental method. The remainder of the literature research presents the relevant aspects to support the development of the experimental setting.

2.4 Behavioral Operations Perspective

Previous sections have discussed the reasons why innovation in product development is worth to be analyzed, and the opposite approaches used to deal with it. The following sections present the foundational elements of Behavioral Operations and correlated applied research. These concepts are used to formulate the hypotheses on how people behave in different PM approaches and the effects of those behaviors on project results.

The Behavioral operations field founds its value in “recognizing that almost all contexts studied within operations management contain people” (Croson et al., 2013). The authors give a definition of the field that implies: (1) Study of potentially nonhyper-rational individuals; (2) An operational context; (3) A behavioral context that consider various patterns of human action, and not only the one devoted to a single monetary goal; (4) Unit of analysis constraints to the micro-level. Bendoly et al. (2015) answer to the question about how much knowledge of BeOps is necessary for the field, underlying the basic fact that people are essential to operations. The implication is that to take effective decisions in real complex systems, all the available knowledge of individuals' behavior is required. BeOps research analyzes the decisions and behaviors of individuals and small groups of individuals that define the micro-level collocation of the field. Given the fact that project management and product development are activities highly influenced by human judgment, they are primary subjects of the Behavioral studies.

2.4.1 Individual Decision Making

In this thesis work, an experiment is developed to compare PM approaches. The following part of the research is useful also to collocate it correctly in the field. The interest is on intra-organizational dynamics, the branch that studies the characteristics of internal operations. Given the empirical landscape, the preexisting theory of other sectors is needed to a complete comprehension. The experiment developed in this thesis involves individuals working alone, that implies the use of concepts from cognitive psychology. This psychology branch recognized that "cognitive limits lead to limited (and often critically flawed) mental models of works contexts" (Bendoly et al. 2015). In the individual decision-making processes the BeOps research distinguishes (Donohue et al., 2018):

- Heuristics: "methods through which solutions are arrived at";
- Biases: "lenses through which problems and solutions are viewed."

In the following these concepts are analyzed as relevant for the experiment developed:

(1) Anchoring and insufficient adjustment heuristics, and planning fallacy;
(2) Procrastination and Parkinson's law;
(3) The macro phenomena of the Behavioral hill;
(4) Effects of multitasking and available information.

Donohue et al. (2018) explain the individual decision-making process as a loop between three elements: "Motivation/Stress – Biases/Heuristics – Perception/Mental models." Regarding Biases and Heuristics, some fundamental studies in analyzing judgment under uncertainty set the foundations of the field. Tversky and Kahneman (1974) presented the anchoring and adjustment heuristics summarized as "different starting points yield different estimates, which are biased toward the initial values." This heuristic affects the general approach people tend to have in estimating a quantity or a value. The authors evaluated that under time pressure the tendency is to start from a known point used as an anchor, and then make an estimation by extrapolation or adjustment. Because adjustments are found to be insufficient in most cases, this procedure should lead to an underestimation. The experiment ran by Aranda and Easterbrook (2005) demonstrates the direct implications of this heuristic within estimations in software development. They proved a "too strong to be ignored effect" on results due to different initial anchors.

Among product development activities an essential step is planning, that is usually the result of intuitive judgments and educated guesses. Kahneman and Tversky (1977) presented biases and corrective procedures related to intuitive judgments. The authors identified two common biases in forecasting activities: non-regressiveness of predictions and overconfidence in the precision of estimates. They evaluated that people, under conditions of uncertainty, tend to have an "internal approach" to predictions. This approach defines the so-called "planning fallacy." It implies a clear tendency to focus more on specific problems and neglect distributional data in similar cases. Optimism leads to thinking that the current project will follow the plan regularly, even though a critical analysis of past similar projects would advise that most of them failed to respect the deadlines. In addition to optimism, subjects make errors due to the anchoring effect and neglect to consider the mistakes in the forecasting cycle time (Tong and Feiler, 2016). Researchers have worked on possible ways to contrast the planning fallacy. The commonly accepted solution is the need for an outside overview, that should aim to learn from past-similar projects when planning new ones.

Kahneman and Tversky (1977) already suggested that the framework characteristics of the decision making should facilitate the use of all information, to overcome the planning fallacy. One studied approach is to ask the individuals to recall for past experiences to make them conscious of what happened before, and it demonstrated to improve estimations' quality (Lovallo et al., 2012). Other researches demonstrated a similar effect. Kruger and Evans (2004) conducted a study on people's private plans and demonstrated that divide main tasks in sub-tasks increase the quality of the estimations. These divisions are typically done with tools such as the work breakdown structure, that is also largely applied in traditional project management.

In Agile teams, many different techniques are used to make reliable estimations, such as the planning poker. These techniques have the purpose of making the team responsible for the planning and help to reach an "outside" point of view, and they are also a successful way to overcome the planning fallacy.

Donohue et al. (2018) suggest some lenses through which examine behaviors of actors working in project management: psychological safety, multitasking, procrastination, Parkinson's law, and the role of information. For the experiment developed in this work, psychological safety is considered not to be relevant. Regarding multitasking, researchers argued that it could lower performance compared with the sequential execution of tasks (Buser and Peter, 2012). This argument is in contradiction with the common beliefs that address multitasking as an effective way to reduce cycle times and improve efficiency. Procrastination, also called student syndrome, and Parkinson's law are consolidated behaviors, known for many years and studied in different fields. Parkinson's law claims that "work expands so as to fill the time available for its completion" (Parkinson and Lancaster, 1958). This phenomenon implies that all the time available for an activity will be used, with rare cases of early completion. The student syndrome goes along with that behavior. Researchers demonstrated that projects are highly affected by this combination. Researches have been done to understand how to control procrastination, and both internal and external imposed deadlines don't show a positive impact on task completion (Bisin and Hyndman, 2014). Wilcox et al. (2016) analyzed the effect of keeping people busy and found a positive correlation with the task completion rate.

2.4.2 Traditional and Agile hypotheses

The above-presented concepts allow formulating the first two hypotheses on how people behave in different approaches to PM. In the experiment developed participants experience one of two treatments. The treatments simulate the characteristics of traditional and Agile PM. In the following, I refer to them as the "traditional framework" and the "Agile framework," to distinguish when I consider the experiment setting. Participants are presented with a list of products' features that they are requested to build using LEGO® bricks. In the following, I refer to these features calling them "sub-items." Each sub-item built is evaluated, and this evaluation constitutes the result achieved by the participant. The first two hypotheses consider the frameworks separately, comparing the results obtained among the set of sub-items.

H1. With the traditional framework, participants reach a significantly higher result in the first sub-item. There is a significant difference between the result of the first sub-item and the others. The results show a decreasing trend from the first sub-item to the followings.

The motivations for the stated hypothesis are due to the correlation between the individual biases and heuristics and the characteristics of the traditional framework. The traditional framework is based on a sequential way of working, that separate the planning and the execution phases and implies ex-ante planning of the entire project. The participant is required to prepare a WBS with the schedule of the tasks she will complete during the experiment. The anchoring and adjustment heuristics has a substantial effect on this type of process. At the beginning no information for the planning is given to the participant, it means that she will use her personal anchors to do the planning. With no previous specific experience on the tasks to do, the estimations are likely to be far from optimal. Besides, planning for what is the long-term in this scenario, makes the request harder. During the execution phase, there are no externally imposed deadlines. The participant has her internally imposed deadlines of the WBS. Optimism, procrastination and Parkinson's law should impact the way the participant works in the following way. Starting to work on the first sub-item, the participant uses all the time planned for it (Parkinson's law). Due to the planning fallacy, that will be probably higher for the first task, the time scheduled for it won't be sufficient. The participant will tend to not respect the deadline, due to the effects of the student syndrome and optimism. The result is more time than planned spent on the first sub-items, that will leave less time for the last tasks. This behavior should impact the sub-item results, with higher results on the first sub-items, particularly the first one, and lower for the following ones.

H2. With the Agile framework, participants reach an average result in all sub-items. There is not a significant difference between any sub-item and the others.

In the Agile framework, the work is organized in iterations that the participant plans one at the time. The planning is made each time only for the following iteration. The whole process is time-boxed, and it's imposed the time respect for the different phases externally. Besides, the participant is required to finish all that she previously planned for the iteration. These characteristics should reduce the power of procrastination and Parkinson's law. They still affect the participant within the iteration, but the additional pressure imposed should force the subject to overcome them to respect the rules. The implication is that the participant shouldn't spend more time than planned on any sub-item, producing no result significantly higher than the others. After each iteration, the participant is required to evaluate the work completed. This moment involves a "think aloud" reflection and forces the subject to take consciousness of how the process is going. As explained in the literature, it should reduce the planning fallacy for the following iteration. The framework requires the participant to consider the distributional data on the project. This should increase the planning efficiency, and again avoid that the participant doesn't respect the plan and spend more time on specific sub-items.

2.4.3 Comparative Hypothesis

In the following, other concepts of the BeOps field and related research are examined to formulate the third hypothesis regarding the two frameworks.

The above mentioned cognitive psychologic concepts are not isolated from one another. Their influences mix with other factors in the Operations environment and create some specific macro phenomena. Bendoly et al. (2015) summarized the composite phenomena identified in BeOps. The behavioral hill is one of the most studied recently and has implications for this work. Bendoly and Hur (2007) presented a comprehensive discussion to explain the phenomena. The name comes from the inverted U shape, that the authors found in the correlation between the challenging level and motivation. It means that there is a peak of motivation due to the challenging level, so the correlation is not monotonic. The authors discuss that the result is due to the bipolar effect of certain motivators. These factors, motivation/stress/work excess, balance each other producing different outcomes in response to the inputs. With an increasing challenge, the motivation grows. Instead, at low challenging levels, Parkinson's law explains why motivation and productivity are lower. This effect is valid until a certain level, that is in general different for each person, where motivation and performance reach a peak. Above this challenge level, the individual's capabilities are exceeded, and the excessive challenge stresses the individual lowering the performances. Bendoly and Prietula (2008) analyze the effect of training on motivation and related performances, demonstrating that training and experience have not always a positive impact on results. The authors observed a typical inverted U in the correlation between results and different input levels, meaning that after a certain degree more training or experience produce worse outcomes. They also found that the length of the work queue affects the motivation level, the work completion rate, and the quality of results. These correlations change at different skill levels, with a counterintuitive tendency for the high skill level. In this case, longer queue means higher motivation, completion, and quality with a monotonic correlation. It says that for individuals with high skills, a perceived higher challenge has a positive impact.

Donohue et al. (2018) summarized the set of decisions and behaviors related to the specific process of developing new products, from conception to execution. The aspects relevant for this thesis are connected to the planning and execution activities, so the behaviors in the conception phase won't be considered. The factors that play a role in planning behaviors are mainly: incentives, motivations, the process itself and the degree of uncertainty. The individual is affected by those factors and develops some specific cognitive processes. The main point is that the subject responsible for the planning need to take critical decisions that involved trade-offs between three summarizing project goals: time, budget and scope. In that environment, planning is highly impacted by the individual characteristics of those carrying it out. Thus, there has been some research to find out who might be an effective planner (Mumford et al., 2001).

Research has been done to study the implications between the planning and execution phases. Choo (2014) demonstrated a U-shaped correlation between the time spent for problem definition and the project duration. This correlation shows that more time spent in the first phase produces a positive effect in reducing the following project duration. However, the correlation is not monotonic, and there is a certain amount of the time spent in planning that assures the lowest value of project duration. With time exceeding this amount, the effect is counterproductive. The study concludes that it is relevant to find the right balance in the first phase to avoid counterproductive effects. A limitation of Choo's research is its focus on Six Sigma projects. It would be desirable to study if these results can be generalized to behaviors in a general product development process. Kagan et al. (2017) analyze with an experiment setting the correlation between the time spent in the ideation phase and the results achieved. The study outlines that there is not a significative difference between different exogenously imposed transition times. Indeed, imposing constraints to the subjects outperformed the endogenous treatment, where participants are free to choose how to manage the time transition. The research also outlines that the number of ideas did not consistently predict performance, implying that quantity doesn't always mean quality.

One main difference between this experiment treatments is the relation between problem definition and project definition, how time is managed and split between those phases. So, the application of the presented researches will be testes.

Some studies give direct implications to PM approaches. Bendoly et al. (2014) suggest to use the elements implemented in Agile product development to reduce the negative effects of task switching. The suggestions are focused mainly on using short and modular tasks. Aranda et al. argued that development "lifecycles such as the spiral model or incremental development are safer than others like the waterfall model." The higher risk of the latter is due to the weight given to deadlines in traditional models. Bendoly and Swink (2007) focused on the effect of information in a multi-project resource management setting. They found that with high levels of uncertainty a lack of information can lead decision-makers to not optimal decisions. The take away is the need for organizational structures that increase information availability and process visibility to gain sufficient transparency.

The discuss literature brings to the definition of the third hypothesis. This one is about the comparison of the two frameworks' results. The hypothesis is split into two parts.

H3.1 With the Agile framework, participants reach an average result significantly higher than with the traditional framework.

The effect of the behavioral hill presented should be in favor of the Agile framework. Given the time-boxed iterations, the perceived challenge by the participants is higher. The effect should be that in each iteration the participants have a higher motivation that allows them to reach higher results. In the traditional framework, the participants are free to organize their time, and this doesn't add any challenging effect. The effect of the work queue is related to the skill level of the participants. Given that participants don't train on the specific tasks before to start, I would argue that their skill level is on average not high. This would mean that a longer queue lowers the motivation and completion rate. In the traditional framework, the work queue is the longest possible for this setting. While in the Agile framework, participants have at each iteration a short queue made of what they planned for that iteration. Besides, the exogenously imposed transition times present in the Agile framework should allow the participants to reach higher results. The characteristics implemented in the Agile approach are found to create a safer product development process. It should imply that participants are less likely to fail or to have problems in satisfying the requests. The combination of these effects should allow agile participants to reach higher results on average.

H3.2 With the traditional framework, participants reach a significantly higher result in the first sub-item compared to the Agile framework

Given the randomization of the participants in the two frameworks, the differences in results should be related only to the frameworks' characteristics. Therefore, the traditional hypothesis H1 implies the same difference between the two frameworks. Participants with the traditional framework should reach a significantly higher result on the first sub-item, also in comparison with the agile participants.

3 Experimental Design

To approach the research questions, I develop an individual laboratory experiment. The purpose is to compare the identifying characteristics of the waterfall and Scrum approaches. As discussed, the management of a NPD process is complex and articulated. In real applications, many complexities are due to people and team management. Clearly, using this individual experiment approach the limits are due to the extreme simplification in comparison to real projects. Nevertheless, this approach allows focusing on how people work to understand if the introduced differences have an impact on that. I analyze a more complex situation, in terms of influencing factors, in chapter 5, that presents the application of a learning game.

