Comparative Performance Assessment of the Wisconsin State Capital Projects

By

Youssef Nagy Lamie Labib

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

(Civil and Environmental Engineering)

at the

UNIVERSITY OF WISCONSIN – MADISON

2019

Acknowledgment

First and Foremost, I thank the One who has given Himself for me. His love, nurturing, trustworthiness and generosity have never failed to impress and satisfy.

Nobody has been more supportive to me in the pursuit of this project than the members of my family. I thank my father, Nagy, whose support and care has always been overwhelming, my mother, Manal, whose unconditional love and wisdom have always been my companions. They are the ultimate role models. I also thank my older sister and childhood friend, Sara, whose tenderness, integrity and hard work have been my inspiration.

Most importantly, I thank my lovely partner, Mariam, for always standing by my side through thick and thin. Your mere presence in my life has been a blessing and an endless delight.

To my advisor, Dr. Awad Hanna, I hope that you might be as proud to be my mentor as I am to be your mentee. Your guidance has brought me to this place through false starts and hesitations. I would not be here without you.

Additionally, I'm extremely grateful to the rest of my thesis committee: Prof. Jeffrey Russell, Mr. John Nelson, and Prof. Wei-Yin Loh for their encouragement, insightful comments, and hard questions. I am also grateful to Dr. Wafik Lotfallah for his imperative guidance and persistent efforts. Your input has been influential to this research.

I extend special thanks to Mr. Nelson for his generosity and affirmative contribution to my career through the John S. Nelson Wisconsin Distinguished Graduate Fellowship (WDGF). I also thank the University of Wisconsin-Madison for supporting my education through the Teaching Assistant and Research Assistant appointments.

I thank all my fellow graduate students who have taken some time to discuss and enrich my work. Your intelligence and dedication have challenged me to grow. Especially, I thank Michael Ibrahim, for introducing me to this great opportunity, and for teaching me fundamental research and statistical techniques during our Construction Industry Institute (CII) project.

My deep appreciation goes out to the Division of Facilities Planning & Management at UW-Madison and the Capital Planning and Budget office at UW-System for their knowledge, patience and support. I also thank Representative Dave Murphy, as well as the Office of the Lieutenant Governor, Rebecca Kleefisch, for their substantial support during the data collection phase.

I'm also grateful to The Boldt Company, where I interned during the summer of 2018, working on the company's Milwaukee Children's Hospital of Wisconsin project and expanding my knowledge of the construction industry. I also thank Affiliated Engineers, Flad Architects, and individual state project managers for their support as I conducted my research.

Finally, I thank my extended family at the St. Mary & St. Rewais Church. Without you, my time in Madison would have been significantly harder.

May the Almighty God richly bless all of you.

Abstract

The Engineering and Construction (E&C) sector of the global economy accounts for $10 trillion in expenditures annually and creates more than 100 million jobs around the world. In the United States alone, $1.25 trillion were spent, and 8.4 million individuals were employed by the E&C sector in 2017. Yet despite the economic significance of this industry, repeated studies have shown that productivity is at best holding steady, and at worst declining relative to the industry's performance in previous decades, and the present performance of comparative non-farm industries. Since 1964, construction industry productivity has been steadily declining, with the industry now operating at only 80% of its 1964 productivity, an annual decline of -0.7%. This is largely attributed to the fragmented nature of the industry, a lack of innovation and adoption of disruptive technologies, strict regulations on construction performance, excessive risk transfer, and lack of coordination between different contracting parties. The impetus for this research effort is the industry-wide problem of declining and stagnating productivity, which frequently leads to delays in construction.

Worldwide, some of the largest owners and clients of building construction are in the public-sector (i.e. state, local, and federal governments and their agencies). Traditionally, public building construction procurement operates on a low-bid model, using the Design-Bid-Build (DBB) delivery method. The method ensures transparency in selection of the prime general contractor and focuses on the initial construction cost, often crucial to annual budgets and public oversight. But DBB is criticized for not considering total cost of construction and ownership, which can be much higher due to delays, quality issues, etc. The Traditional DBB delivery method can be divided into three types:

- DBB Multiple-Prime (MP): the state solicits bids for each major type of work on a project and engages in separate contracts with each of the winning firms.
- DBB Single-Prime: the state solicits bids for the entire project and contracts directly with a general contractor, who in turn contracts with the pertinent subcontractors or suppliers.

- DBB Hybrid Single-Prime (SP): the approach in current use by the State in which the State solicits bids from MEP contractors 1 week before the GC bid is due. The state then selects the lowest specialty contractors and the GC is bound to contract with the winning MEP contractor.

Other innovative project delivery methods (PDMs) have been developed that can increase early participation of the contractor(s) in the design stage and encourage collaboration. These include, for example, Design-Build (DB), Construction Manager at Risk (CMAR), and Integrated Project Delivery (IPD). These PDMs involve collaboration among all stakeholders, including the general contractors and subcontractors, much earlier in the process than under the DBB system. This allows for more effective problem-solving. Furthermore, studies have shown that in some situations those delivery methods can reduce change orders, delays, and cost overruns.

This thesis seeks to investigate the delivery methods used in Wisconsin public building projects and make recommendations that will enable state decision makers to improve outcomes on future projects. The data collected in this study shows that the majority of the building projects were executed under Hybrid Single-Prime (SP) and DBB Multiple-Prime (MP). However, some of Wisconsin's public building projects used other delivery methods. This research compares the results of these delivery methods to the Traditional DBB method using a number of performance metrics.

The initial step was a comprehensive data collection effort. The objective of this phase was to eliminate coincidental inconsistencies, maximize sample size, and capture as much project diversity as possible. Data was collected from 189 capital projects completed between 2000 and 2017, with a total construction value of $5.2 billion (in 2018 dollars). This amounts to over half of state-funded building construction projects (by value) over this period. There was a negligible difference in the population of projects by state agency (i.e. UW-System, Department of Corrections, etc.), thus the sample collected is considered an appropriate representation of Wisconsin public building construction projects executed under the supervision of the Department of Administration (DOA).

Next, projects were evaluated in an objective, quantifiable manner. Eight metrics were identified from the literature and used to assess performance in schedule, communication, change management, and spending – both individually and collectively. From these metrics, a mathematical model, the Project Performance Index (PPI), was developed. The PPI is a single numeric score that ranges from 1 to 10, facilitating the comparison of projects that may significantly differ in their scope, cost, length of contract, etc. The scoring and analysis were then validated using two independent statistical techniques: Confirmatory Factor Analysis (CFA) and Partitioning Clustering Analysis, ensuring consistency of the findings and the confidence level of the conclusions. Finally, the project performance was broken down by UW-System campus (if applicable) to determine the impact of this factor on the performance of a project.

The results show that 97% of projects in this study cost more than the bidder originally proposed. This cost growth averaged 10.3%. However, it should be noted that some cost data includes contingency while others do not. The degree to which the cost growth observed has resulted from the normal use of contingency is unknown. Specifically, the data for DB and CMAR contain contingency, whereas the data for DBB does not.

In terms of schedule, 69% of projects were delivered behind schedule. These delays extend the schedule by an average of 39%, increasing indirect costs of administration of the projects, and the opportunity costs of projects delivered later than initially committed. However, these costs were not available and are therefore not considered in this thesis. A review of current literature shows that these costs can be substantial and that PDMs other than DBB manage the costs resulting from these delays more favorably.

This thesis compares the performance of DBB, DB and CMAR based on the metrics defined in this study. It was found that DB statistically outperformed SP and MP in all eight of the performance metrics. CMAR outperformed SP in 5 metrics and MP in 6. It was further found that SP outperformed MP in terms of 3 metrics, and DB outperformed CMAR in terms of only one metric. These findings were validated statistically at the 95% confidence level. Nevertheless, these findings should be used with the understanding that multiple factors were not captured by the analysis, including the complexity of projects, market conditions, unforeseen conditions, as well as other intangible factors.

Application of the PPI model developed in this research found that Collaborative delivery methods—that is, Design-Build (DB) and Construction Manager at Risk (CMAR)—performed 40% better on average than the more traditional Hybrid Single-Prime (SP) and Multiple-Prime (MP) methods. That is, Collaborative projects are 40% more capable of reducing changes, executing quickly, meeting schedule, and maintaining communication among vested parties than Traditional methods.

Furthermore, due to the lower performance of both SP and MP as compared to DB and CMAR (based on the metrics), this research recommends the State investigate the potential benefits from the use of alternative and innovative delivery methods. These methods may be particularly suitable for advanced projects such as hospitals and research labs, or complex, large-scale projects which may require early coordination of multiple trades and design. More Collaborative delivery methods have the potential to improve communication and allow for innovative practices. One possible approach would be to develop a pilot program that allows for the selection of more appropriate methods on specific projects, those characterized by high levels of system complexity and a high risk of schedule slippage, and evaluate the benefits for the State of Wisconsin as a proof of concept.

Table of Contents

List of Figures

List of Tables

1 Chapter 1: Introduction

1.1 Background & Motivation

The Engineering and Construction (E&C) industry has a profound effect on people and economies throughout the world. In 2016 alone, $10 trillion USD was spent on construction-related goods and services. This accounts for 6% of the global gross domestic product (GDP) in that year (World Economic Forum (WEF) 2016). Additionally, it is estimated that over 100 million people worldwide are employed in the construction sector (Global Construction Perspectives 2013, Oxford Economics 2013). This amounts to 7% of the working-age population (McKinsey and Company 2017). The construction industry also has an immense appetite for raw materials. Construction consumes 50% of global steel production annually, and approximately 3 billion tons of raw materials are consumed globally to manufacture building products, such as concrete, dimensional lumber, etc. (World Steel Association 2015).

The United States is no exception to the worldwide trends in construction. The econometrics are as follows: $1.25 trillion in construction spending, amounting to 4% of GDP in 2017 (U.S. Census Bureau 2017); 8.4 million individuals employed in the Engineering and Construction (E&C) related industries (U.S. Bureau of Labor Statistics 2017).

Yet despite the massive spending and high employment numbers, the E&C industry has been facing significant adversity from stagnating or declining productivity over the last several decades. In turn, this widespread productivity issue has led to low margins. A 2016 study examined the financial performance of 30 public engineering and construction companies from 2005 to 2015. Only 15% of these firms experienced consistent double-digit growth and margins. On average, it was found that these companies experienced a compound annual growth rate (CAGR) of only 6.3%, and earnings before interest, taxes, depreciation and amortization (EBITDA) of only 10.4% (McKinsey & Co. 2016a). This amounts to severe underperformance when weighed against other industries, such as manufacturing, fashion, etc., who are regularly able to achieve operating margins of 30 to 40% (Ibid.). Furthermore, and perhaps most worrying, E&C firms provide only a 5% return on investment (ROI) for investors and shareholders, a figure regularly outperformed by other S&P 500 companies (WEF, 2016). This low ROI dissuades investment and impedes business development in the E&C sector.

The source of these tight profit margins on which the E&C sector operates is the inherent tight competition and constrained supply and demand companies face. However, it is forecasted that effective management of operations can lead E&C firms to margins of 20 to 30% (McKinsey & Co. 2016a). But rather than focusing on efficiency improvement and innovation, or on productivity enhancement, many E&C firms focus on utilization by entering into low-margin business to keep resource and manpower consumption high, worrying about lost surplus when the time comes to reconcile claims and change orders. While this can be profitable in cases of nonstandard or costly specifications, it limits the company's profit centers overall due to competition and furthermore, creates an adversarial relationship with the owner due to the inherent nature of claims and change reconciliation (Franz et al. 2017; McKinsey 2016a).

Furthermore, the E&C sector as a whole struggles with poor labor performance. Many publications agree that productivity is declining, with the only point of contention being the magnitude of decline. Teicholz et al. (2001 and 2013) and the WEF (2014), are conservative, forecasting -0.4 to -0.5% decline annually. McKinsey & Co. (2015) and the Bureau of Labor Statistics (2017) cite stagnation, rather than decline, of 0 to 1%. Optimistically, McKinsey & Co. (2016a) and the Bureau of Labor Statistics (2018) found positive improvement in specific sectors and/or demographic areas of 5 to 20%. Yet all these sources concur that manufacturing and other non-farm industries significantly outperform the construction industry by a factor of 1.5 to 2.5.

These factors (significant productivity loss, tight profit margins, poor management) and others are the impetus behind the overall poor performance of construction projects. Studies have shown that 98% of megaprojects (i.e. those projects with total cost of $1 billion or more) suffered cost overruns greater than 30% and 77% of these projects fell behind schedule by at least 40% (McKinsey & Co. 2015). Globally, this amounts to $1.6 trillion USD (equivalent to the GDP of Canada) in lost value annually, and projects in the United States account for 33% of this loss – approximately $530 million (McKinsey & Co. 2017). Three significant issues help explain the poor performance of the construction industry: fragmentation of the industry at-large, slow innovation and adoption of technology, and encumbering government regulations.