Regarding the experimental design, I need to first discuss the characteristics that define the experiment and its purpose, this is done in section 3.1. Given the focus on frameworks, their characteristics are explained in detail in section 3.2. Section 3.3 presents the requests that are asked to the participants and discuss their effect on the results. Section 3.4 gives details of the specific activities done during the experiment sessions.

3.1 Experimental Setting

The experiment is a real-effort physical task with a strategy space limited by some initial requests. The participants are all asked to build the same structures using LEGO® bricks. In doing so, they have to follow a set of rules, that identifies the approach to PM. There are two treatments that simulate traditional and Agile frameworks. LEGO® bricks are used in this experiment because they allow participants to easily build something and be creative in doing so. In addition, the bricks ensure, with a good enough probability, that participants do not lose interest during the experiment. The general characteristics of the setting simulate the following relevant aspects of product development:

1. Fixed launch date;
2. Multiple products to be developed simultaneously;
3. Final customers evaluate the products;
4. An individual decision maker;
5. Design a product starting from not too specific customer requirements.

These conditions are generally true for the majority of business sectors, where there is a competition that implies certain time to market. Participants have a fixed time of one hour to satisfy the requests of a customer. As mentioned in the previous chapter, the requests are presented as sub-items. There is a total of seven sub-items, that constitute three main buildings. In the following, the set of requirements is referred to differentiating between the three items and the seven sub-items. While working, people have to independently organize their time among sub-items and items, as normally happens in projects where workers are required to work on different tasks. The individual decision maker or a team leader that takes decisions for the project is a common situation. The product development phase starting from customer requirements is an implication of having different levels of innovation management. Products' specifications come from the effort done at the exploring and absorb levels. The requirements could also simulate a product developed started on a specific order of a client. The analysis of the experiment is done in two different ways. During the experiment, I observed each participant to evaluate the behavioral aspects. In addition, the buildings, and specifically each sub-item, are photographed and evaluated by a third party. These evaluations constitute the basis to discuss the approaches' performances and to test the previously presented hypotheses.

The main general decisions for the experimental design are discussed in the following. The underlying general idea is to keep the setting simple in its variables, due both to the resources available and the qualitative/explorative nature of the experiment. The experiment was organized at the TUM SoM[1] and it was divided into individual sessions for each participant. The subject pool is made of 20 subjects, they are all students of TUM except for three LMU[2] - medicine students. The average age of participants was 23.6. It was not possible to pay each participant, so the experiment was presented as a contest with a monetary reward just for the winner. The total number of participants was decided in advance, given the difficulty to recruit more than about twenty people and the time length required for the experiment. It took two working weeks to run all sessions. The treatments are different in many characteristics and constitute the only dimension analyzed in the experiment. It means that the comparison is held between the frameworks as they are. An additional dimension should be analyzed in future implementations, to understand which frameworks' factors affect the results and how. In this experiment, it is used just one dimension due to the small sample available and the willingness to not add variables that could

[1] Technische Universität München – School of Management

[2] Ludwig Maximilian Universität

complicate unnecessarily the setting. This should be seen as the first exploration of experimental differences between the two frameworks, to understand if the matter is worth to be analyzed in this way.

It is used a between-subject design, that means each subject experience only one treatment. It would be interesting to use a within-subject design, but it is not doable in this case because it would require participants to spend more than two hours for the experiment. The priority is given to have participants working for an amount of time long enough to appreciate the framework's characteristics. To achieve randomization is applied a factorial design, that in this case is obvious given only two treatments. The subjects are split randomly between the two treatments. The result is ten measures for each treatment. Given the little number of participants, I was able to monitor each participant. My presence there could have influenced the participants' behavior. Nevertheless, the priority was to gain direct and qualitative insights on individual behaviors. Afterward, I am confident to say that there has not been any sign of biases due to my presence, this was due mainly to the not formal environment. Croson et al. (2013) argued that observed behavior in the laboratory provides only limited information about the processes of human judgment and decision making. An available alternative would be to utilize a verbal protocol analysis. Given the specific setting, that approach would influence significantly the individuals' behavior. An examined difference between treatments is how subjects realize and evaluate how the project is going. Having people express verbally their decisions might interact with the framework's characteristics and produce biased results. More importantly, it would arguably attenuate the differences between the treatments.

3.2 Frameworks Description

Frameworks' characteristics simulate the PM activities identified by the PMI, previously described in section 2.2, that are: (1) Initiating; (2) Planning; (3) Executing; (4) Monitoring and controlling; and (5) Closing. It's fundamental to highlight again that those processes follow that numeric order just for the TPM, in Agile PM the alternation has a completely different nature. The translation of these activities in the experimental setting means that participants are not free to just build as they prefer, but they are forced to respect the steps as they would do in a real project. This is ensured using a set of rules that structure the process and identify the treatments' differences.

The comparison between the two paradigms is made specific under these aspects: (a) approach to planning; (b) relation between planning and execution in how the time is managed; (c) required documentation to support the process; and (d) flexibility to changes during development. For the elements to use in simulating these aspects experimentally, I took the basic idea used by Siemsen[3] in his teaching method on PM, that in turn took inspiration from the learning game Lego4Scrum[4]. In general, to make the treatments comparable the same net building time is imposed on both frameworks. To reach that, different phases have a precise duration that participants have to respect. To decide which characteristics from Agile incorporate, I consulted the Scrum guide™ (2017). While for the traditional approach I used very basic common practices from different papers and teaching material. In the following specific characteristics of each framework are presented.

3.2.1 Traditional – Waterfall Framework

The basic principle of TPM is the net separation between design plus planning and execution. It is achieved with an initial heavy upfront planning, that implies deciding in detail everything that will be done at the beginning. To allow participants to do this in the experiment, the supporting tools are an excel spreadsheet where to do the Work Breakdown Structure and standard documents to define the sub-items design. The rules define the two main phases:

1. Design and Planning – 15 minutes

   Participants use this time to read the customer requirements and decide the design of each sub-item. They have to compile a standard preliminary design form with details on main features, colors, and dimensions for each sub-item. Besides, they have to decide on a plan for the whole building time. They are asked to write at least one task for each sub-item specifying the start and end time. The excel sheet shows them the WBS plan that updates with the time running and points out at every moment what they should be doing according to the plan.
2. Building – 40 minutes.

   Participants build following the WBS plan they wrote.
3. Design and plan changes – 5 minutes

   In addition to the building time, participants can use this time to make changes to what they decided in the first phase, both for the design and the plan.

[3] E. Siemsen reported on this exercise during 2017 INFORMS annual meeting (Houston, Texas. October, 2017).

[4] Teaching practice used to explain to practitioners how Scrum works. Available at web-site www.lego4scrum.com

The basic rule is that, at every moment of the execution phase, the participants have to respect the design and WBS. It means that they are not allowed to build something not schedule for that moment, they have to be always in agreement with the plan. To make participants respect it and at the same time have a comparable net building time with the Agile framework, I introduced the five additional minutes for changes. Participants can use them in the following way. To work on a task that is not scheduled for that moment, they must stop building and change the WBS according to what they want to do. The only exception is for delays on a certain task. If a participant is building something but realize the scheduled time is not enough and she wants to continue working on that, she is free to do it. But prior to moving on to the next task, the participant has to correct the time spent on the activity, basically recording the happened delay. Participants have also another option to change the plan: scheduling a rework. They are presented with two example cases when they might want to use it. If they evaluate that more time is needed for a certain task, but they want to complete it in a later moment they can schedule it. Besides, if at a certain moment a participant decides that a previously considered done sub-item requires changes, she can schedule a rework and change it.

3.2.2 Agile – Scrum Framework

For the Scrum framework, the typical events and artifacts are used. Before to go in detail with characteristics and rules, I mention the definitions of elements took from the Scrum Guide™ (2017) that are applied in the experiment. The Scrum events are time-boxed, such that every event has a maximum duration. “Once a Sprint begins, its duration is fixed and cannot be shortened or lengthened.”

- The sprint is the heart of Scrum, a time-box during which a "Done", useable, and potentially releasable product Increment is created. Sprints have consistent durations throughout a development effort. Each sprint has a goal of what is to be built, a design and flexible plan that will guide building it, the work, and the resultant product increment.
- The work to be performed in the Sprint is planned at the Sprint Planning. Sprint planning is time-boxed. During this event, it has to be decided what can be delivered in the next Sprint Increment and how the work to do it will be achieved. A Sprint Goal is also created, that is an objective set for the Sprint that can be met through the implementation of the Product Backlog.
- A Sprint Review is held at the end of the Sprint to inspect the Increment and adapt the Product Backlog if needed. The result of this event is a revised

Product Backlog that defines the probable Product Backlog items for the next Sprint.

Scrum Artifacts:

- The Product Backlog is an ordered list of everything that is known to be needed in the product. It is the single source of requirements for any changes to be made to the product. Product backlog items have the attributes of a description, order, estimate, and value. They often include test descriptions of items that will prove their completeness when "Done."
- The Sprint Backlog is the set of Product Backlog items selected for the Sprint, plus a plan for delivering the product Increment and realizing the Sprint Goal.
- The increment is the sum of all the Product Backlog items completed during a Sprint and the value of the increments of all previous Sprint. At the end of a Sprint, the new Increment must be "Done", which means it must be in usable condition and meet the Scrum Team's definition of "Done."
- The definition of "Done" implies that team members must have a shared understanding of what it means for work to be complete, to ensure transparency. This definition is used to assess when work is complete on the product Increment.

In the experiment framework the one-hour total time is divided in:

1. Define the Product backlog – 12 minutes;
2. Work in four sprints of 12 minutes each – 48 minutes.

   Each sprint is divided in:

   - Previous Sprint Review and next Sprint planning – 2 minutes;
   - Building time – 10 minutes.

The participants work with the support of an online working platform called Atlassian[5], here it is used the term "issue" to identify a virtual ticket or post-it with something needed in the product. On the site, people can create and define the Scrum artifacts and manage their work. While defining the Product Backlog participants are asked to create at least one issue for each sub-item. For each one, they have to decide its main features and do a time effort estimation. The description they add will constitute the definition of Done. The Sprint Review constitutes in deciding which issues are completed and which ones are still in progress. The Sprint planning is made by moving issues from the Product Backlog to the Sprint Backlog.

[5] www.atlassian.net is a platform used by Scrum teams.

3.3 Customer Requirements

The characteristics of the customer requirements presented to the participants influence many aspects of the experimental purpose. The items and sub-items were namely the followings:

1. high rise building: (a) nice rooftop (b) big main entrance
2. residential house: (a) at least two floors and four windows (b) front yard (c) separate garage
3. castle: (a) at least two towers (b) drawbridge

The choices made for the requirements have two main purposes: keep the treatments comparable while allowing to relevant differences to show. The requirements' characteristics that need to be discussed are their level of correlation, the total number, the detail level and how they are presented.

Regarding the correlation between requests, the choice was to use separate items, that simulate the development of different products that don't have implications on one another. In a real case scenario, the items could be independently commercialized. The implications could not apply to a situation where different project's parts are strictly correlated to one another. The results' evaluation is independent, so that differences can be seen among them. The three items are designed to require approximately the same effort, and the sub-items of each one should balance the time required. The detail level of requirements is also a matter of discussion because it could affect the innovation level potentially reachable. This choice has to balance between the necessary free space to reach an innovation and the need to have comparable results. The correlation between requirements' specifications and the aim for innovation results is still an open discussion. Nevertheless, one aspect commonly accepted is that requirements should not be neither too wide nor to narrow to foster a radical innovation. With this view, it was then evaluated the option to frame the requirements as "user – story" or "jobs to be done". That would leave space for creativity and researchers identify them as a good starting point for innovation. The drawback is that it would shift the focus on the concept phase, that is not the experiment main purpose. This would have also required more time and it wasn't feasible. For those reasons, it was decided to use specific building names. Building with LEGO® allows nearly infinite possibilities anyway, so there is arguably still space for creativity in the ideation phase.

The number of requests was decided in relation to the net building time, to make the building phase challenging and keep participant at a good enough level of motivation. To understand these influences, I ran some experimental tests with different combinations of time and requests to roughly understand the effect on challenging and completion rate. Besides, the total number should not interfere in a biased way with neither one of the frameworks. The Agile treatment divides the building time in four time-box of ten minutes each. So, three items with seven sub-items were chosen to not be predictably divided among the Sprints. That would make the planning easier for participants in the Agile treatment compared with the traditional. Moreover, it would hide some behaviors in more complex situations.

The way and timing of giving information to participants is a relevant factor in PM simulations. The discussion was between giving participants all the information at the beginning or distribute them during the experiment. To introduce Agile's focus on customer opinions, it was discussed to give additional feedback after each sprint. Nevertheless, feedback would introduce complications to keep the treatments comparable. To test the implications of this factor it would be required an experimental design with multiple dimensions. That possibility should be discussed for further implementations. For this experiment, all the information available are given to the participants at the beginning, they are standardized for both treatments and do not change during the experiment. Since the differences between the frameworks should be the reasons for differences in the results, it is of major importance to have rules to make participants respect the identifying features. With this purpose, it was evaluated to use some sort of penalties to keep the treatments well defined and different. That again would introduce more complications than benefits. Considering that I stay with each participant individually, I constantly check the respect of the framework rules and verbally remind the rules if needed.

3.4 Experimental Procedures

Participants were given ten initial minutes to read the instructions carefully and do some practice with the tools. The instructions explain the framework rules and how to use the tools, with some examples. The actual experiment lasts for sixty minutes. Participants were given the customer requirements at the timer start. As said, it was done one session for each participant, so there are not problems of session contaminations or interference. I was with each participant for the whole time and they were free to ask clarifications about the framework or if they had problems with the tools. The timer running was visible to the participants. They were reminded when to move to the next "phase". I announced at certain moments the remaining time, with differences between the frameworks. For the traditional:

- every five minutes during the "Design and planning" phase

- after 20, 30, 35 minutes during the forty minutes building time.

For the Agile:

- every five minutes during the initial design phase
- after 5, 7 and 9 minutes during each ten minutes sprint.

This difference was done to emphasize the characteristics of the two different time management approaches. In the Agile framework, each sprint end is an intermediate deadline externally imposed that has to be respected. So, participants were reminded the time more frequently also to simulate the attitude that characterizes the rush to reach the sprint goal in the Agile framework. Instead, during the traditional treatment, there were less frequent reminders since I wanted to observe the participants attitude to respect their initial plan. Participants in the traditional framework set up their own intermediate deadlines for each task, but they were also free to delay them and change the plan. Since they hadn't compulsory deadlines while building, it had no meaning to remind them about the time.

At the time end, the participants were required to present their LEGO structures. Pictures were taken specifically for each sub-item, trying to avoid having two sub-items of the same building captured in the same picture. How these pictures were then used to evaluate the results is explained in the next section.

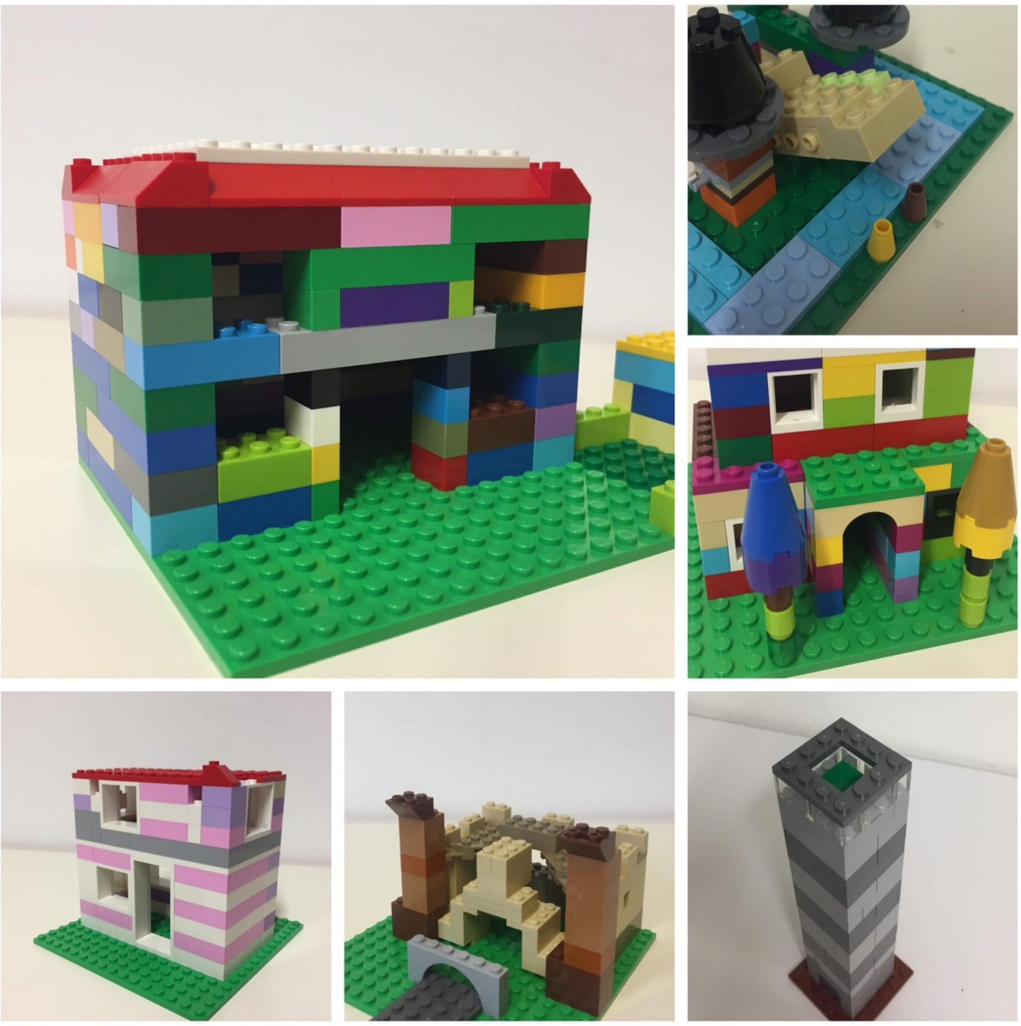
4 Experimental Results

The remainder of this thesis is organized as follows. In this chapter, I present the experimental results, that are of two types. The first analysis is about the performance results that are used to test the experimental hypotheses. Firstly, I present the data source and discuss the statistical characteristics of the analysis in section 4.1. Then the three experimental hypotheses are tested in sections 4.2-4.3-4.4. The second approach to experimental results is the qualitative one. It is based on the observations I did on participants' behaviors. In section 4.5 the relevant aspects are presented and insights on frameworks' characteristics are highlighted. In chapter 5 I present the learning game approach to the matter, reporting about experience done with a specific learning game on PM approaches.

4.1 Data Set and Statistical Approach

To test the hypotheses on how the results are affected by the frameworks, it is necessary to evaluate each sub-item built by participants. It would be desirable to have a standardized and objective way to evaluate the structures. Some options were discussed, mainly related to the number of bricks used, but no valid solution was found. Given that the LEGO® buildings have nearly infinite possibilities and the requests don't have an objective aim, the solution used is a subjective evaluation from a large sample of third persons. This evaluation was done via an online survey, that was completed by 120 responders. As mentioned in the previous section, each sub-item built by participants was photographed. Every picture was taken in the same spot with a neutral background. Those pictures were used in the survey to make people evaluate them. Figure 4.1 shows an example of the pictures used. A crucial point discussed for the evaluation is which type of question should be asked because it has an impact on the answers. Since the interest of the thesis is about the innovative results, it would be meaningful to ask responders to give a vote on the innovation level. The problem with this type of question is that innovation is a concept without a clear definition, and people don't have a shared understanding of it. This would have introduced uncontrolled effects that would compromise the validation of results. The same problem would affect a question about the beauty of the buildings. For those reasons, the evaluation question was framed in a more general way. Responders were asked about the best picture for each sub-item category.

Figure 4.1 Examples of pictures used in the online survey



Using a general question randomize the effects of people's personal biases. Another crucial decision for the survey is about how to make people evaluate, that has also an impact on the following statistical analysis. Since the hypotheses' tests are on the differences between sub-items, it would be desirable to use a way that accentuates the differences. This could be done by giving a total amount of points to redistribute among the pictures, the drawback is the risk to have many pictures with zero points. Another valuable option would be to make responders rank the pictures, that ensures to not have equal results but doesn't show the magnitude of the difference between two consecutive results. The decision was affected by the specific and practical constraints of the case. The number of votes required is considerable because for each one of the seven sub-items there are twenty pictures. In addition, twenty pictures on the same page don't allow easy and fast navigation throughout the survey. For these reasons, the distribution of points and the ranking options were evaluated as not feasible, because they would excessively irritate the responders and affect the evaluation. The choice was to make responders give points on a scale from 0 to 100 to each picture. It was evaluated to be an easier way for responders to complete the survey and the availability of a large scale would still leave space to differences to show. The survey results constitute the raw-data set for hypotheses testing. The set is made of 120 votes between 0 and 100 to 140 pictures, divided into the seven sub-items categories. Each sub-item result is evaluated as the mean of the 120 votes received. Table 4-1 reports the sub-items results divided by framework and participant. The drawback of these means is that they have very high variances. This is due to the spread of different votes that responders gave on the same picture. The standard deviations have a min value of 19, max value of 31, and a mean of 25. These values are considerable compared to the means they referred to, and it could be argued that it has no meaning to consider the mean as the sub-item result. To understand if the means are representative of the real trend of votes, I rank-transform each responder votes and calculate the rank-position of each sub-item. The purpose is to eliminate the effect of the different range of points used by responders. Then I compare the ranks with this approach and the ranks with the means. There is a match between the two in more than 88% of the cases. This result reassures that using the means as sub-items results is significative enough.

Table 4-1 Sub-items results used in the analysis

Average performance on each sub-item for participants								
		Items						
		1. High Rise Building		2. Residential House			3. Castle	
		Sub - Items						
framework	ID	nice rooftop	big main entrance	floors and windows	yard	garage	towers	drawbridge
Agile	1	48	62	36	64	59	43	65
	2	42	37	44	36	51	27	33
	3	40	30	56	17	39	26	0
	4	60	54	67	34	54	73	50
	5	38	24	37	44	27	62	44
	6	49	51	39	32	77	46	60
	7	56	57	38	72	48	30	64
	8	63	40	46	61	40	47	72
	9	42	47	60	65	62	52	48
	10	36	48	32	72	24	35	45
Traditional	1	35	41	73	24	0	53	0
	2	27	26	16	58	27	32	46
	3	27	37	41	48	48	20	41
	4	32	38	51	57	39	64	47
	5	47	52	30	34	44	25	36
	6	42	54	59	45	47	55	52
	7	47	31	52	39	50	48	51
	8	67	56	39	40	0	66	52
	9	46	37	59	64	38	47	37
	10	43	31	30	43	35	34	37

Before to move to the analysis, some statistical matters need to be discussed. The dependence of observations is a basic argument for experiments that involve decision-making. In this experiment, the outcomes are not decisions, as it would be in a classic newsvendor behavioral experiment. The observed outcomes are the survey evaluations, that still are affected by the decisions made by participants. It means that the 140 sub-items results cannot be considered as independent observations. In fact, the seven results that each participant got depend on one another, due to the experiment constraints that relate them. That correlation is at the basis of the experiment hypotheses, that test the dependences among the results of each participant. For these reasons, the dependencies between observations have to be considered. All tests of the "classic" hypothesis testing assume that the observations are independent draws from some distribution. One common approach to deal with dependencies between observations is the so-called ultraconservative. This approach consists of taking subject or session averages. The downside is that it drastically reduces the number of observations to analyze. In this experiment, sessions' and subjects' dependences collide, since it was done one session for each participant. Moreover, there are not enough participants to use the ultraconservative approach, but the issue is taken into account in the following anyway.

Regarding the statistical power of the analysis, the matter to be discussed is if the introduced variations in treatments' parameters are enough to show a difference. The treatments' parameters and the comparison to do are more complex than in the typical case of a newsvendor problem. In that case, it is analyzed the effect of factors that can be easily varied on a continuous scale, such as the number of bidders or the frequency of feedback. In this experiment, the parameter of comparison is just one – the framework for product development – but it's composed of many different characteristics. Some of them are the duration of each phase, the documents to be compiled, the possibility and the way to make changes during the experiment. The issue is how to determine those characteristics to obtain a significative difference. It is not obvious because the aim is to compare two overall approaches, that are also not defined in a unique way. For these reasons, I suggest considering the variable T – treatment as a continuous variable in the scale of the many possible approaches to product development. Among all those possibilities, this experiment chooses two of them as identifications of the traditional and Agile approaches. The choices were made to have a significative comparison even with just two treatments. To gain a more powerful analysis, it would be needed to have more values of the T variable to understand how the characteristics affect the results.

4.2 Traditional Hypothesis Testing

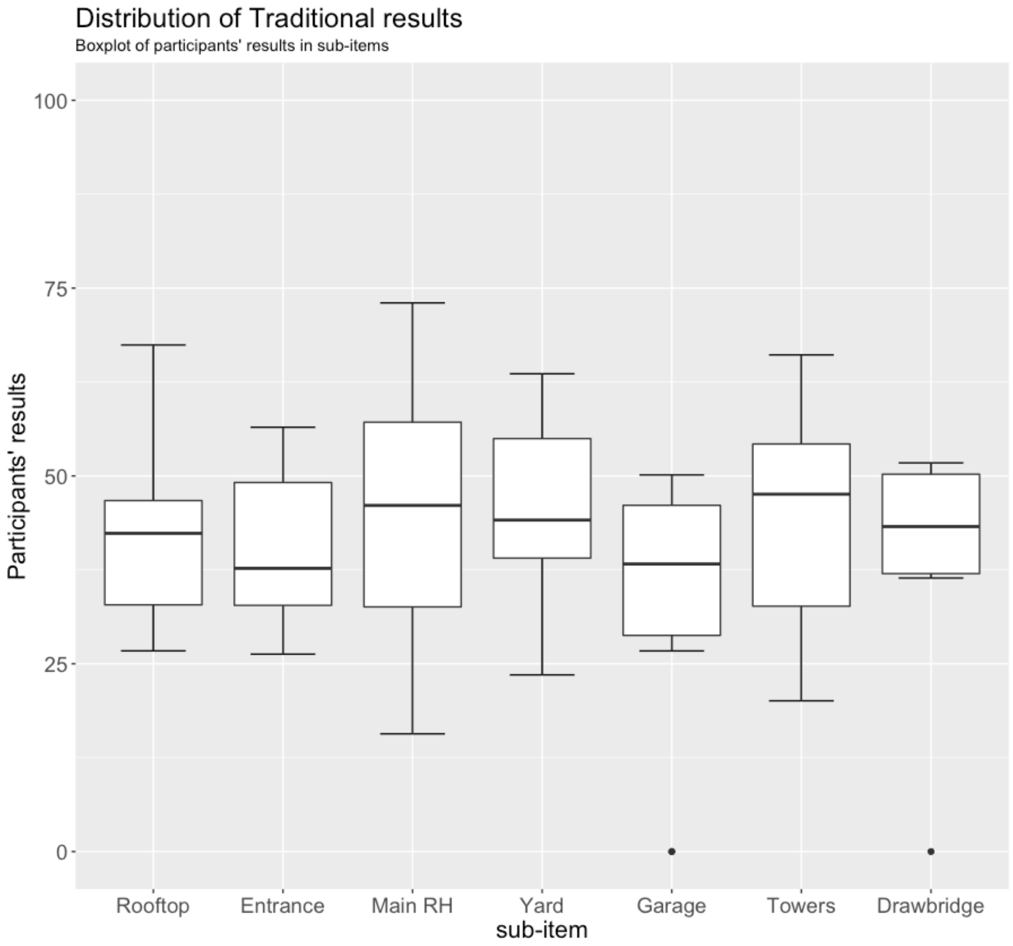
In the following analysis, the focus is on sub-items and items, so I recall them here with the shortened names that will be used. The items and sub-items are presented in the following order to all participants, and all the considerations about the order of sub-items refer to this one. Customers' requirements:

1. High rise building:
    a. Rooftop
    b. Entrance
2. Residential House:
    a. Main RH – Residential House
    b. Yard
    c. Garage
3. Castle:
    a. Towers
    b. Drawbridge.

The sub-item "main RH" was presented to the participants as a request that states "at least two floors and four windows." In the taken pictures this sub-item was basically the main building of the residential house, for that reason now is reported like that.

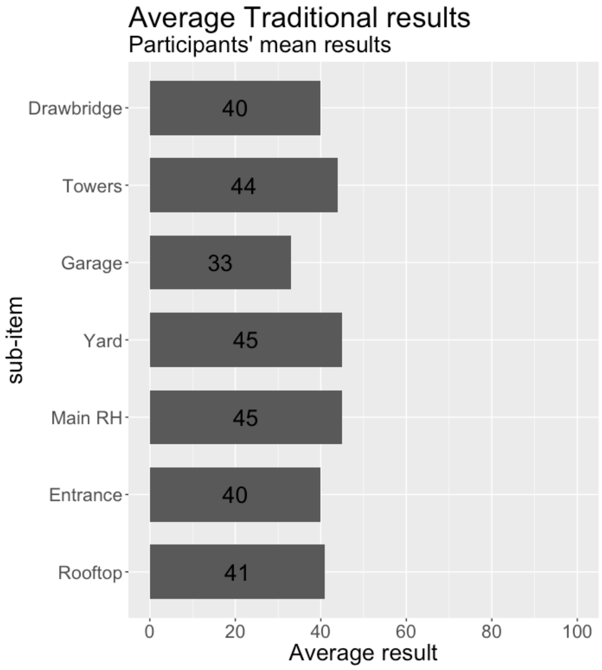
Prior to frame the hypothesis in a statistical way and test it, I present the results to make some overall considerations. Figure 4.2 shows the results obtained by the ten participants that used the traditional framework. There is a boxplot for each sub-item, that gives the distribution of results. In the garage and drawbridge sub-items, there are some results that received zero by default. It means that some participants (two for the garage and one for the drawbridge) didn't deliver at all those buildings. From a first visual inspection of the distribution, there are not eloquent differences between sub-items. The medians are all included in a range of twelve points and there is not a boxplot that is clearly higher or lower than all the other ones. Moreover, it's not visible the hypothesized trend from the first to the last sub-item. The hypothesis would expect a higher result on the first sub-item and a decreasing trend towards the followings.