1.2 Industry Challenges

1.2.1 Fragmentation

The construction industry is severely fragmented yet relies on the seamless collaboration of all participants in the value chain. It was reported that in the United States there are over 700,000 E&C firms, yet only 2% of these employ more than 100 people, and 80% employ 10 or fewer (WEF 2016). As the execution of a construction project typically encompasses numerous stakeholders (e.g. designers, contractors, etc.) each bound by their own contract(s) with no unified management, stakeholders necessarily have different and sometimes competing priorities. Owners, for example, seek generally to deliver the project on time and on or under budget. However, contractors may prioritize their revenues and margins over timeliness, particularly in cases where there exist no penalties for delay or in situations where contractual caps for penalty have been tapped out.

The result of this fragmentation can be an adversarial on-site environment (Pishdad-Bozorgi and de la Garza 2012; Carpenter and Bausman 2016). In this situation, project parties approach each project as not a long-term investment or learning process, but rather as an opportunity for continued survival. As a result, lessons learned on one project are often lost or ineffectually transferred to future projects. Thus, additional efforts must be made to train and instill shared standards and values in teams which could have come from lessons learned previously (WEF 2016; Franz et al 2017).

1.2.2 Digitization & Innovation

The E&C industry is one of the least digitized industries globally. The MGI digitization index ranks construction second to last in the United States (ahead of only agriculture) and last in Europe (McKinsey & Co. 2017). This is due to an industrial culture that has not adopted digital solutions. Although numerous software applications have been published and made available in recent years, the industry continues to overly rely on paper documents such as blueprints, design drawings, procurement and supply-chain documentation, progress reports, etc. (McKinsey & Co. 2016b). Even in situations where one or more software applications are in use, there is little integration among them.

One potential reason for this lack of innovation in technology and techniques is the significantly low figure for research and development (R&D) spending in construction. Only 1% of revenue is applied to R&D in construction, while automotive and aerospace apply 3.5 and 4.5%, respectively (Ibid). Information technology (IT) spending follows much the same pattern, with construction again spending less than 1% of revenue (Ibid). As a result, construction professionals have difficulty capturing and analyzing data. This is a significant problem because historical data can be used to improve risk management and forecasting in future projects. As construction projects frequently involve binding arbitration or litigation as a means to resolve disputes, mismanagement of project documentation contributes to disagreements between owners and contractors on matters of project progress, change orders, requests for information, contractual notice, etc. Lack of IT support and integration further hamstrings construction, as siloed software applications generate differing conclusions from the same set of data, leading to competing versions of reality which in turn negatively affect information sharing (Ibid).

1.2.3 Regulation

A recent study by KPMG (2011) found that regulation was the most significant cause of increased complexity in construction. The construction industry is subject to numerous regulations from various regulatory bodies charged with protecting the environment, public safety, etc. Such regulations can compromise profitability and schedule by delaying the approval of funding. Presentations at the World Economic Forum in 2015 concluded that a balance must be struck between the need to ensure compliance and the need to quickly secure funds (WEF 2015).

Public procurement of construction projects primarily relies on the design-bid-build delivery method (DBB), with projects typically awarded to the lowest bidder. As a result, public procurement of construction projects focuses primarily on controlling initial construction costs (the bid), while not adequately considering the total cost of ownership (WEF 2016). For example, DBB separates the design and construction phases to such a degree that there is no opportunity, in many cases, for the architect and the contractor to collaborate, which can result in a less integrated project team (Franz et al. 2017). As DBB is typically used with the selection of a contractor based on the lowest bid, the focus of the project is incorrectly shifted from project-based to profit-based, further misaligning the team and the jobsite (Pishdad-Bozorgi and de la Garza 2012; Franz et al. 2017).

While these practices and methods were appropriate in the past, when projects were less complex, companies less fragmented, and technologies less developed, the time has come to innovate. Progressive governments have begun to reform the procurement process and adjust bidding requirements to allow for more collaborative delivery methods and outcome-oriented bidding, which stimulates optimization and productivity (Col Debella and Ries 2006; Rojas and Kell 2008; Hale et al. 2009).

1.3 Reflections on The State of Wisconsin

Within the state of Wisconsin, construction has been growing at an annual rate of 5% in the past half-decade. In 2017, the industry employed over 100,000 workers and had total expenditure of $5.2 billion, 43% of which was allocated by public entities including state, local and/or municipal agencies (U.S. Census Bureau 2017; U.S. Bureau of Labor Statistics 2018). Despite the size and positive growth of this sector of the Wisconsin economy, there are some inefficiencies which will be elaborated upon in the forthcoming sections.

1.3.1 Current Funding Approval Process

The scope of this research is to explore whether the benefits of collaborative delivery methods (DB and CMAR for instance) reported in the literature are applicable to public projects in Wisconsin. However, a review of the data revealed that significant time elapses between when a potential project is first identified and when funding in secured. Particularly in the case of facilities executed on behalf of the University of Wisconsin at Madison, the approval process for funding can take up to 30 months (2.5 years) before funds are allocated and an architect can get hired, as explained in Figure 1 below. Per the collected data, a project with a value of $100 million could take upwards of eleven years from initiation to execution (2.5 years for approvals and 8.5 for design and construction).

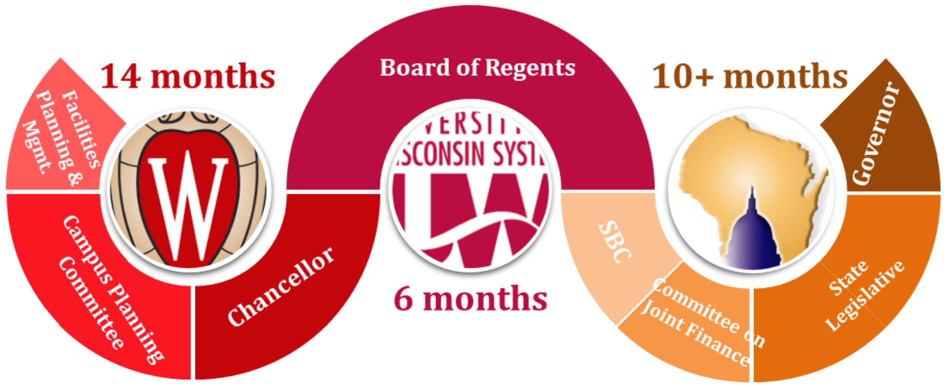


Figure 1: Projects Funding Approval Process (UW-Madison Website)

While the current approval process is necessary to ensure the best use of public funds, the results of this study indicate that potential savings could be realized if this time could be reduced. Reducing the approval process by one year could have theoretically saved $160 million (or 3.8% of the original total projects cost on average) simply due to the rate of inflation. By the same token, 2 years reduction in approvals returns $290 million in savings (or 6.8% of original total projects cost).

1.3.2 History of Design-Build in The United States

It should be noted that collaboration can be achieved, theoretically, under any delivery method as it is a function of project participants' experience, culture and practices. However, innovative delivery methods such as DB and CMAR facilitate a collaborative environment by allowing project participants to interact early on during the design and planning phases.

Public owners first began experimenting with Collaborative delivery methods like Design-Build (DB) in the early 1990's. In 1993, only 3 states (Idaho, Virginia and Florida) permitted the use of DB, as seen in Figure 2.

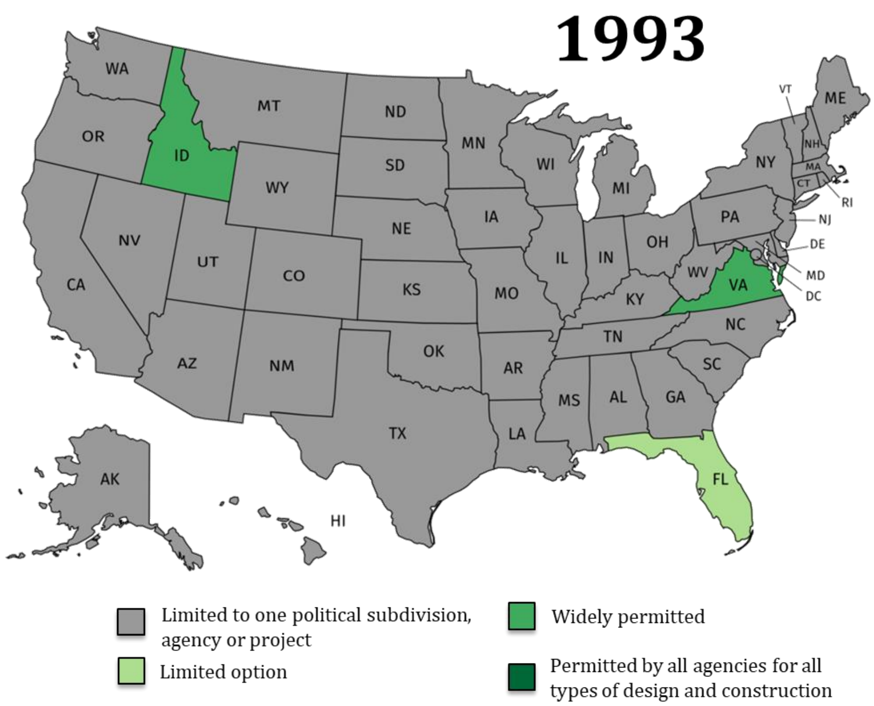


Figure 2: Design-Build Authorization Level by State in 1993 (DBIA 2018)

Ten years later, public-owner use of DB was more widespread. Wisconsin was an early adopter in the Midwest, permitting the use of DB as an alternative delivery method to DBB (Figure 3).

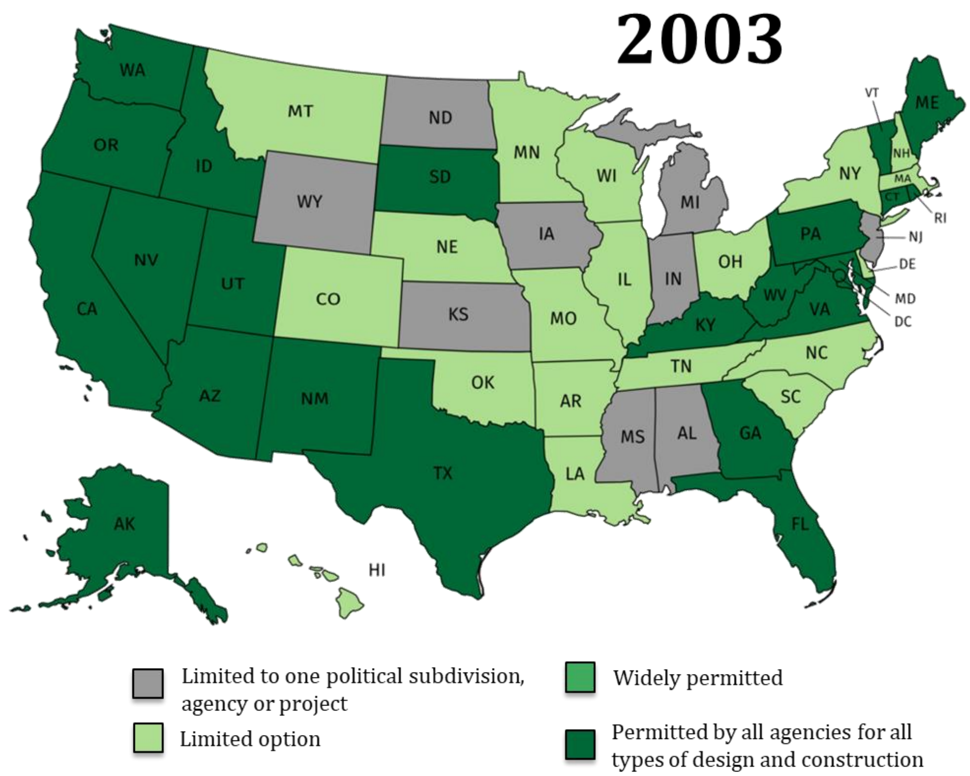


Figure 3: Design-Build Authorization Level by State in 2003 (DBIA 2018)

By 2013 (Figure 4), DB delivery methods had been adopted by over 90% of public owners in the United States, with only four states restricting the use of DB. One of these four states was Wisconsin, which, after an initial foray into DB, had chosen to restrict the use of it for public projects (DBIA 2018).