Figure 4.2 Traditional framework results



Analyzing the means of each sub-item category the outcome is the same. The means are very close to one another, all seven values are in a range of twelve points. Given a scale of 100 points, the mean results don't show differences of high magnitude. Fig 4.3 shows the average results.

Figure 4.3 Average Traditional results

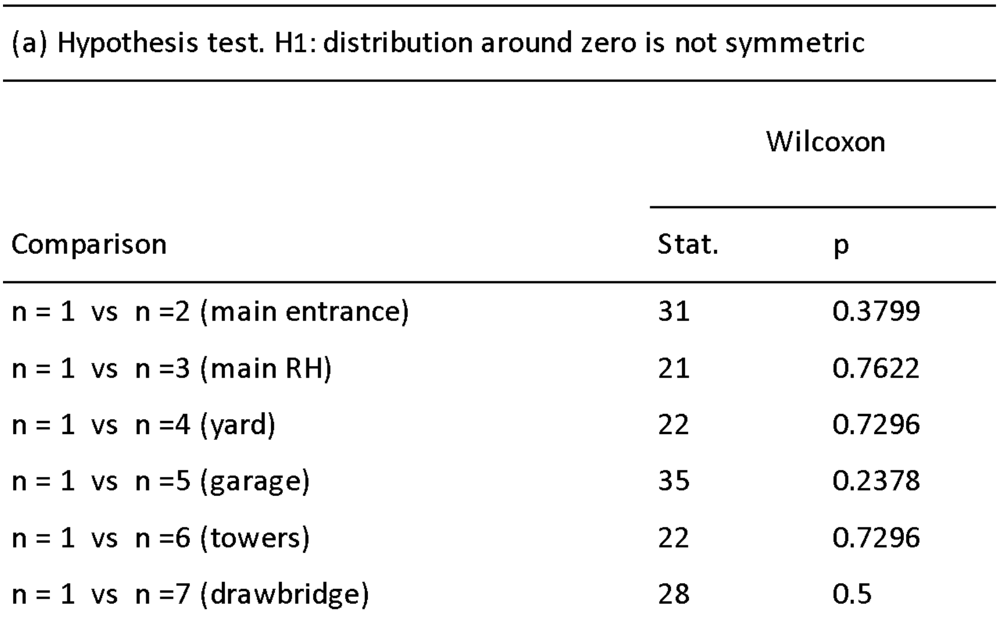


In the following the Traditional hypothesis is tested:

H1. With the traditional framework, participants reach a significantly higher result in the first sub-item. There is a significant difference between the result of the first sub-item and the others. The results show a decreasing trend from the first sub-item to the followings.

For statistical analysis, it means that the elements in the first – rooftop sample are on average higher than the elements in the other samples, at any significant level. Regarding the trend, a regression analysis should show a linear trend from the first to the last sample. I begin with comparing the rooftop sample with each one of the other samples using a two-sided nonparametric test. A nonparametric is used because I am not willing to make normality assumptions for these samples, and they are not big enough to use the central limit theorem. In each comparison, there is a dependence between the two samples, because every participant has one result in both samples. It means that the samples are made of paired observations for the participants. Given this sample composition, I use a Wilcoxon signed – rank test and do six comparisons. The Wilcoxon test calculates the differences between the paired observations. The null hypothesis is that the distribution of the differences is symmetric around zero, that means differences are due only to chance. The alternative hypothesis is that the distribution is not symmetric and is shifted in favor of the first sample. The comparisons' results are reported in table 4-2.

Table 4-2 Comparison of sub-items results with a two-sided test



(a) Hypothesis test. H1: distribution around zero is not symmetric

	Wilcoxon	
Comparison	Stat.	p
n = 1 vs n =2 (main entrance)	31	0.3799
n = 1 vs n =3 (main RH)	21	0.7622
n = 1 vs n =4 (yard)	22	0.7296
n = 1 vs n =5 (garage)	35	0.2378
n = 1 vs n =6 (towers)	22	0.7296
n = 1 vs n =7 (drawbridge)	28	0.5

Notes: n=1 is the rooftop, the first sub-item in the presented order. The Stat. column reports the value of the test statistic.

There are no significant differences between the first sample and any of the others, the lowest p-value is equal to 0.2378. It means that the differences are due only to the case.

This result is sufficient to reject the traditional hypothesis. It means that on average the traditional participants didn't reach a significantly higher result on the first sub-item. The result is even stronger because the rooftop sample has a mean that is higher only than three samples, and a median value higher than just two other sub-items.
Given the fact that there are more than two groups and it can occur the issue of multiple hypotheses testing, additional tests are done to evaluate the distribution of the sub-item samples. To test if there is a trend in three or more groups, given the dependence across groups and the non-parametric condition, there are two solutions (Donohue et al., 2018). One solution is the Friedman test, that is a non-parametric test, the other is repeated measures ANOVA on the rank-transformed data (Baguley, 2012). The ANOVA test is a parametric one, for that reason the data are transformed. I did both tests to check if there is some difference. The null hypothesis for the Friedman test is that the location parameter of each sub-item sample is the same. The alternative hypothesis for the ANOVA test is that at least two means differ. Table 4-3 reports the results. The null hypotheses cannot be rejected in both cases. It means that there is not a sample that significantly differs from the others. This implies that the participants not only didn't reach a peak result on the first sub-item, but also in any one of the others. I give an explanation of this unexpected outcome with the observed behaviors presented in the qualitative analysis section.

Table 4-3 Comparison across all sub-items' traditional samples

(a) Hypothesis test

	Friedman test		Repeated measures ANOVA	
Test	Stat.	p	Stat.	p
Sub-item order influence on results	2.7076	0.8446	0.063	0.803

Since requirements were presented to participants as three building items, it could be that participants focused on them as structures and not sub-items. This brings to frame the traditional hypothesis for the three items. Specifically, that participants reach a significantly higher result on the first building item. Therefore, it is meaningful to compare the results grouped in this way. I use three samples made with each participant average result in each item. In this way, the samples have the same dimension and the elements in each one of them are independent because they are average results from different subjects. The comparison can be done with the Wilcoxon signed-rank test because the samples are made of paired observations, as for the sub-items' samples. I did two tests between the first sample and the other two. The results are reported in table 4-4 and are not significative, the lowest p-value is 0.5472. This result confirms the previous findings of the absence of a peak.

Table 4-4 Performance comparisons between items with two-sided comparisons

(a) Hypothesis test. H1: distribution around zero is not symmetric

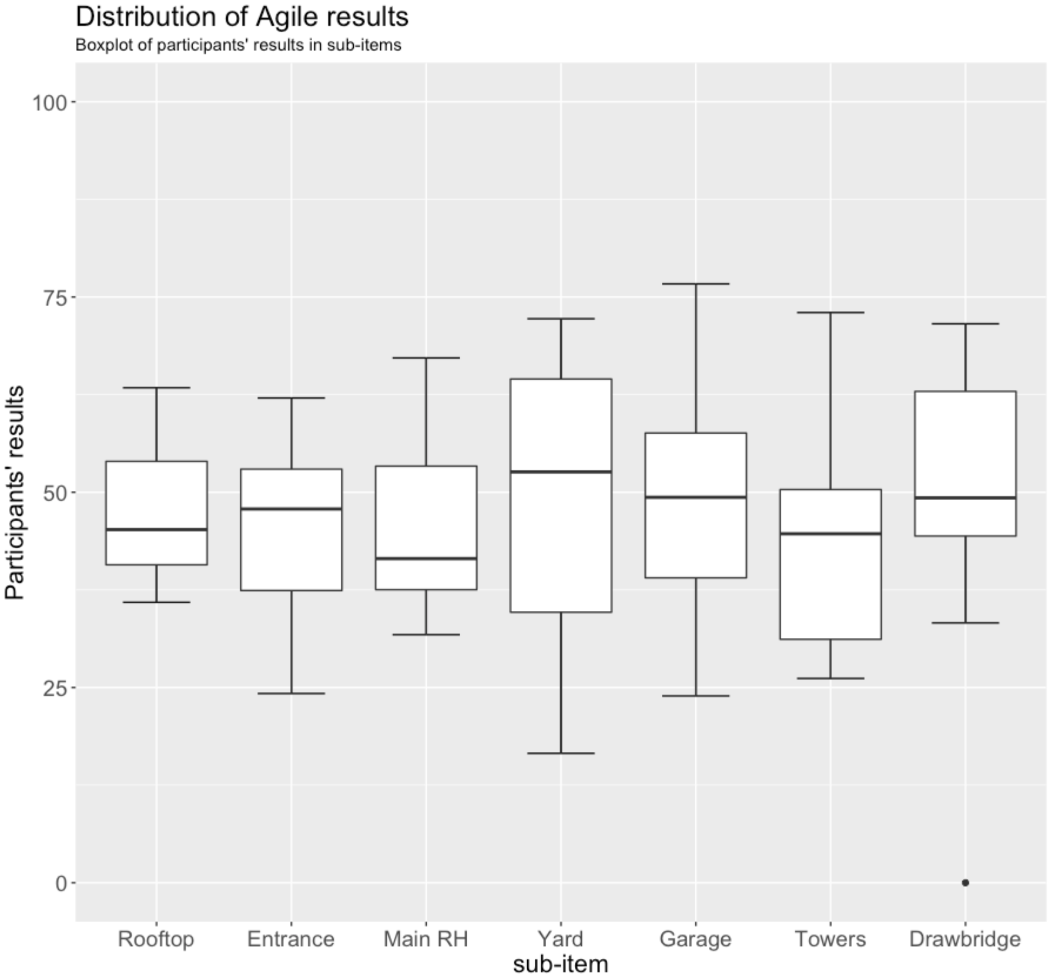
	Wilcoxon	
Comparison	Stat.	p
n = 1 vs n =2 (Residential house)	21.5	0.7465
n = 1 vs n =3 (Castle)	24.5	0.6394
n = 2 vs n = 3 (Castle)	22	0.5472

The results presented are strong evidence against the traditional Hypothesis – H1. The p-values in each test are so high that they don't leave space for further analysis. Besides, there is no meaning in testing the trend between the sub-items since it's clear that there is not the expected trend. The distribution of results and the comparison of means and medians show already that the rank of results is not as hypothesized.

4.3 Agile Hypothesis Testing

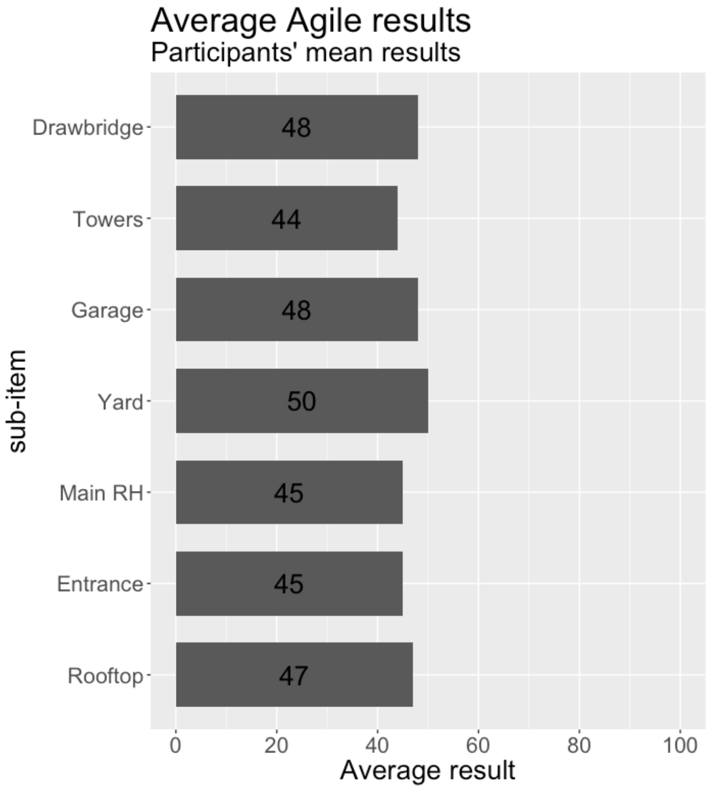
Before to move to the hypothesis testing, the following figures give an idea of the distribution of Agile participants results. Figure 4.4 shows the results obtained divided in sub-item categories. There is not an outstanding result, and the seven boxplots seem to be distributed on an average level. The range of medians is comparable with the traditional seen above, but it is shifted at a slightly higher level. One sub-item, the yard, has a median higher than 50 points, no one of the traditional sub-items reaches that level. Regarding the sub-items not delivered, there is one drawbridge not done, so with result equal to zero by default.

Figure 4.4 Agile framework results



The comparison of means clarifies furthermore the situation. Figure 4.5 presents the mean performance for each sub-item. The range is smaller than the traditional one, it is just 6 points, and the average level is higher with 46.7 points. These distributions say that the participants with Agile reached an average level among all the requests. This outcome is in agreement with the Agile hypothesis. Nevertheless, the fact that the situation is very similar compare to the Traditional one weaker the result. It could be argued that the outcome is not due to the specific framework's characteristics. This aspect is discussed in the following.

Figure 4.5 Average Agile results



The Agile hypothesis to test is:

H2. With the Agile framework, participants reach an average result in all sub-items. There is not a significant difference between any sub-item and the others.

To verify this statement, it has to be proved that there is not one sub-item sample with elements that are on average significantly higher than all the other samples. Given the dependences among the samples and the non-parametric characteristic, the situation is the same as for the traditional samples. The tests that can be done are again the Friedman test and the repeated measures ANOVA with rank-transformed data. With these tests the purpose is to understand if there are significant differences between the samples or are due only to chance. The tests' results are reported in table 4-5. The null hypotheses of both tests cannot be rejected. It means that there is not a sample that differs significantly from the average. This result brings to not reject the Agile hypothesis H2. The statistical analysis confirms the expected outcome for this framework.

Table 4-5 Comparison across all sub-items' Agile samples

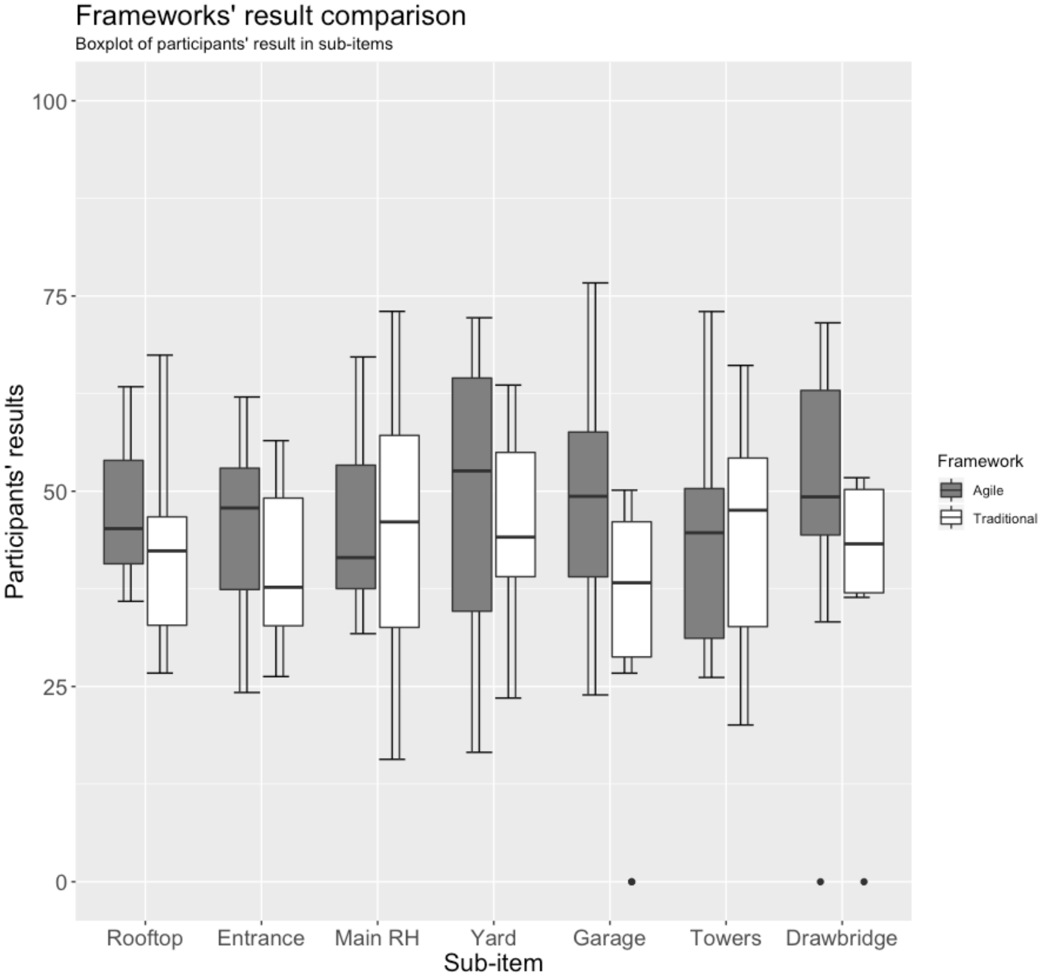
(a) Hypothesis test

	Friedman test		Repeated measures ANOVA	
Test	Stat.	p	Stat.	p
Sub-item order influence on results	1.8851	0.93	0.123	0.727

4.4 Comparative Hypothesis Testing

To roughly compare the frameworks' results, the following figures are helpful. Figure 4.6 represents the comparison between frameworks for each sub-item. In the categories, it is reported the distribution of the frameworks' samples, that are made of ten elements each. Looking to the median values, the Agile is higher than traditional in five out of seven sub-items. In two cases, garage and drawbridge, the outcome seems highly in favor of the Agile sample. In the other cases the comparison appears more balanced.

Figure 4.6 Traditional and Agile frameworks' results

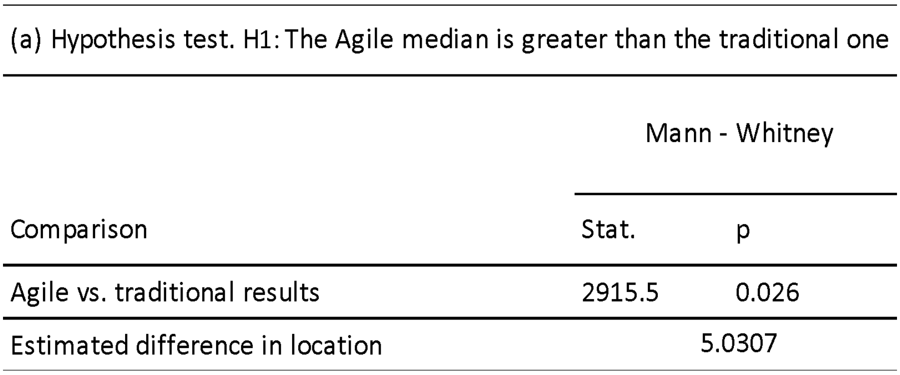


The comparative hypothesis is split in two parts. The first one is:

H3.1 With the Agile framework, participants reach an average result significantly higher than with the traditional framework.

To test the whole distribution of frameworks' results, I use a Mann – Whitney U test. It is the equivalent of the t-test but without the normality assumption. It is used to test two independent random samples, taken from two independent distributions F1 and F2. The test is used to verify the assumption that the distribution F2 represents a location shift of F1. The null hypothesis is that the medians of the two samples are equal. The alternative hypothesis is that they are not, and the first sample's median is greater than the second one. The mechanics of the test consists in ordering all the observations from lowest to highest. The lowest is given a rank of 1, and successive observations are given higher numbers. The ties are being given the average rank. The ranks are summed for all the observations of the first sample, and this value is called R_1. The statistic value is calculated as $U = R_1 - n_1*(n_1 + 1)/2$, with n_1 denoting the number of elements in sample 1. The distribution of U is tabulated for small samples, and for large samples is approximately normally distributed. In this case, the two samples have seventy elements each, that are all participants' results. The first sample is made with the Agile results, the second with the traditional ones. The test results are reported in table 4-6**.** The null hypothesis is rejected with a p-value equal to 0.026. It means that the Agile average result is significantly higher than the traditional one. This result doesn't reject the hypothesis H3.1.

Table 4-6 Performance comparison between frameworks

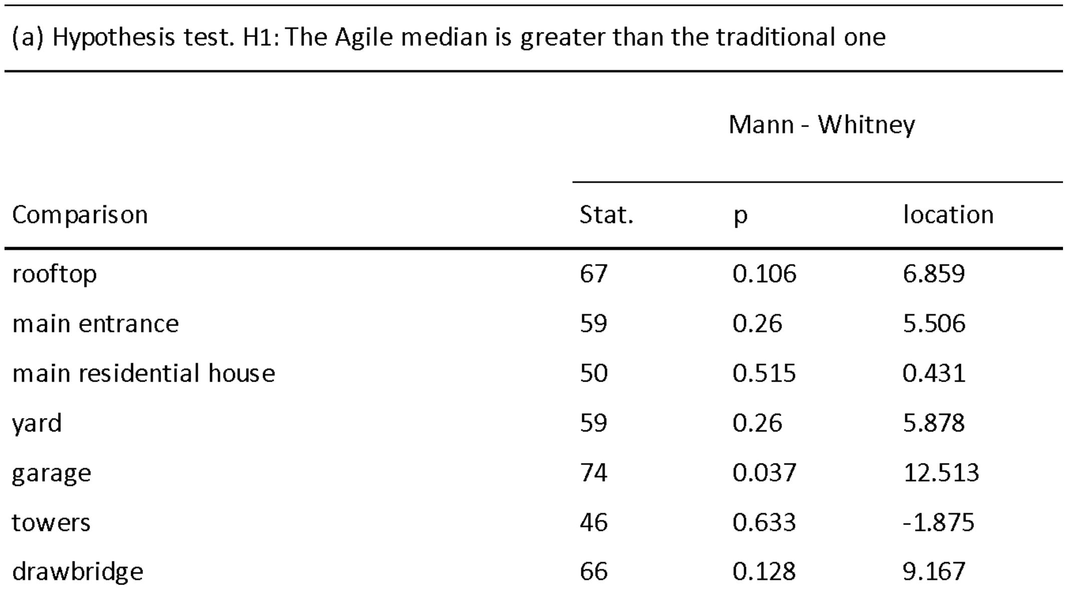


(a) Hypothesis test. H1: The Agile median is greater than the traditional one

	Mann - Whitney	
Comparison	Stat.	p
Agile vs. traditional results	2915.5	0.026
Estimated difference in location	5.0307	

The interesting further analysis to do is about how this result happened. The first comparison done is between frameworks for each sub-item. The aim is to understand if Agile participants ranked significantly better in some specific sub-item. It is done a Mann-Whitney U test for each sub-item, comparing the two treatments' results. The samples are made with the participants' result in each sub-item, seven samples for traditional and seven for Agile treatments. Each sample is made of independent elements, because there is only one result for each participant. Moreover, in each comparison, the samples are not dependent, so the test's assumptions are respected. The alternative hypothesis of each test is that the median of the Agile sample is greater than the traditional. Table 4-7 presents the results.

Table 4-7 Frameworks comparison for each sub-item



(a) Hypothesis test. H1: The Agile median is greater than the traditional one

	Mann - Whitney		
Comparison	Stat.	p	location
rooftop	67	0.106	6.859
main entrance	59	0.26	5.506
main residential house	50	0.515	0.431
yard	59	0.26	5.878
garage	74	0.037	12.513
towers	46	0.633	-1.875
drawbridge	66	0.128	9.167

Among the seven comparisons, there is just one case with a significant difference. In the garage comparison, the null hypothesis can be rejected with a p-value equal to 0.03. It means that the Agile garage results are on average significantly higher than the traditional ones. This result is highly affected to the fact that two traditional participants got zero by default in the garage. Also, the result for the rooftop is interesting, even if the comparison is not significant at any conventional level. The test has a p-value of 0.106, slightly higher than a confidence level of 90%, and compared with the values of the other tests is relevant. Additionally, this result is unexpected because in strong contradiction with the second part of the comparative hypothesis, that is:

H3.2 With the traditional framework, participants reach a significantly higher result in the first sub-item compared to the Agile framework.

From the test result, the opposite happened, and the hypothesis H3.2 can be rejected. The Agile participants reached results in the rooftop that are on average almost significantly higher than the traditional ones.

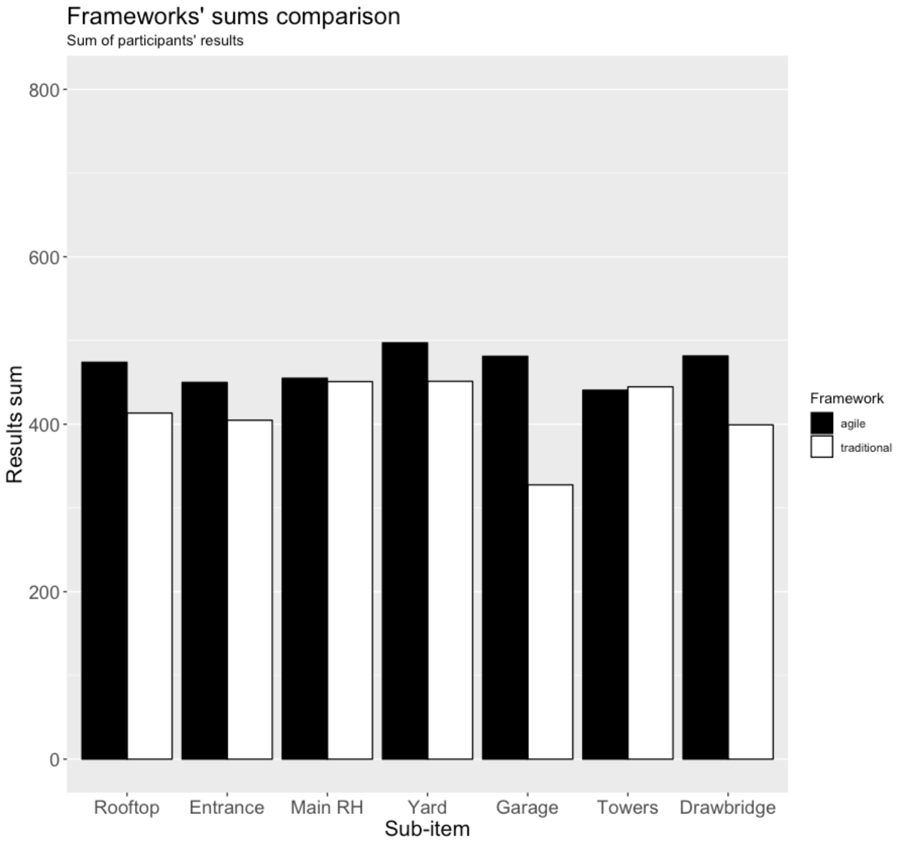
To gain a better understanding of sub-items comparison, table 4-8 shows the frameworks' means for each sub-item. The Agile's means are higher than the traditional ones for each sub-item, except for the towers and the main RH, where the average result is the same. It means that, even if not statistically significant, the Agile participants reached an average result higher than Traditional in every sub-item.

Table 4-8 Frameworks' average results

(a) Mean comparisons between frameworks		
	Mean performance	
Comparison	Agile	Traditional
1. rooftop	47	41
2. main entrance	45	40
3.main RH	45	45
4. yard	50	45
5. garage	48	33
6. towers	44	44
7. drawbridge	48	40

Another useful comparison is between the sums of points in each category. Figure 4-9 shows these results for the frameworks. As expected, the Agile overcomes the traditional result in each sub-item, except for two cases where they are comparable. This representation is meaningful also to understand the trends in the frameworks. The Agile results are comprised in a smaller range highlighting the fact that the results are balanced among all sub-items. The traditional results are more irregular due to poor results in the garage. In addition, the distribution shows clearly that the hypothesized trend for the traditional framework is not respected. That would be a decreasing trend from the first to the last sub-item. Instead, the first two sub-items have a total result that is lower than the following ones.