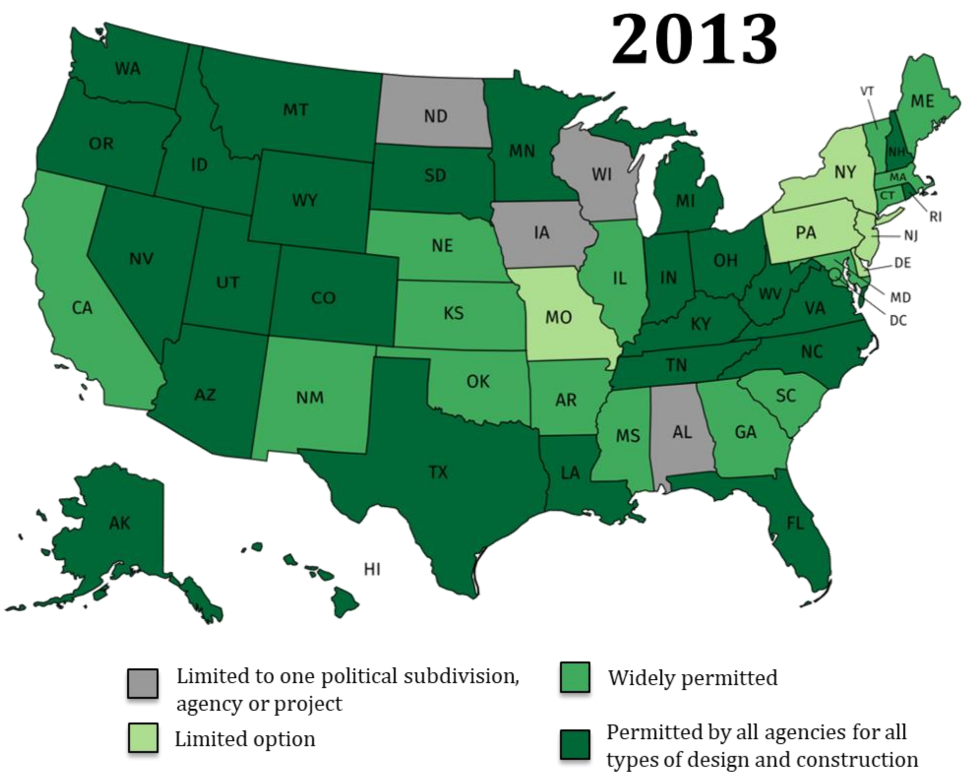


Figure 4: Design-Build Authorization Level by State in 2013 (DBIA 2018)

In the present day, only 3 states restrict the use of DB—North Dakota, Iowa, and Wisconsin (Figure 5). Note that Figure 5 divides states that use DB into three categories, according to the amount of flexibility allowed for DB processes. These categories are as follows: limited, widely permitted, and permitted by all agencies for all types of design and construction. Of all the states that do allow DB, 90% are within the two most flexible permission categories.

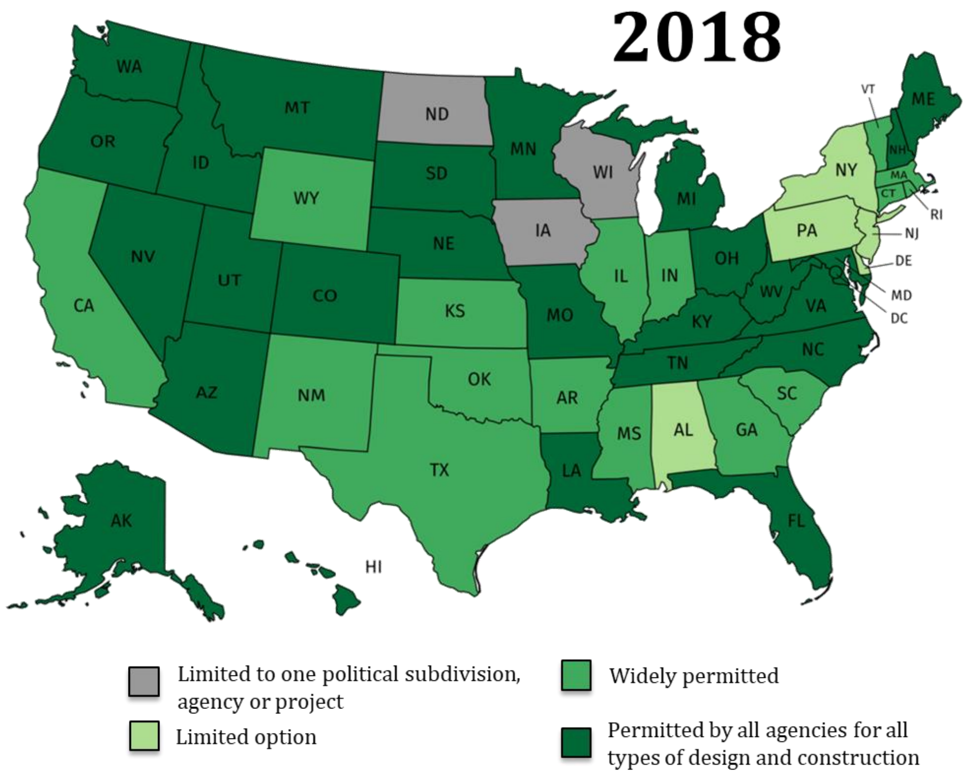


Figure 5: Design-Build Authorization Level by State in 2018 (DBIA 2018)

Figure 6, below, highlights the significant reduction in states that prohibit the use of DB since 1993. A similar trend was also found when examining the prevalence of the Construction Manager at Risk (CMAR) delivery method, which is also currently prohibited in Wisconsin while it is permitted in other states.

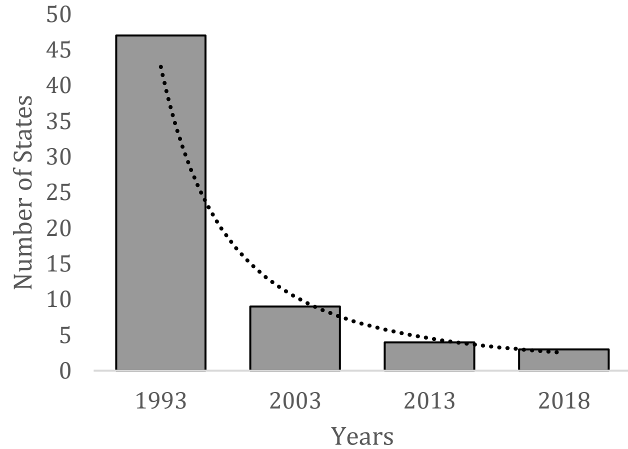


Figure 6: Number of States with Limited Design-Build Authorization Over the Years

1.3.3 The Hybrid[1] Single-Prime (SP) Delivery Method

Prior to 2013, the principal method of delivery for all Wisconsin state construction projects over $185,000 was DBB Multiple-Prime. Under this model, the state solicited bids for each major type of work and engaged in separate contracts with each of the winning firms, as explained in Figure 7 (DBIA 2014).

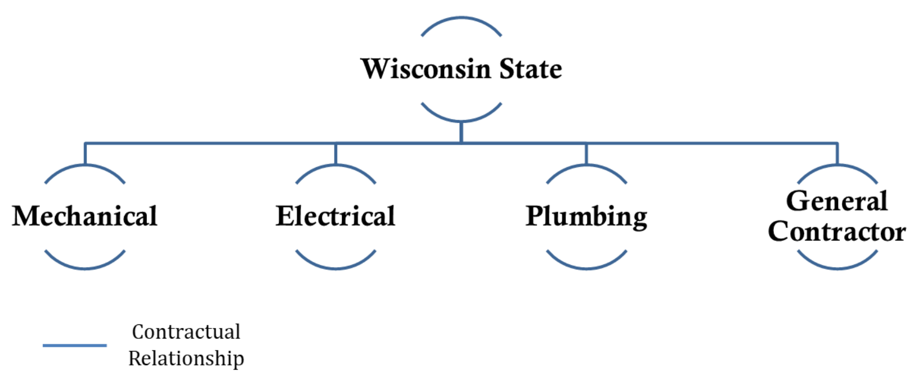


Figure 7: Structure of Traditional Multiple-Prime Contracting, Used in Wisconsin Before 2013

[1] The term Hybrid Single-Prime is used by this author to differentiate between the traditional Single-Prime structure typically applied in the industry, as explained below, and that currently utilized by the State of Wisconsin. However, both terms 'hybrid Single-Prime' and 'Single-Prime' are used interchangeably in this document to represent the system currently utilized by the State

In 2013, new Wisconsin legislation mandated a different delivery method based on the traditional Single-Prime model shown in Figure 8. In the Single-Prime model, which is widely used throughout the construction industry, the owner contracts directly with a general contractor, who in turn contracts with the pertinent subcontractors or suppliers (DBIA 2014).

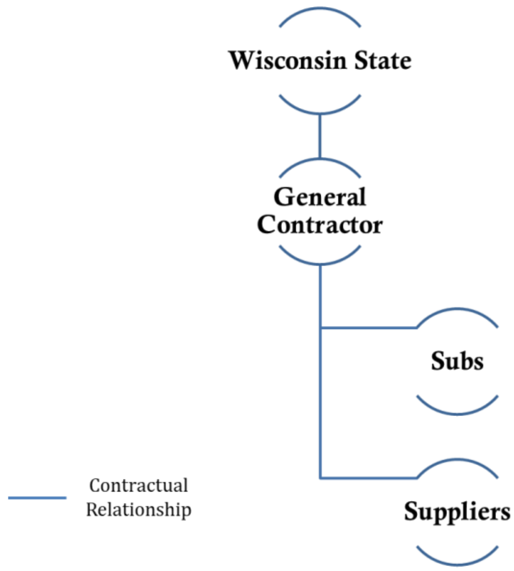


Figure 8: Structure of Traditional Single-Prime Contracting

However, in its 2013 legislation, the State of Wisconsin added a stipulation designed to prevent bid shopping by specialty contractors. The result is Wisconsin's hybrid version of Single-Prime illustrated in Figure 8. Under hybrid Single-Prime, the state preliminarily solicits bids from Mechanical, Electrical and Plumbing (MEP) contractors 1 week before the General Contractor (GC) bid is due. The state then selects the lowest specialty contractors and the GC is bound to contract with the winning MEP contractors (Wisconsin Act 20 2013).

This creates a situation in which GCs are forced to contract with subcontractors who could be unknown to them, or with whom they may have had difficulty working in the past. Often GCs have no knowledge of the subcontractors they will be working with until after the GCs have completed and submitted their bids. Forcing companies to work together like this often results in adversarial working relationships. Potentially this could lead some GCs to shy away from bidding on state projects, due to the risk involved.

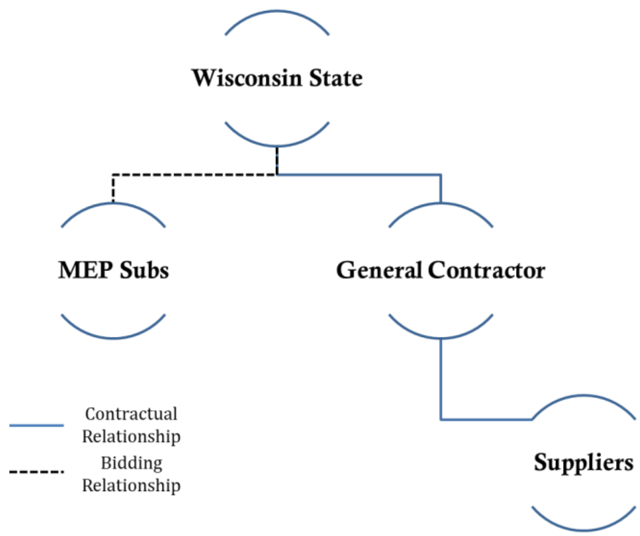


Figure 9: Structure of The Wisconsin Hybrid Single-Prime Contracting

1.4 Experience of Other States

1.4.1 North Carolina

In the year 2000, the University of North Carolina (UNC) was faced with a monumental task: completing a $4.5 billion construction program under budget and on time, with a scheduled duration of six years. State regulations at the time dated from the 1930s and permitted only the use of DBB Multiple-Prime. This restriction precluded many of the largest, most innovative, and most sophisticated general contractors from bidding on state projects, as these firms disliked the delivery method prescribed. However, in response to pressure from public entities within the state, CMAR was authorized in 2002. Since that year, UNC has initiated 94 CMAR projects with a total value of $3.3 billion. Project stakeholders realized great benefit from earlier contractor involvement, such as validating cost estimates, design modifications, and common expectations of schedule. The use of CMAR was also reported to reduce adversarial relationships and permit a merit-based selection process for CM selection (UNC 2009).

1.4.2 South Dakota

The University Center at Sioux Falls in South Dakota was also able to use CMAR delivery on a 50,000 sq. ft. Science and Technology classroom building. The project was originally bid under DBB with an estimated completion date of May 2012. However, the use of CMAR delivery reduced the planned schedule by 35%, and the project went from approval in March 2010 to occupancy in August 2011. This schedule savings was achieved in spite of extreme weather conditions due to heavy snows and harsh winter weather, as well as a notably wet and rainy spring. CMAR also permitted the university to vacate its leased space earlier, saving over $300,000 in leasing costs. In addition, CMAR allowed the university to take advantage of competitive bidding (McCluskey 2013).

1.4.3 Minnesota

The North Carolina and South Dakota examples focus on the advantages of CMAR, which is far more collaborative than Wisconsin's hybrid Single-Prime delivery method. But that is not to say that a hybrid Single-Prime method is necessarily unworkable. Indeed, Minnesota has successfully employed a hybrid version of the Single-Prime that eliminates bid-shopping (the goal of the Wisconsin hybrid Single-Prime model) and yet also allows flexibility in selection of contracting partners.

The Minnesota model works as follows:

- The mechanical and electrical contracts are bid 1 week prior to the Single-Prime due date and submitted directly to the state.
- Prime (general) contractors have 1 week to review the scope of work with participating subcontractors and are free to choose among any of the subcontractors that have submitted a bid proposal to the state.
- The posted price of the selected subcontractor cannot be changed by either the prime contractor or the subcontractor.
- Subcontractors must submit a joint bid bond payment to the state and the award prime contractor, but subcontractors can identify in writing individual general contractors with whom they will not work.
- The state selects the lowest responsive and responsible bidder among the prime contractors (Rojas 2008).

1.5 Research Statement

The scope of this research is comparing the performance of various delivery methods and the overall performance of state-funded building projects in Wisconsin. The two goals of this research are: 1) to determine if the studied performance metrics suggest that one method is objectively better than the others; and 2) make recommendations to decision makers based on these findings. The performance metrics used in this research are pegged to cost, schedule, communication, change management, and spending. They were summarized from the literature and examined both individually and collectively. Note that metrics related to complexity, quality, safety, etc. were not addressed due to lack of relevant data and/or difficulty in quantitatively representing those factors.

This study includes a data-driven mathematical model validated by statistical techniques that evaluates the relative performance of delivery methods. The model assigns each delivery method a single numerical score that is representative of all examined performance metrics. These scores provide a simple means of comparing the performance of one project or one contractor to another, allowing decision makers to identify highly performing entities.

In addition, this study includes an in-depth investigation of projects executed by UW-System campuses to determine how the studied performance metrics affected the success of these projects. This document provides empirical data and analysis thereof that UW-System decision makers can use to ensure the success of future projects.

1.6 Definition of Terms

1.6.1 Definition of Time-Related Terms

- AE Selection: the date on which the Architect is selected.
- Notice-to-Proceed: the date on which the Contractor is directed by the State to commence construction.
- Substantial Completion: the date on which the construction phase is deemed by the State (or its representative) sufficiently complete.
- Contractual Completion Date: the date stipulated in the contract as a deadline for the completion of the construction phase.
- Completion Date due to Change Orders: the date on which the contracting team is contractually responsible for delivery the project in a sufficiently complete status after including extensions of time approved by the State to the original contractual completion date.
- Final Payment Date: the date on which all payments are made by the State to all the Contracting Firm(s).
- Contractual Construction Duration: the initial contractual duration of construction calculated from the Notice-to-Proceed date to the Contractual Completion Date.
- Extended Contractual Construction Duration: the initial contractual duration of construction calculated from the Notice-to-Proceed date to the Completion Date due to Change Orders.
- Actual Construction Duration: the actual duration of the construction phase calculated from the Notice-to-Proceed date to the Substantial Completion date.
- Total Approved Extensions: the total number of extended durations to initial contractual duration in days.
- Actual Delivery Duration: the actual duration of the design and construction phases calculated from the AE Selection date to the Final Payment Date.

1.6.2 Definition of Cost-Related Terms

- Time Adjustment Year: the year 2018, the baseline for cost adjustment for time
- Location Adjustment: the city of Madison, Wisconsin, the baseline for cost adjustment for location.

- Project Adjustment Year: the year of Substantial Completion (in the case of construction contracts) or Last Encumbered Date (in the case of design contracts), as the reference for time adjustment.
- Adjustment Size Factor: the factor generated to adjust the project cost by GSF, based on the understanding that bigger projects have less cost per GSF than smaller ones. See detailed explanation in Chapter 4.
- Base Bid: the original cost stipulated in the contractual agreement between the Contracting Firm(s) and the State. In some Collaborative delivery methods, it refers to initial Guaranteed Maximum Price (GMP).
- Final Construction Cost: the actual total payment made to the Contracting Firm(s) by the State after the completion of the construction phase.
- Adjusted Final Construction Cost: the actual payment made to the Contracting Firm(s) by the State, adjusted for Time, Size and Location.
- Initial Design Cost: the original cost stipulated in the contractual agreement between the Architect and the State.
- Final Design Cost: the actual total payment made to the Architect by the State after the completion of the design phase.
- Adjusted Final Design Cost: the actual payment made to the Architect by the State, adjusted for Time and Location. See detailed explanation in Chapter 4.
- Initial Delivery Cost: the summation of Base Bid and Initial Design Cost.
- Final Delivery Cost: the summation of Final Construction Cost and Final Design Cost.
- Adjusted Final Delivery Cost: the summation of Adjusted Final Construction Cost and Adjusted Final Design Cost.

1.6.3 Definition of Other Terms

- State: the State of Wisconsin.
- Agency: any of the following: UW-System, Military Affairs, Corrections, Administration, Health Services, Historical Society, Natural Resources, or Veterans Affairs.
- Facility Type: the function of a facility; possible types include: Research, Public Assembly, Industrial, Academic, Recreational, Office, Detention Center, Food Service, Mixed Use, Maintenance, Parking, Residence Halls, Infrastructure, Treatment, Museum, Outreach, Nursing, Library, Historic, and Other.
- Architect: the architectural and engineering firm that was selected by the State to execute the design phase for a specific project.
- Contracting Firm(s): the construction team that was selected by the State to execute the construction phase for a specific project. Under certain delivery methods (such as DB) this refers to one entity, under others this refers to all the contractors and subcontractors (such as Multiple-Prime).

- Construction Type: either New Construction, Remodeling, or Addition / Remodeling projects.
- Project Delivery Method (PDM): the utilized system under which a specific project was procured. This includes: Design-Build (DB), Construction Manager at Risk (CMAR), Design-Bid-Build (DBB) Hybrid Single-Prime (SP), or Design-Bid-Build (DBB) Multiple-Prime (MP).
- Design-Build (DB): an approach in which one entity—the design-build team—works under a single contract with the project owner to provide design and construction services. The design-builder provides a single point of responsibility for the owner. (DBIA 2014).
- Construction Manager at Risk (CMAR): an approach in which the construction manager acts as a consultant to the owner in the development and design phases, but assumes the risk for construction performance as the equivalent of a general contractor holding all trade subcontracts during the construction phase (CMAA 2012).
- DBB Single-Prime: an approach in which the owner solicits bids for the entire project and contracts directly with a general contractor, who in turn contracts with the pertinent subcontractors or suppliers.
- DBB Multiple-Prime (MP): an approach in which multiple packages are separately bid and awarded to the lowest qualified and responsible prime contractors. The owner holds all prime contracts and is responsible for coordination during construction (DBIA 2014).
- DBB Hybrid Single-Prime (SP): the approach in current use by the State in which the State solicits bids from MEP contractors 1 week before the GC bid is due. The state then selects the lowest specialty contractors and the GC is bound to contract with the winning MEP contractor (Wisconsin Act 20 2013).
- Project Delivery Type (PDT): refers to the collaboration level under which the project was delivered. This includes: Collaborative and Traditional delivery methods.
- Collaborative Delivery Methods: refers to Design-Build (DB) and Construction Manager at Risk (CMAR) delivery methods due to their nature, as the general contractor is engaged early on during the design phase.
- Traditional Delivery Methods: refers to Design-Bid-Build (DBB) Single-Prime (SP) and Design-Bid-Build (DBB) Multiple-Prime (MP) delivery methods.
- Delivery: the design and construction phases of a project.
- Percent Difference: applied when comparing two experimental quantities, where neither of which can be considered the standard value. The percent difference is the absolute value of the difference over the mean times 100 (Wenning 2004-2014).
- Percent Error: applied when comparing an experimental quantity with a theoretical quantity, which is considered a standard value. The percent error is the absolute value of the difference divided by the standard value times 100 (Wenning 2004-2014).

- Project Change Order Reasons: a set of standard reasons for change orders that are allowed by the State. No clear definitions by the State for each of these reasons were found—such as what constitutes Major and Minor Scope Changes.
- Unforeseen Conditions: defined by Federal Acquisition Regulation (FAR) as subsurface or latent physical conditions that differ materially from those indicated in the contract documents and that are not able to be anticipated.
- Design Oversight: Project changes made due to design coordination issues or design errors.
- Scope Changes: the owner's modifications (additions or deletion) to the original scope.
- Initial Project Area: true project area as constructed.
- Adjusted Project Area: the project area after being adjusted for size.
- RS Means National Average: a value based on the following 30 cities:

Atlanta, GA	Detroit, MI	New York, NY
Baltimore, MD	Houston, TX	Philadelphia, PA
Boston, MA	Indianapolis, IN	Phoenix, AZ
Buffalo, NY	Kansas City, MO	Pittsburgh, PA
Chicago, IL	Los Angeles, CA	St. Louis, MO
Cincinnati, OH	Memphis, TN	San Antonio, TX
Cleveland, OH	Miami, FL	San Diego, CA
Columbus, OH	Milwaukee, WI	San Francisco, CA
Dallas, TX	Minneapolis, MN	Seattle, WA
Denver, CO	Nashville, TN	Washington, D.C.

1.6.4 Definition of Factors

- Project Percentage Change: percentage of change (increase / decrease) in project cost as compared to the Base Bid (CII Performance Metric Formulas and Definitions).
- Schedule Growth: percentage of change (increase / decrease) in project duration measured against the Contractual Construction Duration (CII Performance Metric Formulas and Definitions).
- Schedule Factor: an index that represents the amount of change (increase / decrease) in project duration measured against the Extended Contractual Construction Duration (CII Performance Metric Formulas and Definitions).
- Number of Construction Change Orders Per Million Dollars: the number of change orders made to the construction contractual agreements, divided by the Adjusted Final Construction Cost.

- Number of RFIs Per Million Dollars: the number of RFIs submitted from the construction team to the design team, divided by the Adjusted Final Construction Cost (Navigant 2013; El Asmar et al. 2015).
- RFI Processing Time: number of days a contracting firm has to wait to receive a final response to its submitted RFI (Navigant 2013; El Asmar et al. 2015).
- Construction Speed: construction spending in thousand dollars per day calculated by dividing the Total Construction Duration by the Adjusted Final Construction Cost (Konchar and Sanvido 1998; El Amar et al. 2015; Carpenter and Bausman 2016).
- Delivery Speed: project delivery spending in thousand dollars per day calculated by dividing the Actual Delivery Duration by the Adjusted Final Delivery Cost (Konchar and Sanvido 1998; Carpenter and Bausman 2016).

1.7 Assumptions and Limitations

1.7.1 Assumptions

The analysis and conclusions presented in this study are founded on the following assumptions:

- Since the Engineering, Procurement and Construction (EPC) method is very similar to DB method, with the exception of the procurement portion, and also since only 2 projects were procured under EPC, these projects were considered as DB projects.
- At the state level, the collected data can be considered a population rather than a sample since most of the projects executed under the research scope were included. Nevertheless, relevant statistical tests were conducted as this population of projects could still be considered a sample of similar projects in the Midwest, and nationally.
- When comparing the performance of individual performance metrics, metrics such as complexity, quality, safety, etc. were not considered as these data was not available.
- When assessing the schedule performance by comparing the initial and final durations, the research assumes that the initial duration set by the State is adequate for the type, size and level of complexity of the project.
- Major and Minor Scope changes were combined to form Scope Changes as no clear definitions were found for each of those two change reasons.
- Most change orders are generally thought as negative yet inevitable, as change disrupts on-site operations and planning and can lead to rework or out-of-sequence work (Hanna 2004; Riley et al. 2005).
- Quality/Value Reevaluation, Program Work Reinstated, and Other reasons were deemed too generic, ambiguous, and/or represented too small a portion of the data as compared to other reasons for change orders. Thus, these reasons were assessed statistically only for completeness. The researcher was unable to draw conclusions. Instead, the analysis was performed in full, with the goal of initiating state-wide discussion.

- Scope Changes, of the three considered in the analysis, have the most impact on the project. Unforeseen Conditions and Design Oversight are beyond the control of the contracting party and in many situations do not add value to the project. They are typically implemented to restore an aspect of the project (safety for instance) that was initially assumed to be included in the project.
- Design Oversight changes are generally the product of errors in design and coordination. Accordingly, delivery methods with higher percentages of those changes are indicative of their poorer design performance during the design phase.
- Unforeseen Conditions are assumed to be independent of delivery method and beyond the control of the contracting party. Delivery methods with higher percentages of Unforeseen Conditions changes are indicative of their reduced flexibility in mitigating and controlling unanticipated risks.

1.7.2 Limitations

Although numerous validation procedures were employed in this study to maximize the confidence level of the findings, certain limitations should be considered when evaluating the results.

The principal limitation of this study is also its strength: it was constrained to Wisconsin State projects. As such, while the results are highly applicable to the state of Wisconsin, they may not be so applicable to other states or locations with different characteristics or regulations. Nevertheless, the methodology is sound should another researcher wish to rework the findings for a different state, and the conclusions in general are applicable to many other locations as well.

Another limitation is that the total cost of construction projects should include both actual costs as well as indirect costs of coordination and management. These costs are borne by the owner—in this case, the public. The latter type of cost is difficult to quantify over a large number of projects and public authorities. The study is limited, therefore, as the unavailability of these costs made their inclusion challenging, in comparison with direct contract cost of the studied projects. Although administrative costs are significant to public owners, the estimation of such costs was beyond the scope of this study.

A mathematical model called the Project Performance Index (PPI) was developed as part of this study to provide a single numerical score to evaluate the relative performance of the various delivery methods. The PPI model does not address overall success of capital projects. Rather, the index measures only the success of the project during construction and only in light of the eight defined performance metrics. Overall success includes design performance, customer satisfaction, post-construction performance, quality, latent defects, safety, and others, for which data was not available and which was beyond the study scope.