Table 4-9 Points sums for sub-items

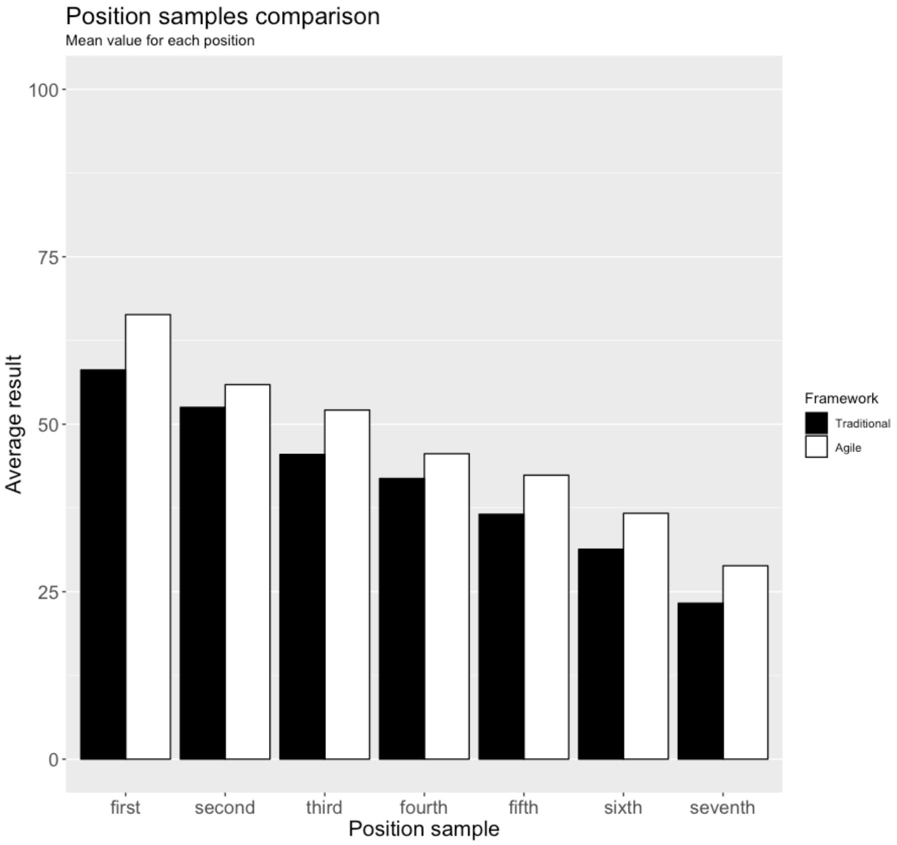


The hypotheses and considerations presented so far group the participants results in sub-items categories. The implications are difficult to demonstrate because they require that all participants had the same behavior in the same sub-item. Given this fact, a lighter type of hypothesis could be tested. The traditional hypothesis is about participants reaching a peak performance in the first sub-item. The behaviors that motivated it, could predict also that each participant reaches an outstanding result compared to her other results. This would be explained from the traditional framework's characteristics and the relation with expected biases and heuristics. Participants are freer to focus on one single sub-item because they don't have externally imposed limits. Effects from procrastination, Parkinson's law and student syndrome could affect people to spend more time on one task neglecting to respect the plan. There should be a difference from the Agile framework, where participants are forced to respect the time-boxed sprints, and this should prevent them to focus on a single sub-item. Thus, the hypothesis is:

H3.3 With the traditional frameworks, participants reach a significantly higher peak result compared to the Agile framework.

To test this hypothesis, I used different samples from the previous ones. I ranked each participant's results and create seven samples with the ranked position results, divided between the frameworks. It means that the "first position" sample contains the best result of each participant, and with the same logic the other samples are made. Fig 4-10 reports the average results of these samples. The picture shows clearly that the hypothesis will not be verified because the average peak is higher for the Agile sample. In addition, it is interesting to notice that the difference between the first and second position it's higher for the Agile framework. It means that not only the Agile peak is higher than the traditional, but it is also more relevant compare to the rest of Agile results. As expected from the fact that the Agile had higher average results, the means comparisons are in favor of the Agile for all samples.

Table 4-10 Position samples comparison



Given this premises, a test is done to understand the magnitude of difference in the peak results. I used a Mann-Whitney U test with the alternative hypothesis that the Agile median is higher than the traditional. The elements are independent within and between the samples, so the test's conditions are satisfied. The result is significant, and the null hypothesis can be rejected with a p-value equal to 0.038. The statistic value of the test is U = 74. It means the Agile participants reached on average a significantly higher peak than the traditional ones. The fact that the difference is even significant with a confidence level of 95% strongly reject the hypothesis H3.3.

The conclusion on the hypotheses' test is unexpected and surprising. The significant outcome is that the Agile participants reached an average and peak result higher than the traditional ones. With the traditional frameworks, participants didn't behave as hypothesized, or at least the effect was not strong enough to impact the results. I give an explanation on why this happened with the qualitative observations, that are reported in the following section.

5 Qualitative analysis

This chapter is based on the qualitative observations I made during the experiment. As explained in the experimental setting, I monitored each participant individually to capture relevant behaviors in relation to the frameworks. The analysis is presented as follows. Firstly, I discuss the most relevant behaviors that I observed in the majority of participants and explain their impact in sections 5.1 and 5.2. Then, I present how the specific frameworks' characteristics affected the way people worked in sections 5.3 and 5.4. In the last section, 5.5, I discuss which framework proved to be safer.

5.1 Agile Common Behaviors

The relevant common behaviors are three and they are presented in the following schematic way.

- **One building for one sprint.** I observed it was natural for participants to plan one single sprint for each one of the three main building items. The majority of participants did this type of planning. They probably saw the item as a whole, even if it was made of different sub-items. They didn't evaluate to redistribute the work among the four sprints from the beginning. For the majority of participants, it did make sense to use one sprint for one building. With this approach, they had one sprint left, but the use they made of it differed substantially. Participants either improved something that was already done or finished incomplete sub-items. One or the other behavior happened depending on the personal abilities to build with LEGO® bricks. Participants that were confident and fast with building managed to finish all that they had planned in three sprints. So, they were able to use the last sprint to make some sub-items really good or improve something that wasn't satisfying for them. In this way, these participants reached their highest results. Instead, participants with weak building skills didn't manage to complete everything in three sprints. Then they used the last sprint to reach an acceptable level on the incomplete sub-items. Only two participants didn't behave as explained, and they planned from the first sprint sub-items of different buildings.
- **High time pressure.** The participants felt the time pressure during all four sprints. In particular, the subjects that were trying to finish one building in each

sprint were more stressed. Close to the end of the Sprint participants rushed to complete what they had planned. Nine out of ten people reported that this time management stressed them during the experiment. This stress factor was in general helpful and had a positive impact on the results. I argue that the way participants planned and the time pressure, made them reach the right challenging level in each Sprint. That allowed them to have a sort of motivation peak in each Sprint. The result is that people were, in general, more productive. Only in a few cases, the stress had a negative outcome. It happened to a participant that, with just a few seconds left in the sprint, he was trying to assemble a structure and instead he broke it in pieces.

- **Incorrect time effort estimation.** In the beginning, subjects struggled to estimate the time effort of each issue. Even if they knew the net building time was forty minutes, they didn't consider having an effort estimation coherent with that. In no case, the sum of minutes for the sub-items was equal to forty. This initial bias affected the way they plan the Sprints, as described above. The fact is that, in the beginning, it was difficult for them to have an overall picture of the whole project.

5.2 Traditional common behaviors

- **Difficulties in the initial plan.** I observed that the majority of subjects found difficult to plan all the forty minutes in advance. The struggled for them was to estimate how much time is needed for each sub-item. They were asked to plan everything without having touch before the bricks. For this reason, they reported to not be sure about how to build and also what kind of bricks they would have found. The consequence of that initial uncertainty, it revealed in two different ways to split the time among the tasks. One way was to create seven or eight tasks and set almost the same time to each of them. Participants thought of doing eight tasks of five minutes each as the easiest way to plan the project. So, they didn't consider that different sub-items could require different amounts of time. The other method participants used required them more effort at the beginning because they tried to estimate precise times for different tasks. To do that they tent to create more tasks, around eleven/twelve. In this second way, participants used less time for designing since they focused more on the planning. With both types of initial plan, participants needed to reschedule during the building time. Almost everyone had to rearrange the initial plan already within the first item. It was either too detailed or too rough to be respected. In the latter case, the problem was mainly with overestimated time for small tasks. Regarding the overall productivity, I evaluate it was more efficient to make a simple and not too detailed plan, given the lack of previous knowledge and the high probability that changes would be needed afterward.

- **Last minute stress.** Participants were not stressed during the building time, even if they were making changes to their initial planning. There wasn’t a constant time pressure and having always a plan for the whole time made them feel on track. In addition, if they were running late with a task, they were allowed to use more time and then adjust the plan. Nevertheless, most of them got stressed around the last ten minutes of the building time. They were running out of time and still have to start the last building. This didn't happen only to two participants, that respected rigidly the initial plan without spending more time than what was planned on any task.

5.3 Agile Framework Implications

- **Sprint focus.** While working in a Sprint, it was easy and fast for participants to check what they had to do in those ten minutes. On the screen, there were only the issues to be done in that single Sprint. For this reason, it was simpler to check what to work on. I observed it was particularly helpful for participants that were struggling to build something. On the other hand, only the Sprint situation was under control. They should have checked the product Backlog to keep the overall under control, but with this little time, nobody did that.
- **Plan do check act.** Participants were forced to stop building for two minutes every ten. Except after the first sprint, when it was needed a re-arrangement of issues, the planning of the next sprint was done very quickly. It took them just a few seconds to drag and drop issues from the Product Backlog to the Sprint Backlog. They would have liked to start to work immediately, but they had to wait those two minutes. So, they were forced to think a bit more on how to build the next thing. In this way, they probably got a clearer idea of what to do when they work again with the bricks. I observed this aspect was helpful for participants with no good building skills. In these cases, the time alert stopped them from overworking. The Sprint review made them realize how much time was left and think about how to respond to all requests.
- **Sub-items interdependency**. The sub-items planned in the same Sprint depend on one another. I observed this correlation especially in Sprints that

turned out to be underestimated, often because some sub-items were completed very quickly. A sufficient level was reached minutes before the Sprint end. Giving the fact that they weren't allowed to work on something not planned for the sprint, the participants kept working on the items and improved them. In this way, some of the high ranked sub-items were done.

So, the results obtained in the same Sprint are correlated. In the case of the castle, this correlation is clearer because the two sub-items required a very different effort. The choices made for one of them affected the other. This case is discussed in more detail in section 5.5 because it's related to the safety level in frameworks.

5.4 Traditional Framework implications

- **Thoughtful initial phase.** In the initial phase of designing and planning, the framework helped participants to think more in detail on how to build the structures because they had to compile some kind of design documents. So, they had a more thoughtful initial phase but then no more during the building time. They never stop to build and make a point of the situation. It wasn't prescribed by the framework and nobody did it. They only moments when they stopped were to change the WBS, but they try to do that as fast as possible to go back to work.
- **Overall under control.** The WBS was always shown on the screen making easier for the participants to keep an eye on the overall project while building. Most of them kept checking the WBS. Even if they changed it sometimes, it made them feel secure about the progress since everything was planned and they just had to respect it. This is also one of the reasons why they weren't stressed during most of the building time.
- **Plan changes.** Two contradicting aspects happened. On one hand, participants found difficult or at least annoying to change the plan while working. Changing one task meant, most of the times, a need to change also the following ones. On the other hand, subsequent tasks were not strictly related to each other. It means that, if participants wanted to spend more or less time than planned on something, they were allowed to simply change the plan. They were not forced to wait a certain time, as in the Agile framework. With this perspective, traditional participants were freer to reschedule the work at every moment.

5.5 Is Agile safer than Waterfall?

The analysis of results presented in the previous section seems to confirm what can be found in the literature, that spiral and incremental development are safer than Waterfall. One aspect is that more participants with the traditional framework didn't deliver some sub-items. I observed that the time planning had a strong influence on this. Agile participants started the last building earlier, on average. This is due to the "one item – one Sprint" type of planning, explained above. That implies the majority of participants started the castle (the last item) in the third sprint. So, on average after twenty minutes of net building time. On the contrary, in the traditional framework participants planned on average to start the castle after twenty-eight minutes. After the plan changes, they actually started the first castle's sub-item after twenty-nine minutes on average. These evidence on times would support the comparative hypothesis. It suggests that Agile can help to reach a safer result because it forces people to work faster and touch with hand all the requests earlier. In case of a critical aspect on the last item, Agile participants could discover it earlier, when there is still time to deal with it properly. In the traditional framework happened the opposite, and often participants had even less time than what was planned for the last item. The hypothesized outcome of this behavior is a significantly lower result of traditional participants in the last sub-items. Looking at the results, this was not completely the case. The towers sub-item is one of the two cases where traditional participants reached a result with a higher median than agile. It is also the only Mann-Whitney comparison where the estimated location is in favor of traditional sample with 1.9 points. For the last – drawbridge sub-item the situation is the opposite, the Agile participants reached a higher result, with an estimated location of 9.2. As seen in the results analysis, these comparisons are not statistically significant but still, they are interesting for qualitative analysis. On this mixed result, not completely expected, the specific experimental setting had an impact. There is a correlation between the results on the last two sub-items, due to the fact that the drawbridge could be seen as a more difficult structure to do compared to the towers. I observed that since the Agile participants had more time available for the castle, on average they decided to focus on the drawbridge. In this way, they spent more time on that sub-item. Given the two sub-items were planned in the same sprint, that reduced the time available for the towers. This choice could explain the higher Agile result in the drawbridge, and lower in the towers. Instead, traditional participants had less time and on average didn't even try to make a proper drawbridge. Participants were running out of time, so they preferred to realize a simple design for the drawbridge and spent the remaining time on towers. It can be observed how in extreme late cases, where time was not enough for the castle, participants chose to not do the drawbridge at all. This affected also the low average traditional result on that sub-item. The pictures presented in the following support this explanation. Figure 5.1 shows five traditional cases out of ten where the drawbridge was built in a very simple way. In addition, in one case it wasn't built at all. With a simple way, I mean that participants used a limited number of bricks and the result is a basic structure. On the contrary, Figure 5.2 shows five cases of Agile results, where the participants tried to build a proper drawbridge.

Figure 5.1 Examples of simple drawbridge structures built by traditional participants

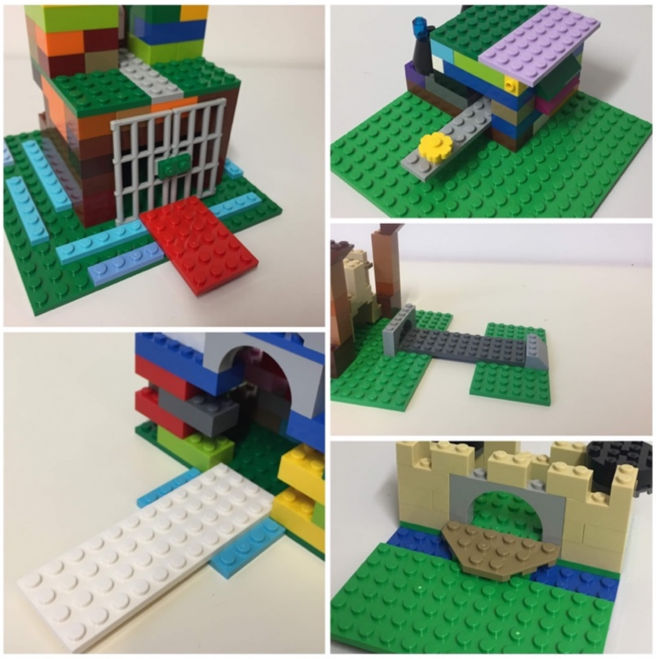
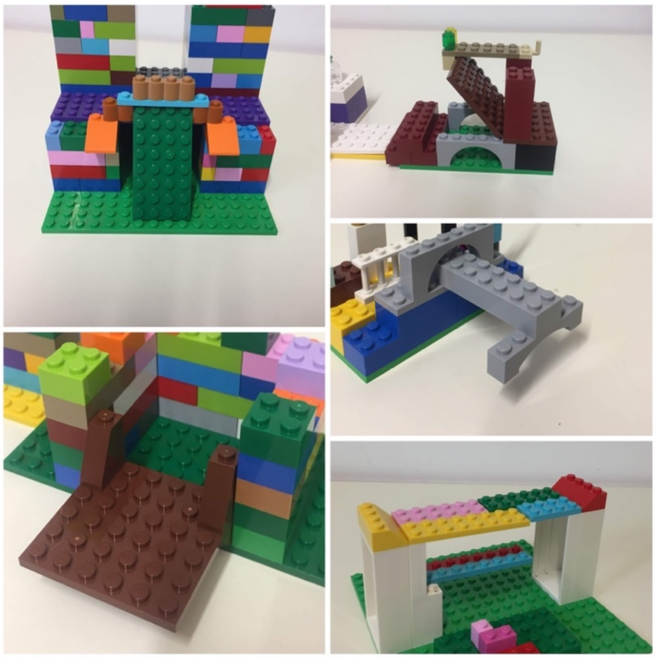


Figure 5.2 Examples of complex drawbridge structures built by Agile participants



6 Learning Game

An additional perspective on the comparison of traditional and Agile project management was gained with a learning game. Three sessions were done with students of two courses at TUM School of Management[6]. The learning game had the purpose to teach what the differences between traditional and Agile project management are. The setting is an implementation of the already mentioned exercise developed by E. Siemsen. The game was not done with an experimental approach. Nevertheless, it showed interesting insights on how a team behaves in those frameworks. The following sections are organized as follows. Section 6.1 explains the learning game design. Sections 6.2. presents the observations made afterward. Then, 6.3 discusses some critical aspects and implementations that could be done in future applications of the game.