The final limitation of this study was that it did not control for facility type in analysis or in the PPI model. As there is great diversity in the type of facility executed in public interest, the sample size of 189 projects was insufficient to make comparisons between different facility types. As such, the findings of this research reflect the relationships found across all projects, regardless of type, executed by the Wisconsin State Agencies, and it should be understood that specific paths may perform better or worse based on facility type and complexity.

1.8 Methodology

1.8.1 Research Methodology

To facilitate understanding of the research questions, the high-level roadmap shown in Figure 10 was created to define the research process and benchmark progress. There are four principal stages to this research effort.

The **first stage** is a comprehensive review of available literature concerning the performance of various delivery methods on both public and private projects. Further emphasis was placed on the various methods used to measure project performance; the statistical and mathematical techniques used. This review identified a pressing question: why does Wisconsin and two other states limit the use of DB? Secondly, this research highlighted several key gaps in the literature that have not yet been addressed, including a focus on public projects in the Midwest generally, and those in Wisconsin specifically.

Alongside the literature review, a rigorous requirement identification process was conducted, considering relevant research findings and refining the scope of the research based on the availability and quality of the stored data provided by the State of Wisconsin. Many performance metrics were identified and refined. The process involved identifying a performance metric and cross-checking it with available data to determine if it was suitable for consideration. This process relied on the experience and expertise of industry experts in State-funded projects, including project managers, architects, contractors, agency representatives, and legislators, to sift and winnow the key concerns about the current procurement and selection systems in Wisconsin and identify a promising research path.

The **second stage** was an exploration of the available means of identifying and obtaining the requisite data. Various paths were explored in collaboration with State agencies such as the Department of Administration (DOA). Three public records requests were submitted, and a workable portion of the data was obtained in April 2018.

The third stage was data preparation and analysis. The data was first prepared using treatment tools that standardized the data, making it more readable so as to facilitate later analysis. This was conducted via popular and widely used statistical techniques, and the creation of a proprietary mathematical model. This stage also featured continued involvement of industry experts to collect feedback and optimize the model for use.

The **final stage** was identifying meaningful findings concerning the relative performance of each delivery method. A set of recommendations was also produced that can act as a guide for decision makers to enhance the current procurement system and project performance.

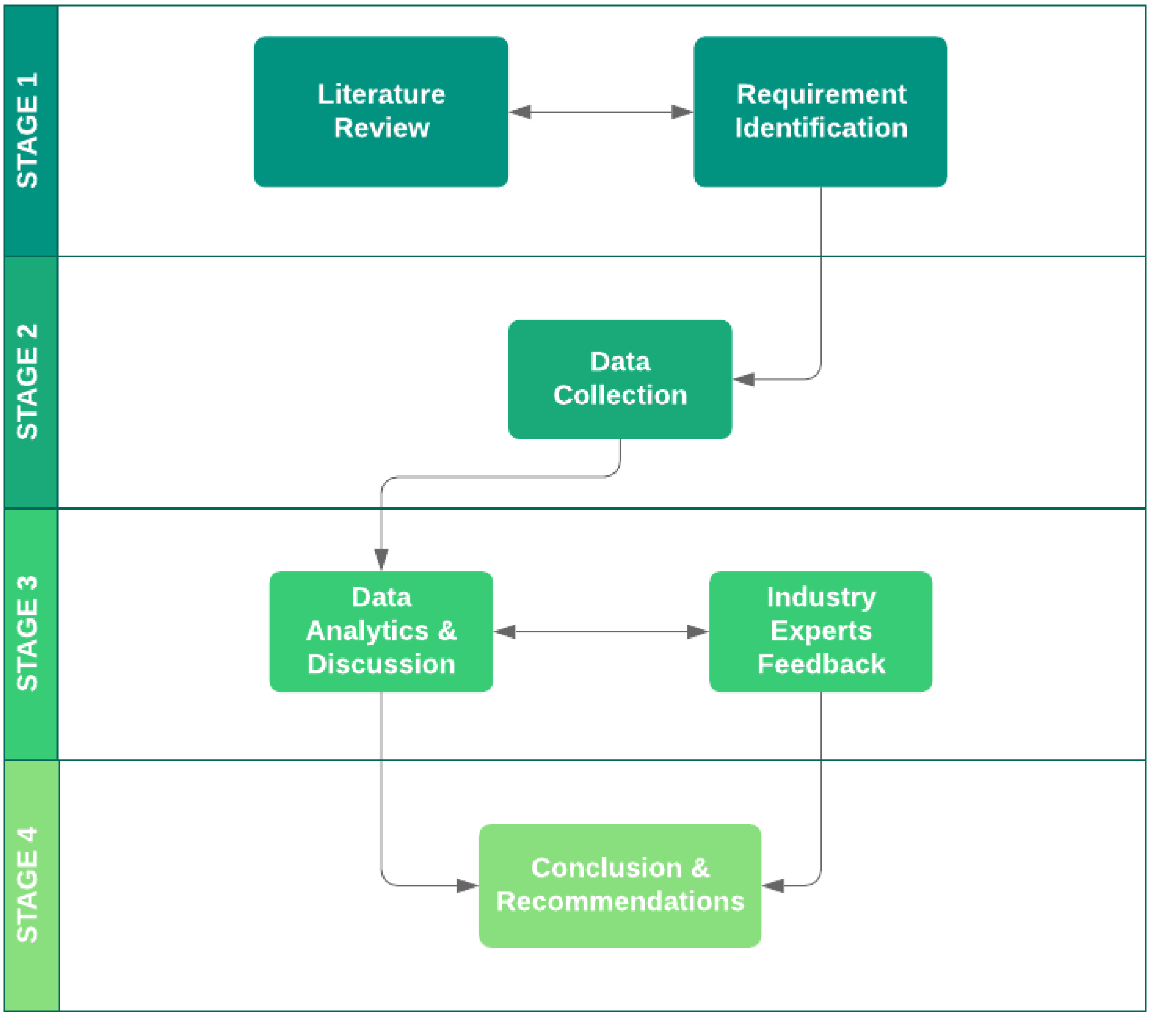


Figure 10: Research Methodology

1.8.2 Data Analytics Methodology

This section elaborates on **stage three** of the research. It lays out the analytic thought process (Figure 11 to Figure 14) so that it can serve as the basis of a higher-level discussion in later sections. It also underscores the comprehensive nature of this research and the level of scrutiny applied before conclusions were drawn.

The bulk of the analysis was conducted on continuous numerical data so as to eliminate conscious bias or unconscious error in evaluation, which is least common in this type of data.

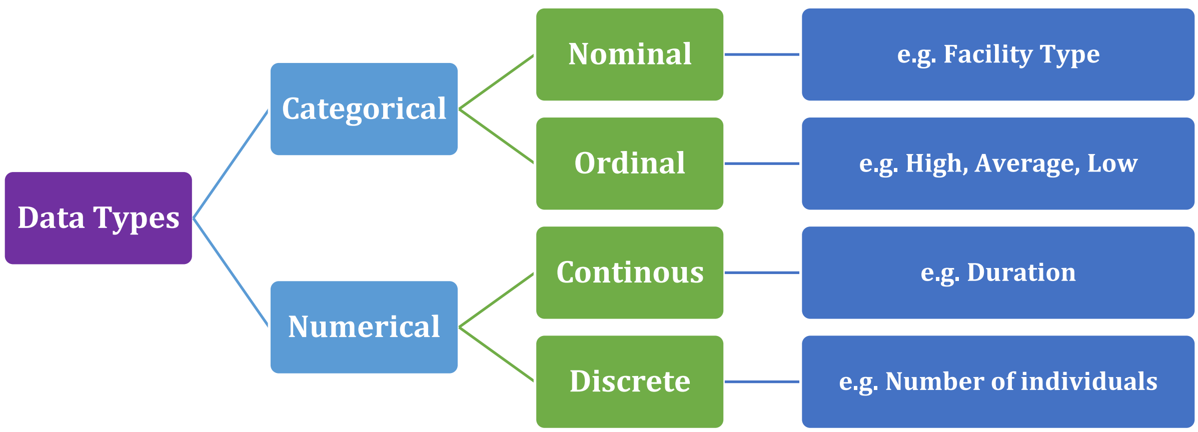


Figure 11: High-Level Division of Data Types

1.8.2.1 Testing for Proportion Significance

In this phase of the analysis, three significance tests were utilized to determine the association of two categories. The principal objective of this study was determining if any statistically significant association existed between project delivery methods/delivery types and cost/schedule changes.

Given that the data sample in this case is nominal categorical represented by the count of projects under each variable, three proportions tests such as Chi-Square (χ^2), Fisher's Exact and Aylmer (West and Hunkin 2008) tests were utilized as appropriate (Figure 12).

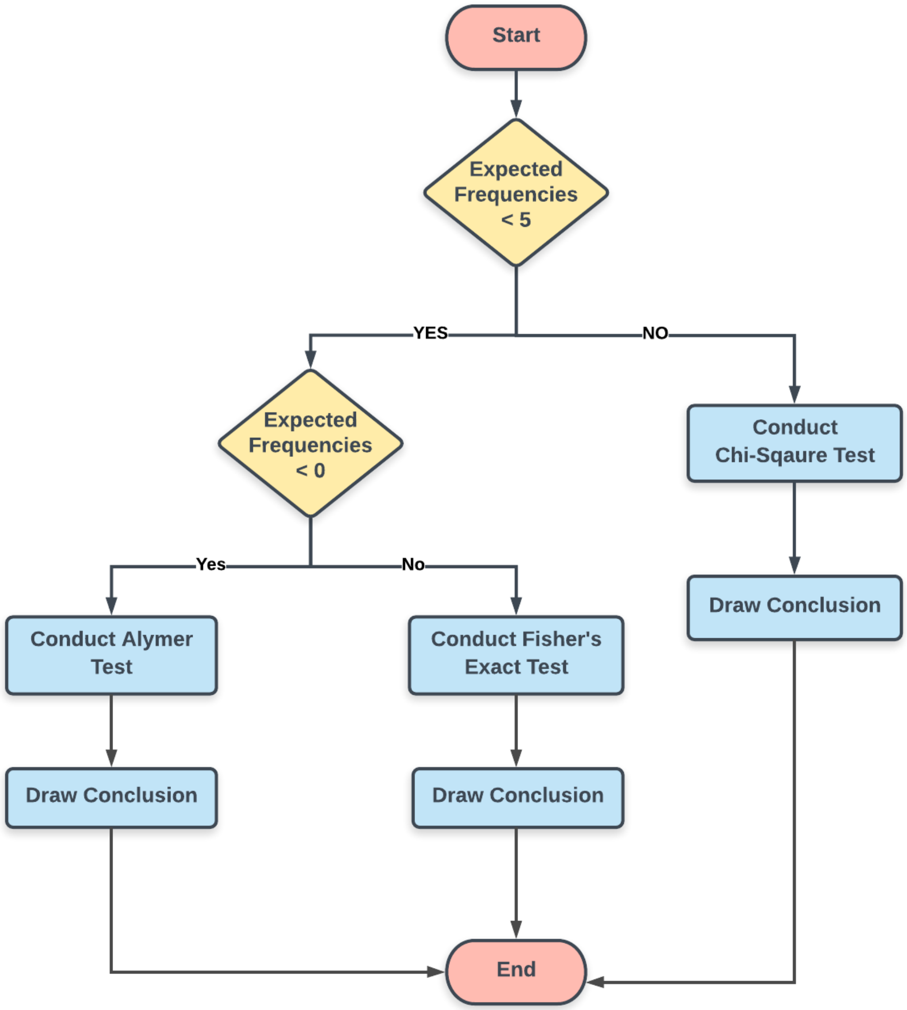


Figure 12: Methodology of Proportion Significance Tests

The Chi-Square test is the most common method for testing association between categorical variables. As such it was given priority – whenever its assumptions were met, it was used. These assumptions are:

1. The sampling method used is simple random sampling.
2. The levels (or categories) under study are each independent.
3. The levels of the variables are mutually exclusive. That is, a particular subject fits into one and only one level of each of the variables.
4. The expected frequencies are at least 5 for each cell.

In such situations where the fourth assumption was violated, Fischer's Exact test was preferred as it is able to compute exact values even with the expected frequencies are less than 5 (note the difference between observations and expected frequencies). In these situations, the Chi-Square test was still utilized for the purposes of validation, augmented by the Yates continuity correction and a Monte Carlo simulation (bootstrapping) of 2000 replicates to account for the violation of the third assumption.

However, in situations where observations had a value of zero, neither Chi-Square nor Fischer's exact test were reliable. Therefore, the Aylmer test, a generalization of Fischer's exact test, was used as it measures the association of two categories based on probability. In these situations, Fischer's exact test was used alongside a 2000 replicate Monte Carlo simulation for validation purposes.

The *p*-values that resulted from these tests were compared at the 95 and 90% confidence levels to determine if statistically significant associations existed between the examined categories.

1.8.2.2 Comparative Analysis of Continuous Data

This section contains perhaps the most important results of this research. It compares the performance of the four delivery methods under examination and seeks to identify statistically significant differences. These results also have bearing on several other key lines of inquiry – whether the type of delivery method has bearing on project performance in terms of the defined performance metrics of this study individually and collectively.

Further analysis was performed which separated the delivery methods into Collaborative (DB and CMAR) and Traditional (SP and MP). This analysis was designed to produce more concise results pertaining to the performance of Collaborative vs. Traditional methods. While it is understood that these pairs of delivery methods do not perform in the exact same manner, the decision to perform this paired comparison was motivated by the pair's intrinsic similarities in terms of collaboration, award selection method, and performance.

Figure 13 is the logical process flow by which the analysis of this section was conducted. Given that the data type in this study is continuous numerical, the appropriate significance tests were performed based on normality and equality of variance.

Normality was the first assumption examined, using both the Quartile-Quartile plot (Q-Q Plot) and the Shapiro-Wilk normality test. Next, a homogeneity of variance test was also performed to detect any significant differences in variability between delivery methods performance with respect to the performance metrics defined in this study. From these results, the appropriate Multiple-sample test was selected, as well as a post-hoc pairwise test in the event of significance. A similar method was followed for comparative analysis by delivery type.

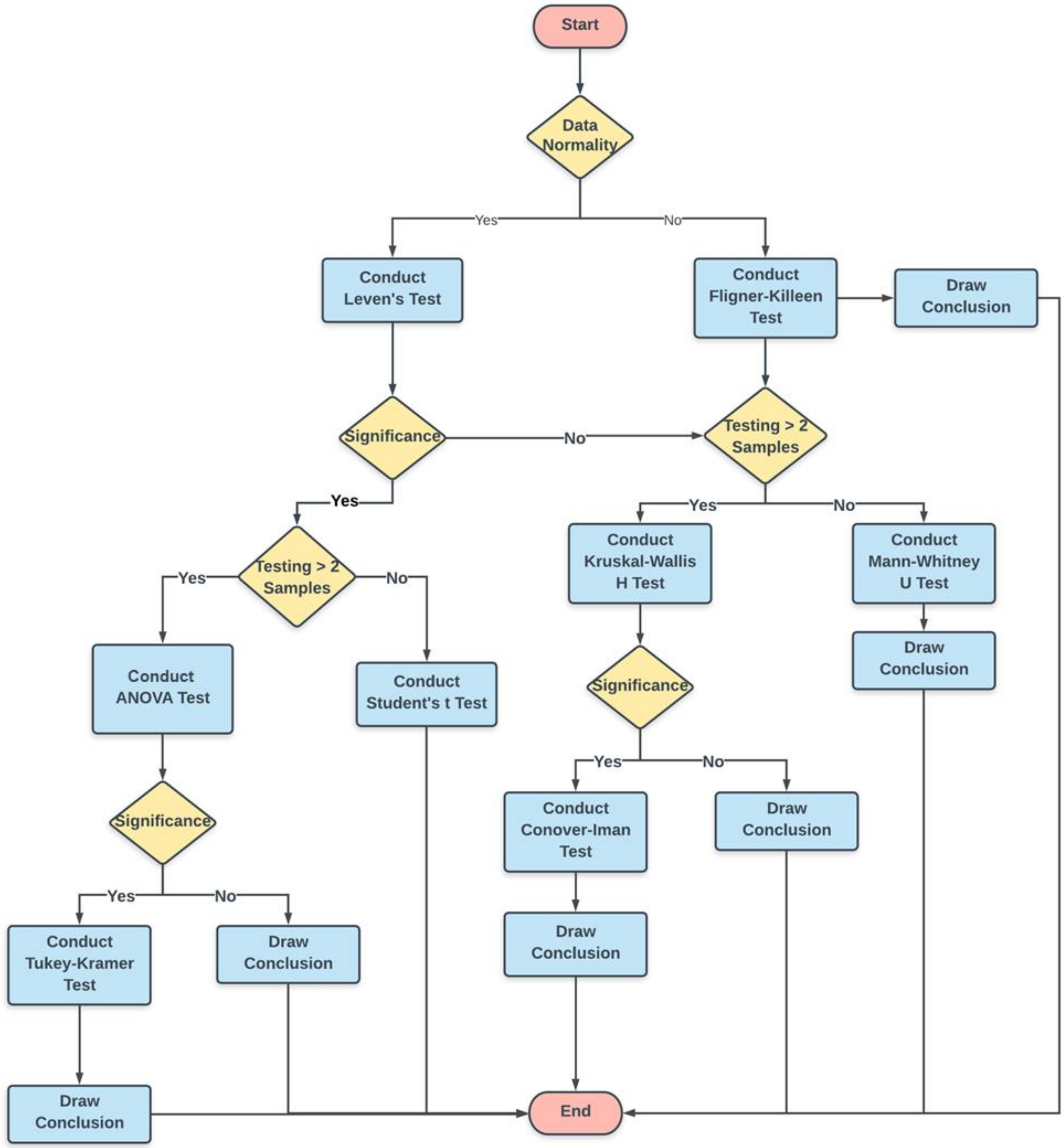


Figure 13: Methodology of Continuous Data Significance Tests

1.8.2.3 Regression Analysis

This regression analysis examines the degree of association between the eight performance metrics and the Project Performance Index (PPI) scores.

The objective of examining this relationship was to validate the contribution of each of the eight metrics to the model. Since the goal of the model was to maximize the correlation between each performance metric and the PPI score, determining the correlation between the metrics and the PPI was a method of insight into the overall success of the model. However, since correlation is dependent on the data itself, investigating the regression charts revealed which project or set of projects might had a disproportionate impact on the correlation, as well as identified specific trends in each performance metric (Figure 14).

It is important to understand that all of the data sampled in this analysis was found to be non-normal, based on normality checks. Therefore, Spearman's Correlation test was used to evaluate the hypothesis of association between the performance metrics and the PPI scores. Spearman's Rho correlation coefficient was also reported to quantify the strength of the relationship. When examining the charts in the coming sections, note that linear regression lines were drawn even though normality was not present. These lines are used solely for trend representation, rather than for any analysis.

Ellipses were added to the regression chart to identify and highlight clusters of data and trends within clusters. While the type of ellipses used are traditionally used under an assumed t-distribution, once again these were drawn solely for representative purposes and the results are approximate.

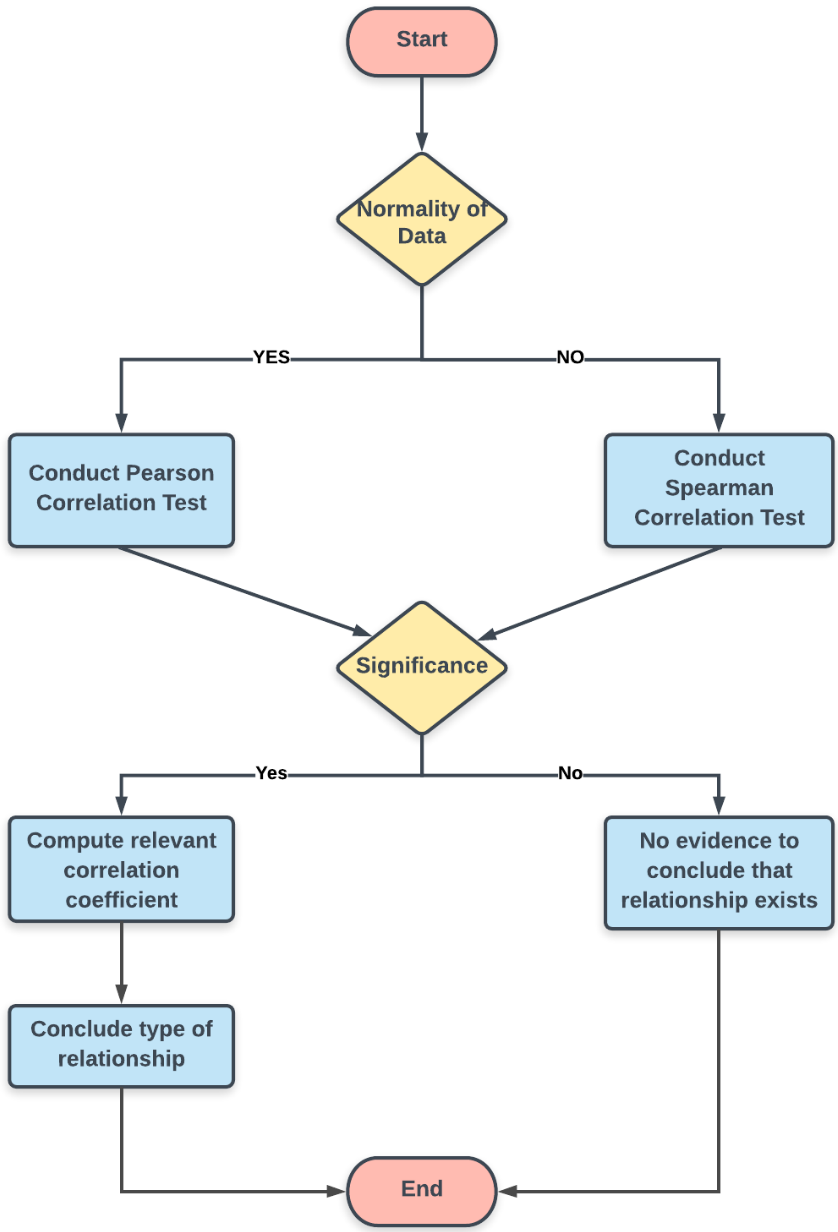


Figure 14: Methodology of Regression Analysis

1.8.2.4 Mathematical Model Validation Techniques

These analyses were performed to confirm the results of the PPI model. While the model is proven mathematically sound in later chapters, it is important to confirm the results secondarily using prevailing statistical methods.

In addition to the iterative mathematical validation explained in Chapter 2, and the previously examined regression relationships, two further independent statistical tests were performed to support the accuracy of the model empirically. These techniques were Partitioning Clustering Analysis and Confirmatory Factor Analysis (CFA), as they are commonly used dimensionality reduction techniques.

Confirmatory Factor Analysis reduces highly-dimensional data without overly reducing the variance. It operates in a similar manner to the mathematical model, with a subtle distinction. This technique produced a comprehensive performance index represented by the Factor component. This was then used to replicate the boxplots for the various project methods and types for comparison to those generated by the PPI. The technique was also used to generate three factors that capture higher variance and provide more detailed insights about the performance of the different delivery methods in a 3-dimensional representation.

Partitioning Clustering Analysis groups projects by their dissimilarity in performance, and then investigates which delivery methods were grouped together based on performance. Thusly, no presumption of delivery method was present, and the analysis focused solely on how a project performed relative to other projects in its cluster and other clusters. This analysis clearly indicates which cluster(s) of projects outperform similarly or differently to other clusters.

Confirmatory Factor Analysis (CFA)

Factor Analysis is a well-known multivariate statistical tool that is used to reduce high dimensional space data into less dimensions, while maintaining a decent proportion of the data variance captured. CFA is the technique used to verify a pre-determined factor structure of a set of observed variables.

Despite being very like Principal Component Analysis (PCA), Factor Analysis is different with the assumption that the variance in the observed variables is due to the presence of one or more latent variables that exert causal influence on these observed variables. PCA instead makes no assumption about an underlying causal model but is rather a variable reduction procedure that aims at accounting for most of the variance in a set of observed variables using relatively small number of components.

Factor Analysis is characterized by a rotation method, which is either oblique or orthogonal, and an extraction method, which encompass a wide range of algorithms, each underlying different set of assumptions. In an attempt to select the most appropriate rotation method, the selection procedure provided by Tabachnick and Fiddell (2007, p. 646) was utilized as follows:

1. Start with an oblique rotation with a standard oblimin algorithm.
2. If generated factor correlations exceed 0.32, there is 10% (or more) overlap in variance among factors, enough variance to warrant oblique rotation.
3. If factor correlations do not exceed 0.32, an orthogonal rotation can be utilized.

Furthermore, since there is no consensus on the application of each extraction technique, trial-and-error technique while keeping the underlying assumptions of each extraction technique in mind were followed, in addition to considering those that generated the most relevant output to a simple structure as reported by Bryant and Yarnold (1995, p. 132-133).

With that, oblimin rotation algorithm was used as one of the most common oblique methods. As for the extraction techniques, various techniques were tested and compared including Ordinary Least Squares (OLS), Unweighted Least Squares (ULS), and a Maximum Likelihood (ML), Generalized Least Squares (GLS) and Principal Axis Factoring (PAF). The best results that satisfied the underlying assumptions and most complied with the following thresholds were selected (Hooper et al. 2008; Kline 2011):

- Empirical Chi-Square *p*-value should be above 0.05.
- Root Mean Square of Residuals (RMSR) should be close to 0.0.
- Root Mean Square Error of Approximation (RMSEA) should be below 0.1.
- Tucker-Lewis Index (TLI) should be above 0.9.

It is worth mentioning that empirical chi-square *p*-value was reported and compared against the benchmark at the 95% confidence level instead of the likelihood chi-square. The likelihood Chi-Square test generates too many Type 1 errors when variables have non-normal distributions; whereas, the empirical approach utilize a nonparametric setting, and thus, can better deal with non-normal data (Owen 1988; Owen 1990).

Partitioning Clustering Analysis

In this section, unsupervised analyses were performed to group a sample of 188 collected projects based on their dissimilarity, with the objective of creating clusters with maximum dissimilarities between the different groups (inter-cluster) and minimum dissimilarity between those projects within the same group (intra-cluster).