6.1 Game Design

The game is designed to make students understand how it is to work in a team following a certain approach. In doing so, the relevant differences between the frameworks are highlighted and explained. It is a multiphase game, with a setting based on building a city with LEGO® bricks. The team has to respect specific procedures, that identify the framework's characteristics. One main aspect is the relation between the working teams and customers. The requirements are not standardized, and students that play the customer role have some freedom to define them. In the following, I discuss the main aspects of the setting. Firstly, the roles that students play, then the type of requirements and the frameworks' characteristics.

In the beginning, students are assigned randomly to different roles. The number of people in each of them depends on the total number of students. The working teams shouldn't be too big to still be manageable by the participants. The other roles can be used to make everyone participate without increasing the number of team members. The roles are briefly explained here:

- ➢ Customers. They are given a list of buildings and infrastructures and they are asked to define the details of those requirements. In the beginning, they have

[6] The courses were held during the winter semester 2018/2019 by Prof. R. Kolisch and S. Schiffels

time to discuss how they want their city to look like. They are asked to prepare a brief presentation explaining to the teams their requirements.

- Team members. Their role is to actually build the city with the bricks. In the sessions done, there were ten students for each team, that was considered to be the maximum number to still keep the situation under control.
- Project manager. The team members elect one person for this role. She doesn't work with the bricks but performs only managing tasks, that differ between the two frameworks. In the Agile framework, the role is called Scrum Master, following the terminology used in the Scrum Guide™.
- Observers. They are asked to debrief the class at the end of the simulation. During the game, they are free to observe both teams and take notes on how participants behave differently in the frameworks. Some specific suggestions are given to them on what they should observe.

The customers' requirements were presented to the students that played the customer role. They included eleven independent buildings. The considerations in deciding the characteristics are almost the same used for the experimental setting, with the purpose to create a challenging game. The differences are in taking into account the number of team members and how this affect the workload. Considering the time available, 1-hour total time and thirty minutes of net building time, eleven requests were evaluated to be the right number to not create an unfeasible project. The idea was to use requests not strictly correlated and not fully standardized. Nevertheless, the teams were asked to deliver a city, so they had to assemble together the buildings on a single big base plate. As in the experiment, customers were presented with a list of buildings names. In this case, they were then free to decide the details of the requirements.

Regarding the frameworks' characteristics, the basic idea is similar to the experiment. Nevertheless, in the game case there is not an interest to have the situations fully comparable. So, for the traditional team the phases don't have a fixed duration. It is done to simulate the reality of traditional PM, where there are not specific indications about time duration. The differences between the frameworks are specified in (a) the way design and planning are done; (b) the tools that teams use; (c) the way of managing the building time; (d) the project manager role and (e) the relation with the customers. For the traditional team, the work is divided into three steps. In the first one, participants prepare a one-page document that outlines the specifications and deliverables of the project. They are asked to present it to the customers, discuss the project with them and find an agreement on the requests. The second step consists in creating a work breakdown structure where they outline the tasks, their duration and a person responsible for each one. The third step is the execution, where the team members build the city. The project manager keeps track of the work monitoring the overall process and makes sure that there is a deliverable at the end of the 1-hour time frame. Regarding the interactions with customers, the team is free to ask an additional meeting with them to have their feedback on what was built.

In the Agile team, each step is time-boxed. The tool available for them is the Agile board, that is printed on paper and attached to the wall. It has different sections: user stories, features, Product Backlog, and Sprint Backlog. The Scrum Master is asked to make the members respect the schedule and keep the process on track. The first three steps are long five minutes each and the aim is to convert the customer requirements in features that constitute the Product Backlog. In the first five minutes, each team member writes user stories on a post-it and attach them in the user story section. Then each team member can suggest features related to the user stories. These features become essentially building tasks. In the next five minutes, they make a time effort estimation for each task. The building time is organized in three Sprints of 15 minutes each, divided in five minutes of Sprint review and Sprint planning, and ten minutes of building. During the planning, the team moves features to the Sprint Backlog and members take ownership of the tasks. After the first Sprint, customers take part in the Sprint review giving their feedback on the buildings. During the building time, team members work on their tasks and the Scrum master makes sure that there are potentially deliverable outcomes at the end of the Sprint.

6.2 Observations

In addition to evaluating the learning game under teaching purposes, the interest in this thesis work was to observe how people behave differently in the two teams. I observed the three sessions to find common patterns and specifically understand the effect of the framework's characteristics on behaviors. In this section, the observations are presented. Firstly, in general for the game results and then in detail for each framework. A first unexpected outcome was that the building results were arguably poor. Given the one-hour time and ten people working with LEGO® bricks, we would have expected to see nicer results. This is simply a qualitative evaluation, but it's interesting if there is a correlation to the game setting. Given the similarities of specific requests, the teams' results can be compared with the individual experiment's ones. The quality of results is really different and in favor of the individual builders. Considering eleven requests and ten people, each person would have roughly thirty minutes of net building time for just one structure. Instead, the individuals have forty minutes for three structures. Clearly, the way of working is very different, and the comparison is not direct. Nevertheless, some causes for teams' poor results can be identified. One main cause is the difficulty in organizing the teamwork. Managing the work of ten people is not easy and requires a lot of effort. In both teams, students struggled to be efficient and they were not highly productive. The other element that added much complexity in the teams' work was the interaction with customers. Students that played the customer role were not prepared and this added a lot of fuzziness. This fact, in itself, is not negative and is part of the game. The drawback was that customers didn't make requirements that add value to the buildings, but simply kept the team busy. This highly impacted the productivity of students. Another relevant outcome, in general, was that teams delivered almost everything closed to the time end. It means that on average participants were not able to respect precisely the rules. This covered some differences that should have been more evident between frameworks. The Agile team was supposed to deliver an Increment after each Sprint, and the traditional team was required to respect the WBS and finish buildings when stated on the plan. These aspects were hardly respected, and this made the approaches more similar than what they were meant to be.

6.2.1 Frameworks' Results

The observations specific to frameworks' characteristics are divided into four groups. In each of them the Agile and traditional are compared to highlight differences. The aspects are about how the initial phase is managed, the effect of the project manager role, the stress level and how this affects the way of working. In the end, also the interactions with customers are discussed.

Management style and behaviors in the initial phase:

- ➢ The traditional teams struggled more in the initial phase, mainly because they weren't provided with any tool to facilitate the design process. Thus, the personal skills of the project manager had a higher impact. Basically, it was her to decide how to conduct this phase. That resulted in difficulties to have all members actively participating in the brainstorming and in the decisions for the design. I found this phase was less productive compared to the Agile team because not everybody was able to express their ideas.
- ➢ For the Agile team, the first phases were carried out more easily. The fact that team members had to write their ideas on post-it an attach them on the storyboard facilitated the process. It wasn't necessary the intervention of the

Scrum Master in this very first moment. Nevertheless, if the student playing this role had the right attitude, she could foster an even richer idea-gathering phase. This brief initial moment was a booster to let the team's creativity express. In addition, the externally imposed deadlines for each phase made the team members interact faster. A sticking point was the time effort evaluation for each task, that was done approximatively. This was expected, due to the lack of experience on how to do that estimation and also on the specific LEGO® tasks.

Project manager impact and way of working during the building time:

- ➢ In the traditional teams, the project manager was not impactful during the game and the result was also a lack of rules respect. In particular, the plan was not well respected, and the team members did work without restrictions. It means that students were free to keep adjusting the buildings because they were not forced to respect the task completion. This was due to the difficulties that the project manager had in making the team respect the plan. Clearly, the plan needed changes during the building time and the fact that it was written on paper make it harder to adjust it in a meaningful way.
- ➢ The Scrum masters had a hard job during the building time. They were required to accomplish many tasks and they were constantly under pressure. They had to make sure there was something deliverable after each Sprint, make the team respect the deadlines and conduct the meeting with customers and the debrief afterward. Given the framework more structured, they were forced to respect these basic requests. Having more tasks to do, they had a stronger impact on how the team worked. In particular, the personal way to approach the role could change relevant aspects, such as the motivation and stress levels.

Perceived level of stress during the game:

- ➢ The traditional teams didn’t appear stressed during almost all the game duration. Nevertheless, during the last circa fifteen minutes, the stress increased a lot. Normally, the team realized that there was still a lot to do and they were struggling to assemble together the various structures built by individuals. This increased challenge made the team rush close to the end and reach their production peak. An implication was not complete respect of the plan in this last phase.

- The Agile teams felt more stressed during the whole duration of the game. The pressure was high close to each sprint and it made members rush to complete the building task within the Sprint limit. The motivation and challenge were perceived higher depending on how much the Scrum master was pushing the team to work harder.

Customers' influence:

- The traditional team had only one meeting with the customers at the beginning. They had the chance to recall them to receive feedback, but no team asked for it. Thus, they were not "disturbed" during the game by the customers, meaning the team didn't have to make unplanned changes to the design of any part. They built what was agreed at the beginning without any further acceptance by the customer on the real structures. Since the initial project charter was not particularly specific, the team had still freedom in what to build.
- The Agile relationship with the customers was very different, as prescribed by the approach. There was more attention to the customers' requirements. The drawback was due to feedback neither clear nor constructive. During the meetings with the team, the customers didn't have a common line and each of them expressed personal opinions. In addition, most of the times their feedback didn't add any value to the structures but still forced the team to keep working on the same buildings for consecutive Sprints. This was probably unsatisfying for the students working on them. In general, the customers slowed down the Agile teams but there was no sign of higher quality in the building results due to this.

6.3 Game Implementations

After the done sessions, the game proved to be valuable to be applied in classes. Nevertheless, some implementations should be done to obtain better results both under teaching purposes and behavioral evaluations. In the following, I present some suggestions. The purpose is to sharpen the differences between the frameworks by making the students more respectful of the rules. In addition, the desired outcome would also be to have comparable results and a way to understand differences due to the frameworks. A relevant aspect that should be modified is the customer role, I start by discussing its characteristics. After that, I present the suggestions divided by framework.

6.3.1 The Customer Role

The way this role is played has major implications on how the game proceeds and especially on participants' behaviors. I observed that there are some critical aspects in the implemented way. The main ones are related to the freedom left to customers. Even if they had a list of requirements, there were still almost infinite possibilities. In itself, this is not a negative aspect. The problem encountered was that, at the end of the game, the building's features were not comparable between the two teams. It was due mainly to the fact that the customers didn't maintain a common line during the game. As explained above, the impact customers had was unbalanced between the two frameworks. The frequent interferences with the Agile team had many negative effects and a few positive ones. Given that they received more feedbacks, they should have been able to satisfy better the requirements. This positive aspect should also reward the Agile team in the end. Instead, it didn't because there was not a final moment were customers gave their final opinion and compare the two results. The debrief was done by the observers, but the team members would have probably been more interested to hear the customers' opinion since they worked for them.

I argue that a more regulated version of the customer role would produce more significant outcomes. In general, I would suggest forcing the customers to decide and maintain a common line during the game. To reach that, they could be required to produce and sign some internal documents where they agree on their decisions. In addition, this would keep them busy during the game. I observed that in some cases they were bored from the game, and this might have a negative impact on their behavior too. Regarding the requirements, they should receive a list of buildings as it was done. In addition, they could be asked to decide on a precise list of sub-items for the buildings. At the beginning of the game, they would present this list to the teams, with a brief description of their requirements. In this way, their common line would be clearer. Regarding the Agile Sprint review, one they should be required to write down the feedback they gave after each one. The purpose is to make them more responsible for their decisions. Moreover, these documents could be used to cross-check the building results at the end of the game. After the game end, they should also be asked to give a final evaluation to the class. When they discuss which team satisfied better their requirements and why. As said, the team members interfaced during all game with the customer, so they would be interested to hear their opinion. The suggested implementations for the customer role should make the frameworks more comparable. Also, the final cross-checking between customer requirements and building results allows evaluating the frameworks' implications.

6.3.2 Agile Framework

The main problem that should be solved is that teams didn't respect the condition to have a viable Increment after each Sprint. Students kept working on items that are "never" delivered. It means that they delivered almost everything at the end of the game without respecting the Sprint structure. I would suggest some differences in the rules for this framework. Firstly, I would make the team decide a definition of done for every feature, that should be done by writing a description on a post-it. Moreover, they should present to customers only items that are done and potentially ready to be delivered, in accordance with the definition. Since they need something to show to customers, they should put more effort into finishing the buildings within each Sprint. Regarding the setting, it would be meaningful to use four Sprints instead than three. Framed as Sprint planning + building, repeated four times, and not the opposite as was done. More Sprints should highlight the differences with the traditional framework. Having less time in each sprint would make students work faster. I would expect an impact similar to the one in the individual experiment. So, team members should reach more performance peaks given a higher challenge and motivation perceived. The Sprint Review requires some small changes too. The customers should not remain for all the time. The team needs time to debrief the feedback received prior to moving on to the next sprint. Moreover, I observed that the review is more meaningful if the builder of each structure presents it to the customers. That could be written as a framework rule. The positive aspect is increased interactions, that clarifies the difference with the traditional approach.