The most common portioning cluster algorithms are k-means, PAM (Partitioning Around Medoids) and CLARA (Clustering Large Applications). CLARA is an extension to PAM specifically designed for larger datasets (several thousand observations) in order to reduce computing time and RAM storage problems. Therefore, k-means and PAM are utilized instead due to their applicability to our research objectives and dataset.

K-means clustering (MacQueen 1967) is one of the most commonly used unsupervised machine learning algorithm for partitioning a given data set into a set of k groups (i.e. k clusters), where k represents the number of groups pre-specified by the analyst. It classifies objects in multiple groups (i.e. clusters), such that objects within the same cluster are as similar as possible (i.e., high intra-cluster similarity), whereas objects from different clusters are as dissimilar as possible (i.e. low inter-cluster similarity). In k-means clustering, each cluster is represented by its center (i.e. centroid) which corresponds to the mean of points assigned to the cluster.

The algorithm starts by randomly selecting k objects from the data set to serve as the initial centers for the clusters. The selected objects are also known as cluster means or centroids. Next, each of the remaining objects is assigned to its closest centroid, where closest is defined using the Manhattan distance between the object and the cluster mean.

After that, the algorithm computes the new mean value of each cluster. With each center recalculated, every observation is checked again whether it can be closer to a different cluster and reassigned to a different cluster if true. This iterative process is repeated until the cluster until convergence is achieved. That is, the clusters formed in the current iteration are the same as those obtained in the previous iteration.

On the other hand, the k-medoids algorithm (Kaufman and Rousseeuw 1990) is a clustering approach related to k-means clustering for partitioning a data set into k groups or clusters. In k-medoids clustering, each cluster is represented by one of the data points in the cluster. These points are named cluster medoids.

The term medoid refers to an object within a cluster for which average dissimilarity between it and all the other the members of the cluster are minimal. It corresponds to the most centrally located point in the cluster. These objects (one per cluster) can be considered as a representative example of the members of that cluster which is useful in identifying the performance of each cluster comparatively.

The algorithm is based on the search for k representative projects or medoids among the observations of the data set. After finding a set of k medoids, clusters are constructed by assigning each observation to the nearest medoid. Next, each selected medoid and each nonmedoid data point are swapped and the objective function is computed. The objective function corresponds to the sum of the dissimilarities of all objects to their nearest medoid.

The swap attempts to improve the quality of the clustering by exchanging selected objects (medoids) and non-selected objects. If the objective function can be reduced by interchanging a selected object with an unselected object, then the swap is carried out. This is continued until the objective function can no longer be decreased. The goal is to find k representative objects which minimize the sum of the dissimilarities of the observations to their closest representative object.

The first step in the analysis was to create a dissimilarity matrix, a square symmetric matrix that expresses the similarity between each pair of observations. The diagonal members are defined as zero, meaning that zero is the measure of dissimilarity between an element and itself. To represent dissimilarity, the distance between each pair of standardized observations was utilized. While Euclidean distance is the most common computation method, Manhattan distance was utilized instead due to its robustness to outliers as compared to the Euclidean distance:

The Euclidean distance is the root sum-of-squares of differences between observations

$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

The Manhattan distance is the sum of absolute distances between observations

$$d_{man}(x, y) = \sum_{i=1}^{n}|x_i - y_i|$$

Following that, partitioning clustering algorithms were applied to determine the most appropriate. During the selection process, various criteria were defined including the silhouette width (Si) approach, which is the mean similarity of each observation to its own cluster minus the mean similarity to the next most similar cluster. With that, the higher the silhouette width, the higher the evidence of a well clustered dataset. This heuristic approach was interpreted as follows:

- Large silhouette Si (almost 1) are very well clustered.
- Small Si (around 0) means that the observation lies between two clusters.
- Negative Si are probably placed in the wrong cluster.

After selecting the most appropriate clustering method, the clustered projects were then re-inspected to investigate which groups of projects performed relative to the others, in light of the original delivery methods and delivery types.

Although k-means is one of the most common methods, it experiences a set of limitations that should be well understood. The final results are sensitive to the initial random selection of cluster centers. This is essential to understand since every different run of the algorithm on the same dataset might result in slightly different results as the set of initial centers change. Nevertheless, this issue was overcome by electing 100 random sets. With that, the clustering algorithm will start by defining 100 initial random configurations and will report only those with the best results.

A second issue with k-means is its sensitivity to outliers. Because the algorithm depends on average values, it is exposed to the influence of outliers. As a result, k-means was elected only when generating two clusters since it reported higher Silhouette value compared to PAM (0.35 vs 0.31), in addition to the lower anticipated influence of outliers when only 2 clusters are under examination. Nevertheless, as it gets more sensitive to classify the data into 4 clusters, PAM algorithm was utilized instead since it is less sensitive to noise and outliers. As opposed to k-means, PAM uses medoids as cluster centers instead of means, which is more robust to outliers.

2 Chapter 2: Literature Review

In this section, previous studies that were reviewed during the development of this research will be discussed. This literature review is divided into several sections: individual performance metrics for different delivery methods, multivariate performance assessment, and the impacts of delivery methods on change management.

2.1 Performance of Project Delivery Methods

Konchar and Sanvido (1998) conducted a study in conjunction with the Construction Industry Institute (CII) to compare the performance of DBB, CMAR and DB projects. Their data included 351 public and private projects in the U.S. The metrics of comparison were unit cost, cost growth, schedule growth, construction speed, delivery speed and intensity, turnover, system quality and equipment quality. In sum, this study examined cost, schedule and quality performance.

Konchar and Sanvido found that DB provided higher performance in terms of cost and schedule, followed by CMAR which in turn outperformed DBB. DB, however, performed equally similar to CMAR in terms of quality, but both DB and CMAR outperformed DBB in that metric. It was also found that DB experienced less change than DBB to a magnitude of 5.2%.

Ibbs et al. (2003) conducted a comparative analysis of DB and DBB projects using a sample of 67 projects from the Construction Industry Institute (CII) database. Their study considered three performance areas: cost (in terms of construction cost change, design cost change, and total cost change), schedule (in terms of schedule change during construction, schedule change during design, and total schedule change), and productivity.

Ibbs et al. found that DB experienced on average 4% lower cost increase during construction, while DBB experienced 9%, and very similar results for the design cost change. This result was further explained by the fact that DB contractors have a higher propensity to innovate during a project, as well as a higher degree of communication with the design team. Both of these factors facilitate better constructability reviews and overall cost savings. Similarly, DB projects were found to experience 7.7% schedule change, compared to DBB which experienced 8.8%.

Col Debella and Ries (2006) compared Single-Prime (SP), Multiple-Prime with Agent (MP-A), and Multiple-Prime (MP) delivery on 94 public school projects executed in the Northeastern United States. This study utilized quantitative metrics of cost, schedule, and litigation, and qualitative metrics for punch-list, startup, callbacks, administrative burden, team communication, team chemistry and project complexity. The study used all identified metrics to test the data set, as well as a subset of only those projects with a total value that exceeded $10 million.

It was found that MP-A outperformed MP by 52% and SP by 73% in the whole dataset in terms of construction speed. In the subset, MP-A outperformed MP by 22%. In terms of change orders, it was suggested that MP-A experienced an increased volume of change orders compared to MP, but the opposite trend in terms of litigation. This suggests that MP-A tends to have more effective project management, resolving cases via change orders before it would be necessary to enter litigation.

No other significant relationships were found between schedule, quality, communication, and the qualitative metrics in the 2006 study. In part, this was due to the researcher's decision to eliminate data subsets that did not follow a normal distribution, rather than employing nonparametric and/or pairwise testing methods. Another limiting factor was the relatively small dataset on which this research was conducted.

Their study also highlighted a phenomenon that is often present in public projects: low response rate and defective information in responses receives. Col Debella and Ries distributed 1,862 surveys to collect data. They received only 334 responses. Of those responses, 60% (194 respondents) were unwilling to provide the requested information. Of the remaining 40% (140), a further 25% (35) provided defective or incomplete information. Many other researchers have experienced difficulties extracting high-quality information from public entities, which consequently limits the research team's ability to fully explore all the significant variables and provide solid conclusions.

Rojas and Kell (2008) performed a deep cost performance analysis on CMAR and DBB delivery on 297 school projects in the pacific northwest region of the US. However, only 24 projects in this dataset were CMAR.

It was found that while CMAR experienced 4.74% cost change as opposed to DBB at 6.29%, this result was not statistically significant at the 95% confidence level. Even after the scope was altered to exclude projects with a total value less than $5 million, clear differences were observed but they were not at the threshold of significance.

Using pre-bid estimate as a benchmark, the 2008 study examined cost growth using a set of 6 CMAR and 222 DBB projects from Washington state. It was found that DBB projects experienced -3.09% cost growth from estimate to contract award and 6.29% growth from award to final cost. CMAR experienced 10.5% from pre-buyout estimate to Guaranteed Maximum Price (GMP) and only 4.74% during construction. This suggests that owners will realize low initial construction price in DBB projects at the expense of increased change order and litigation costs. CMAR minimizes these after-the-fact costs at the expense of higher upfront costs.

Although Rojas and Kell's study was well-executed, providing in-depth insights and analyses, there were limitations. The projects studied had a wide range of execution, 1987 to 2005, and no time adjustment factors were reported. Also, t-test was used to study significance, even though the central limit theorem was violated. In that case, nonparametric tests ought to have been used. Finally, various sources were used in the data collection, which increases the potential of error in reporting comparable figures from one source to the other.

Rojas (2008) conducted a separate study in which he examined the cost performance of Single-Prime and Multiple-Prime delivery methods on an expansive dataset of 1000 projects in 11 states. Three metrics were identified for this study: final cost v. executed cost, bid cost v. executed cost, and final cost v. bid cost.

It was found at the 95% confidence level that MP experienced 5% lower final v. estimated costs than SP. It was explained that 80% of this difference was attributable to the lower bid cost to estimated cost ratio in the case of MP, which was on average 4% at the 95% confidence level. Further, SP was found to have increased costs to the GC acting in the prime capacity, as subcontractors had a tendency to increase their prices to factor contingency against late payments and bid shopping. As such, SP bids include overheads and administrative costs that MP bids do not. MP bids also do not include the increased involvement of the owner and the higher risks of litigation and coordination.

The results of this study were expected under the Theoretical Building Model, which returns the probability of obtaining similar bid prices under both SP and MP delivery. Under this model, it is overwhelmingly unlikely that SP and MP will have equal bid costs. This is a result of the low probability of selecting the lowest bidding prime contractor who in turn had also selected the lowest bidding MEP contractors. As such, the best SP bid could only strive to be as low as the lowest MP bid, and seldom could beat it. Nevertheless, the Theoretical Building Model has many limitations, including market conditions and associated risk.

Furthermore, this study identified another key issue in studying the performance of public projects – data storage and management. The study found that while many states did choose to participate, other states did not have the requisite infrastructure to store data in a sufficient way for researchers to operate. Further, some states did not maintain records even if infrastructure was present.

Hale et al. (2009) investigated and compared the performance of DBB and DB projects executed on behalf of the Naval Facilities Engineering Command (NAVFAC). Data was collected from 39 DBB and 38 DB projects executed between 1995 and 2004, from which performance analysis was performed on three cost metrics (cost per bed, fiscal year duration per bed, and cost growth) and four schedule metrics (project duration per bed, fiscal year duration per bed, construction start duration per bed and time growth).

Cost growth was found to be statistically significantly different at the 95% confidence level – DB experienced only 2.0% while DBB experienced 4.0%. Also found, though not at a significant level, was the favorable performance of DB in terms of cost per bed and other cost metrics.

DB also statistically significantly outperformed DBB in all of the schedule performance metric: project duration per bed, fiscal year duration per bed, construction start duration per bed, and time growth, with values of 2.6, 3.6, 2.6 and 76.4, compared to 7.0, 5.1, 3.7 and 193.8 respectively.

While the outcomes of this study are consistent with the mainstream body of literature, it is crucial to understand that some inaccurate practices embedded in the analysis may have impacted the results. The metric of time growth, for example, was measured and tested in days, with no note of standardization. Since growth in days inevitably varies with project size, this metric ought to have been standardized as a percentage or by unit. Nevertheless, the significance of the findings is sufficient to provide evidence for DB over DBB.

Pishdad-Bozorgi and de la Garza (2012) studied the impact of DBB and DB on claims from and against the owner. This study highlighted the adversarial relationship among project parties and identified that it stems from the natural fragmentation of DBB as a delivery method – poor communication and low collaboration among project parties, and the trend of passing-off risk from party to party. DB, by contrast is characterized by higher levels of collaboration and integration.

When examining claims submitted against the owner, the study found that DB delivery reduced claims relating to errors and omissions, misrepresentations, and differing site conditions. In terms of claims stemming from change orders, it was found that DB is capable of reducing those claims, compared to DBB, when the owner was well equipped – with scope and requirements. However, it was found that uncertainty about project scope increases change orders, and correspondingly increases claims.