6.3.3 Traditional Framework

The suggested implementations have the aim to make the teams respect the rules more strictly. In particular, for the traditional framework, the main problem is in respecting the WBS plan. Not respecting it made the team deliver almost all buildings at the end of the game. To give more support to participants in respecting the plan, they should use a "live" WBS. It can be done with a simple excel sheet and projected on the wall. The purpose is to make it easier to control and adjust the plan for the project manager. Moreover, it would be more feasible to check the real progress and respect of the plan also for external observers. Regarding the interactions with the customer, I would argue that there is no meaning in not having an additional meeting with the clients. I suggest implementing two fixed meetings with the customer. Fixed in the number but not in the duration. The first to give an acceptance review of the preliminary design at the beginning, it should happen to be a relatively long meeting where they discuss the design. While close to the end of the game, a short meeting should be done for the final acceptance review. Even in the most traditional implementations of the waterfall approach, it was suggested that not having a final review from the customer is risky and counterproductive. To make the game more realistic, it should then be implemented. I would add a rule for the meetings to sharpen the difference with the Agile framework. The project manager should be the only one to present the design to the customers, to reduce the possible interactions. In addition, for the project manager role, it should be asked to compile standardize sheets to report on the progress of the project. While the team members are building, the project manager should update the WBS plan and also keep track of the planned and real-time schedule. The documents could be framed in a way that underlines the precise responsibility of the project manager in making the team respect the plan. In this way, the project manager should be more dedicated to that. The expected positive impact should bean higher respect of the framework rules. As said, this aspect is fundamental to keep the frameworks well distinguished.

7 Conclusions

This thesis is the first experimental attempt to my knowledge to study the effects of different PM approaches on project results. Concepts from the BeOps research are used to formulate the hypotheses on how people would behave in a certain framework. From these expected behaviors hypotheses on the outcomes are drawn. I developed an experiment with a real-effort physical task to compare the traditional and Agile frameworks in a NPD project. The main experimental result is that with the Agile treatment participants reached a significantly higher peak result, compared to the traditional treatment. The outcome is surprising because it was expected to happen the opposite. The explanation can be found in the way participants organized their work. In the Agile framework, participants ended up focusing on some specific requests. The reasons were either to obtain a higher result or to reach a deliverable level. Moreover, the results confirmed the Agile hypothesis, with participants that reached a significantly higher average result. I observed that the Agile framework enabled participants with good building skills to excel, but also helped less skilled people to obtain a sufficient overall level. This was possible thanks to its characteristics that keep the challenge and stress level high for all the experiment duration. Participants were forced to work faster from the beginning, and this was a key element given the type of tasks required. The experiment confirms the research that argues that incremental and spiral approaches reduce the risk of failure in a project. With the Agile framework, just one participant didn't deliver one requests, and it was not due to framework characteristics but simply because the participant was not able to build the specific sub-item. Instead, two participants with the traditional framework failed to deliver two and one sub-item. Even if not statistically significant, these outliers are interesting because the participants failed due to the specific frameworks' features. Both of them were so concentrated on one particular sub-item that they forgot to respect the plan. No characteristics of the frameworks stopped them. It resulted in not enough time to deliver all the requests. This peculiar behavior was the foundation of the traditional hypothesis. It happened only in two cases and not systematically as expected; the explanation could be related to the specific experimental setting. The tasks were fairly easy to complete and the building time to manage short. It was clear to participants the need to respect the overall plan in order to deliver all the requests. With a closed deadline as one hour, it is less impactful the effect of procrastination and Parkinson's law. In addition, it could be argued that for this type of project, that can be somehow related to the construction field, the traditional approach is very appropriate. This correlation might have stopped the expected behaviors to happen.

The implications of the experiment for the innovation research question are interesting. The Agile approach could allow its users to reach faster and safer an average result in the project, that means normally a less expensive result. Furthermore, with the correct calibration between intermediate deadlines and work distribution, Agile could foster an outstanding result achievement. The outstanding characteristic depends on many factors more than the framework, but the Agile approach can be used to leave these factors to happen because it doesn't stifle them. For management practices, it means that Agile can be used to reach a base level on requirements and from that point implement an innovative result. On the other side, the traditional treatment proved to not bias participants to focus on specific requests. That is commonly recognized to be one primary cause of delays in traditional project management. In real-long term projects, the situation can be significantly different, and the expected behaviors have a higher impact. It is not clearly explained why traditional participants didn't reach the expected peak. The causes may lie also in the fact that participants were not challenged enough and worked with a level of stress too low. I would argue it happened because in the traditional framework there are not essential characteristics that motivated the participants. Thus, in real applications, the management should be aware to introduce some motivational sources with external factors.

This work addresses the differences between frameworks regarding the approach to planning, the correlation between planning and execution, the required documentation in support of a project and the flexibility to changes. Given the complexity of product development features, the results may not apply to differences in other aspects. Further implementations of the experiment would be needed to analyze more elements of the approaches and reach a complete understanding.

8 Reference List

Aranda, J., S. Easterbrook. Anchoring and adjustment in software estimation. Eur. Software Engineering Conference/ACM SIGSOFT Symposium.

Arrichiello, V., M. Rossi, S. Terzi. Is systems engineering a stifler or an enabler of innovation? A contribute to the ongoing debate. 2014 International Conference on Engineering, Technology and Innovation (ICE), 1-7.

Baguley, T. 2012. *Serious stats: A guide to advanced statistics for the behavioral sciences.* Macmillan International Higher Education.

Bassler, D., J. Oehmen, W. Seering, M. Bendaya. A comparison of the integration of risk management principles in product development approaches. 306-316.

Beck, K., M. Beedle, A. Van Bennekum, A. Cockburn, W. Cunningham, M. Fowler, J. Grenning, J. Highsmith, A. Junt, R. Jeffries, J. Kern, B. Marick, C. R. Martin, S. Mellor, K. Schwaber, J. Sutherland, D. Thomas. 2001. *Manifesto for Agile Software Development.* Available: http://agilemanifesto.org.

Bendoly, E., D. Hur. 2007. Bipolarity in reactions to operational ‘constraints’: OM bugs under an OB lens. *Journal of Operations Management* 25 1-13.

Bendoly, E., M. Prietula. 2008. In “the zone” The role of evolving skill and transitional workload on motivation and realized performance in operational tasks. *International Journal of Operations Production Management* 28 1130-1152.

Bendoly, E., M. Swink. 2007. Moderating effects of information access on project management behavior, performance and perceptions. *Journal of Operations Management* 25 604-622.

Bendoly, E., M. Swink, W. Simpson. 2014. Prioritizing and monitoring concurrent project work: Effects on switching behavior and productivity. *Production Operations Management* 23 847-860.

Bendoly, E., W. Van Wezel, D. G. Bachrach. 2015. *The handbook of behavioral operations management: Social and psychological dynamics in production and service settings.* Oxford University Press.

Biazzo, S., R. Panizzolo, A. M. De Crescenzo 2016. Lean Management and Product Innovation: A Critical Review. *In:* CHIARINI, A., FOUND, P., RICH, N. (eds.) *Understanding the Lean Enterprise: Strategies, Methodologies, and Principles for a More Responsive Organization.* Cham: Springer International Publishing.

Bisin, A., K. Hyndman 2014. Present-bias, procrastination and deadlines in a field experiment. National Bureau of Economic Research.

Brown, T., B. Katz. 2011. Change by design. *Journal of product innovation management* 28 381-383.

Bullinger, H. M., K. P. Fahnrich, T. Meiren. 2003. Service engineering - methodical development of new service products. *International Journal of Production Economics* 85 275-287.

Buser, T., N. Peter. 2012. Multitasking. *Experimental Economics* 15 641-655.

Chesbrough, H., W. Vanhaverbeke, J. West. 2006. *Open innovation: Researching a new paradigm.* Oxford University Press

Choo, A. S. 2014. Defining problems fast and slow: The U-shaped effect of problem definition time on project duration. *Production and Operations Management* 23 1462-1479.

Croson, R., K. Schultz, E. Siemsen, M. L. Yeo. 2013. Behavioral operations: The state of the field. *Journal of Operations Management* 31 1-5.

Curlee, W. 2008. Modern Virtual Project Management: The Effects of a Centralized and Decentralized Project Management Office. *Project Management Journal* 39 S83-S96.

Dewar, R. D., J. E. Dutton. 1986. The adoption of radical and incremental innovations: An empirical analysis. *Management science* 32 1422-1433.

Donohue, K., E. Katok, S. Leider. 2018. *The handbook of behavioral operations.* John Wiley & Sons.

Eppinger, S. D. 2001. Innovation at the Speed of Information. *Harvard Business Review* 79 149-158.

Erickson, J., K. Lyytinen, K. Siau. 2005. Agile Modeling, Agile Software Development, and Extreme Programming: The State of Research. *Journal of Database Management (JDM)* 16 88-100.

Feng, J., T. Sedano. Comparing extreme programming and Waterfall project results. 24th IEEE-CS Conference on Software Engineering Education and Training (CSEE&T).

Hebert, J. E., R. F. Deckro. 2011. Combining contemporary and traditional project management tools to resolve a project scheduling problem. *Computer & Operations Research* 38 21-32.

Highsmith, J. A. 2002. *Agile software development ecosystems.* Addison-Wesley Professional.

Highsmith, J. R. 2009. *Agile project management: creating innovative products.* Pearson Education.

Institute, P. M. 2008. *A guide to the project management body of knowledge (PMBOK(R) Guide).* Project Management Institute.

Joore, P., H. Brezet. 2015. A Multilevel Design Model: the mutual relationship between product-service system development and societal change processes. *Journal of Cleaner Production* 97 92-105.

Kach, A., A. Azadegan, K. J. Dooley. 2012. Analyzing the successful development of a high-novelty innovation project under a time-pressured schedule. *R&D Management* 42 377-400.

Kagan, E., S. Leider, W. S. Lovejoy. 2017. Ideation–execution transition in product development: an experimental analysis. *Management Science* 64 2238-2262.

Kahneman, D., A. Tversky 1977. Intuitive prediction: Biases and corrective procedures. Decisions and Designs Inc Mclean Va.

Kerzner, H. 2013. *Project management: a systems approach to planning, scheduling, and controlling (11th ed.).* Hoboken, NJ: John Wiley & Sons.

Kim, W. C., R. Mauborgne. 2014. *Blue ocean strategy, expanded edition: How to create uncontested market space and make the competition irrelevant.* Harvard business review Press.

Kruger, J., M. Evans. 2004. If you don't want to be late, enumerate: Unpacking reduces the planning fallacy. *Journal of Experimental Social Psychology* 40 586-598.

Kuchta, D., D. Skowron. 2016. Classification of R&D projects and selection of R&D project management concept. *R&D Management* 46 831-841.

Liberatore, M. J., B. Pollack-Johnson. 2013. Improving Project Management Decision Making by Modeling Quality, Time, and Cost Continuously. *IEEE Transactions on Engineering Management* 60 518-528.

Lovallo, D., C. Clarke, C. Camerer. 2012. Robust analogizing and the outside view: two empirical tests of case-based decision making. *Strategic Management Journal* 33 496-512.

Mcdermott, C. M., G. C. O'connor. 2002. Managing radical innovation: an overview of emergent strategy issues. *Journal of Product Innovation Management* 19 424-438.

Mumford, M. D., R. A. Schultz, J. R. Van Doorn. 2001. Performance in planning: Processes, requirements, and errors. *Review of General Psychology* 5 213-240.

Parkinson, C. N., O. Lancaster. 1958. *Parkinson's Law or the Pursuit of Progress.* Murray London.

Pellegrinelli, S. 2011. What’s in a name: Project or programme? *International Journal of Project Management* 29 232-240.

Petersen, K., C. Wohlin, D. Baca. The Waterfall Model in Large-Scale Development. *In:* BOMARIUS, F., OIVO, M., JARING, P., ABRAHAMSSON, P., eds. Product-Focused Software Process Improvement: Springer Berlin Heidelberg, 386-400.

Pichler, R., S. Schulze. 2005. Book Reviews: Agile Project Management: Creating Innovative Products by Jim Highsmith, and Agile Project Management with Scrum by Ken Schwaber. *Journal of product innovation management* 22 371-373.

Relich, M. A computational intelligence approach to predicting new product success. 11th International Conference on Strategic Management and its Support by Information Systems, 142-150.

Royce, W. W. Managing the development of large software systems: concepts and techniques. Proceedings of the 9th international conference on Software Engineering: IEEE Computer Society Press, 328-338.

Salgado, E. G., R. Dekkers. 2018. Lean Product Development: Nothing New Under the Sun? *International Journal of Management Reviews* 20 903-933.

Schwaber, K., M. Beedle. 2002. *Agile software development with Scrum.* Prentice Hall Upper Saddle River.

Serrador, P., J. K. Pinto. 2015. Does Agile work? — A quantitative analysis of agile project success. *International Journal of Project Management* 33 1040-1051.

Shenhar, A. J. 2001. One size does not fit all projects: Exploring classical contingency domains. *Management Science* 47 394-414.

Sloan, A. P. 1990. *My years with general motors.* Crown Business.

Smith, P. G. 2005. Book Reviews: Agile project management: Creating innovative products. *Journal of Product Innovation Management* 22 369-369.

Spalek, S. 2016. *Traditional vs. Modern Project Management Methods. Theory and Practice.* Smart and efficient economu: Preparation for the future innovative economy, 21st International scientific conference.

Sterman, J. D. 2000. *Business dynamics: systems thinking and modeling for a complex world.*

Sutherland, J. 2014. *Scrum: the art of doing twice the work in half the time.* Currency.

Sutherland, J., K. Schwaber. 2017. The definitive guide to scrum: The rules of the game.

Takeuchi, H., I. Nonaka. 1986. The new new product development game. *Harvard business review* 64 137-146.

Tong, J., D. Feiler. 2016. A behavioral model of forecasting: Naïve statistics on mental samples. *Management Science* 63 3609-3627.

Tversky, A., D. Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185 1124 - 1131.

Ulwick, A. W. 2016. *Jobs to be done: theory to practice.* Idea Bite Press.

Vojak, B. A., R. L. Price, A. Griffin. 2012. Serial Innovators: How Individuals Create and Deliver Breakthrough Innovations in Mature Firms. *Research-Technology Management* 55 42-48.

Von Hippel, E. 2010. Chapter 9 - Open User Innovation. *In:* HALL, B. H., ROSENBERG, N. (eds.) *Handbook of the Economics of Innovation.* North-Holland.

Wilcox, K., J. Laran, A. T. Stephen, P. Zubcsek. 2016. How being busy can increase motivation and reduce task completion time. *Journal of personality social psychology* 110 371.

Womack, J. P., D. T. Jones, D. Roos. 1990. *Machine that changed the world.* Simon and Schuster.

Wyrozebski, P., S. Spalek. 2014. An Investigation of Planning Practices in Select Companies. *Management and Production Engineering Review* 5 78-87.

Zang, J.-J., G.-Q. Wang, W. Zang. Research on four-electrical railway project cost estimate based on the WBS standard templates. 2013 International Conference on Management Science and Engineering 20th Annual Conference Proceedings, 720-726.