On the other hand, claims against contractors were found to be reduced under DB delivery, especially claims related to cost overrun and schedule delay. However, claims related to nonconformity and defective work remain arguable. It was found that DB delivery can increase these types of claims, given that less design information is initially available when fixed costs are set, which can lead to claims when the design-builder feels its profits are threatened. The study further found that there was a higher risk of scope-related disputes due to poor communication between owner and contractor and high-level specifications by the owner. Conversely, it was found that DB can occasionally act as a performance incentive and quality motivator, as the design-builder is unable to blame the designer solely for poor quality. Also, DB's high level of communication and interaction among project parties can be very beneficial.

Carpenter and Bausman (2016) reported slightly different results. They studied 137 public school projects in Georgia, Florida, North Carolina and South Carolina. The study compared the performance of CMAR and DBB using metrics of cost, schedule and quality. Although this study reported that CMAR was outperformed by DBB in all cost metrics, the tabulated results highlight no statistically significant difference between DBB and CMAR in terms of construction cost growth and project cost growth. However, the results highlighted that DB performed better than CMAR in terms of unit cost and student cost. It was further found that CMAR was lower on average in construction schedule growth, project schedule growth and project intensity (m^2/day) but the sample size was not sufficient to conclude a statistically significant difference. Additionally, CMAR outperformed DBB in terms of project intensity ($/day).

Quality performance, on the other hand, was the most evident sector of CMAR's better performance, as it outscored DBB in terms of overall service, planning, cost control, schedule control, quality control, communication and cooperation. Similar results indicated the higher performance of CMAR in factors relating to design team service quality, including capture owner vision, complete documents, clearly defined documents, timely responses, communication and cooperation and collaboration. Although this data was provided from personnel feedback, and as such may be subjective, it is beneficial as it captures owner satisfaction on those projects.

Finally, this study examined performance in terms of claims and warranty issues. The data did not have sufficient sample size to examine the number of reported claims, but the study reported that CMAR projects had a higher percentage of zero claims and a lower percentage of 1-3 claims as compared to DBB. It was further reported that CMAR outperformed DBB in terms of costs associated with warranty issues, with a higher mean of claims with $0 in associated costs but a lower mean for claims with more than $1 million in associated costs – this represents CMAR's greater ability to control the cost implications of warranty and callback issues.

Franz et al. (2017) studied latent factors that differentiate the performance of delivery methods. Their study raised the simple to state but complex to understand question: why and how does the performance of a project change? To answer this question, 204 projects were collected from across the United States, 62% publicly funded and 38% privately funded. Six quantitative metrics pertaining to cost and schedule performance, along with six qualitative metrics related to quality performance. Six further team integration factors and five group cohesion factors.

Their results indicated that group cohesion had a positive impact on cost growth, facility quality, and turnover experience, while projects with high team integration performed highly in terms of schedule growth and intensity. Furthermore, public owners experienced statistically significant increases in unit cost, schedule growth, significantly slower delivery speeds and construction speeds, and poorer turnover performance as compared to private owners. As such, publicly owned projects performed worse than privately funded projects in terms of these metrics.

This research further divided their dataset into five sub-classes based on project characteristics. This division allowed the researchers to evaluate projects based on their levels of integration and cohesion, rather than their pre-defined delivery methods, thus standardizing their definitions of each class type to reduce variation of owners and states. The study found that delivery methods that involved the builder early on in the design process were associated with higher team integration, which strongly suggested that team integration is highly dependent on the involvement of project parties in timely manner. Furthermore, delivery methods that included open-book contract terms and qualification-based selection had a strong association with group cohesion, suggesting that a more merit-based selection process and cost transparency are prerequisites for project team cohesiveness.

Sullivan et al. (2017) made an excellent contribution to the body of literature, comparing the performance of DB, CMAR, and DBB projects by compiling 30 studies and over 4,600 projects. This study focused on five performance metrics: unit cost, cost growth, schedule growth, delivery speed, and quality. Per the analysis of Sullivan et al., no single delivery method was better in terms of unit cost, which is insightful as there is a prevailing opinion that DB and CMAR are inherently more expensive than DBB, thus justifying a higher cost growth of a DBB project. Not so.

It was found that DB performed the best in terms of controlling cost, with a weighted average cost growth of 2.8%, compared to CMAR (5.8%) and DBB (5.1%). The results were also investigated statistically, and significant differences in cost growth was found between DB and DBB. Moreover, DB was found to be an outperformer in terms of schedule growth at 10.2% (CMAR 10.4%; DBB 18.4%). DB was also found to be statistically significantly quicker, by a factor of 35%, in terms of delivery speed.

Shrestha and Fernane (2017) attempted to evaluate the performance of DB vs. DBB on public university building projects. They studied 77 projects completed since 2000 in 10 states. Four cost metrics, three schedule metrics and four change-order metrics were used.

The analysis found that DB was statistically significant in terms of contract award cost growth (-10.4%) compared to DBB (-3.5%) at the 95% confidence level. However, this study confirmed the previous conclusion that there is no statistically significant difference between DB and DBB in terms of cost per square foot, a finding also reported by Sullivan et al. (2017).

Furthermore, the study concluded DB outperformed DBB in all schedule-related metrics at the 95% confidence level, experiencing -5.3% design and construction schedule growth (DBB: 7.5%), 5.3% total schedule growth (DBB: 30.1%) and 194 ft^2 per day of construction intensity (DBB: 68 ft^2)

In terms of change orders, DB was found to statistically outperform DBB in the number of change orders in both design and construction. These conclusions, however, require further investigation the number of change orders are a factor of project size, which the study did not account for.

Ibrahim et al. (2018) compared four delivery methods, Integrated Project Delivery (IPD), DB, CMAR and DBB over a set of 109 projects. This study considered 12 performance metrics over 5 performance areas: cost, schedule, quality, safety, communication and change management.

This study found that in the twelve metrics, IPD outperformed DBB in ten. IPD was also found to outperform CMAR and DB in two metrics each. DB was also found to outperform DBB in eight metrics, while CMAR outperformed DBB in five metrics. The table below summarizes these comparisons.

Table 1: Statistical Significance Summary of Pairwise Comparisons

Compared PDSs	IPD – DBB	IPD – DB	IPD – CM	DB – CM	DB – DBB	CM – DBB
Cost growth	**				**	**
Schedule growth	**				**	**
Punch-list items	**				**	
Rework						
Overall systems quality	**	**				
OSHA recordable incidents						
# of RFIs per $ million	**		**		**	**
RFI processing time	**				**	**
Project percentage change	**				**	**
Design-related changes	**				**	
Quality/value-related	**				**	
Change orders processing time	**	**	**			

**statistically significant at the 95% confidence level

2.2 Development of Project Success Indices

Griffith et al. (1999) examined the association of pre-project planning and project success using 5 industrial projects with values that exceeded $5 million. The study identified seven success variables, which were then entered into an index in a stepwise fashion, in conjunction with Chronbach's alpha test for index reliability. As a result, the analysis found that only four success variables had a positive contribution to the index: budget achievement, schedule achievement, design capacity, and plant utilization. Using the responses to a series of interviews conducted by the research team, weights for each variable were determined as a success index equation was developed:

$$Success\ Index = 0.6 * (0.55 * B + 0.45 * S) + 0.4 * (0.7 * C + 0.3 * U)$$

A similar approach was used to develop a pre-project planning index using six variables, then correlation between pre-project planning index and success index was performed, resulting in a R^2 of 0.42. The study, however, did not test the developed models on actual projects nor attempted to compare the different delivery methods.

Lam et al. (2007) gathered data from DB projects executed in Hong Kong, from which they derived 11 criteria of success. By focusing on the highest ranked criteria, the study selected time, cost, quality, and functionality as success KPIs for DB projects. Principal component analysis was then used to weight each of these criteria and the following project success equation was then derived:

$$PSI - D\&B = 0.54 * Time + 0.55 * Cost + 0.47 * Quality + 0.42 * Functionality$$

Furthermore, Lam et al. compiled 40 responses from DB experts on their project performance, which were used alongside the aforementioned success equation to compute benchmark scores for each project and for DB projects as a whole.

El Asmar et al. (2015) developed an innovative unified score, the Project Quarterback Rating (PQR), which was derived from 23 performance metrics in 7 areas: customer relations, safety, schedule, cost, quality, profit and communication. With the help of 35 participants, weights for each of the seven performance areas were obtained using the Project Quarterback Rating (PQR), resulting in the following equation:

$$PQR = 0.45 * Relations + 0.34 * Safety + 0.31 * Schedule + 0.25 * Budget + 0.23 * Quality + 0.22 * Profit + 0.17 * Communication$$

Using the developed weights, the study computed a success score for each of the 35 projects and compared them based on their delivery method (DBB, CMAR, DB and IPD). The results showed an increasing trend in the average PQR score starting with DBB (-0.66), followed by CMAR (-0.03), then DB (0.25) and finally the IPD with the highest average PQR score (0.53). Furthermore, the study was able to validate the results using factor analysis and multidimensional scaling techniques, which generated very similar results.

Chen and Osei-Kyei (2017) studied public-private partnership (PPP) projects in Ghana. This study attempted to identify success criteria for these projects using statistical tools like mean-score ranking (MS), factor analysis (FA) and fuzzy synthetic evaluation (FSE). 120 PPP experts provided data, which highlighted 9 success criteria that were grouped into three success areas, including local development and disputes resolution, cost, and technical specifications. The study further highlighted that meeting cost project budget, and specification requirements was the most important factor with a weight of 35.2%, followed by profitability and meeting project schedule (34.7%), and then by ability to provide local development and minimize disputes (30.1%). Thus, the following equation was generated:

$$\begin{aligned} \text{Success Index} &= 0.301 * \text{Local development and dispute reduction} + 0.347 * \text{Profit} \\ &\quad + 0.352 * \text{Cost and technical specifications} \end{aligned}$$

2.3 Impact of Change Orders on Project Performance

Zink (1986) conducted a study using the measured mile approach to quantify lost productivity. He compared similar activities in areas that are both impacted and unimpacted by a particular change order to quantify the lost productivity.

Leonard (1988) used graphical representation to demonstrate the relationship between the percentage of project changes and the associated productivity loss. He found that the cumulative impact of changes caused out-of-sequence work, low labor morale, and lack of engineering support, which resulted in loss of productivity.

Hanna et al. (1999) studied the impact of change orders for electrical and mechanical construction and labor efficiency. They found that the impacts include a break in planned workflow, resulting in increased costs due to rework and decreased efficiency to base work.

Hanna (2001), under the CII research 158, studied the cumulative impact of change orders. He found that 75% of change orders are due to additions, design changes, and design errors for mechanical contractors, 80% for electrical contractors.

Hanna et al. (2004) reported on the cumulative impact of change orders on labor productivity on mechanical and electrical contractors. The study highlighted that although the changes are typically compensated in terms of direct cost and time extension, the change causes disruption to site operations which can have indirect cost implications, such as extended unabsorbed home office overhead, and disruption to the flow of work that can have a net effect on loss of productivity (called by the author %delta).

As such, this study quantified the cumulative impact of change orders in terms of six factors: percentage change, change order processing time, overmanning, percentage of project manager's time on the project, percentage of changes initiated by the owner and whether the contractor tracks productivity. The study used 69 data points, 33 from electrical contractors and 36 from mechanical contractors, resulting in the following quantification of the %delta:

$$\begin{aligned} \%detla = 0.37 + 0.12 * Percent\ change - 0.08 * Project\ manager\ \%\ time\ on\ project \\ - 0.17 * Owner\ initiated\ CO - 0.09 * Productivity - 0.05 * Overmanning \\ + 0.02 * Processing\ time \end{aligned}$$

An example presented in this study cites the cumulative impact of change orders in the following manner: a 7% increase in scope can result in a 25% loss of productivity. The study concluded by mentioning that while changes are inevitable, changes due to additions, errors, or changes in design can and should in theory be eliminated prior to construction.

Riley et al. (2005a) specifically demonstrate the damaging impacts of design coordination change orders, including loss of productivity, schedule compression, overtime, and trade stacking. The damaging effects of those change orders were also found to increase the later they occurred on a project.

Riley et al. (2005b) compared the occurrence and magnitude of change orders in DB and DBB projects, examining nearly 600 change orders over 120 projects executed by a single mechanical contractor. The study highlighted negative implications of field-generated changes (defined as changes due to design errors or lack of coordination) including loss of productivity, schedule compression, overtime and trade stacking. While other change orders related to scope are inevitable, the study claimed that field-generated changes are the most disruptive and are typically avoidable.

It was further found that DB projects experienced 87% fewer field-generated changes than DBB and 86% lower dollar value of changes than DBB. Only 5% of the change orders on DB projects were initiated from unforeseen conditions, compared to 38% on DBB projects.

It was also found that DB projects experienced an average cost growth of 4.7% compared to 16.6% for DBB projects. Only 0.1% of that 4.7% was due to unforeseen conditions, while 5.6% of changes on DBB projects were due to unforeseen conditions, concluding the better performance of DB over DBB in terms of cost change.

2.4 Section Summary

There is an underpinning urge to understand and optimize project performance, and as such project delivery methods are frequently studied, compared, and analyzed. While many studies have focused on cost and schedule growth, as these are the most visible metrics of performance, some more detailed studies further quantified metrics for quality, claims, disputes, team integration and cohesion, and others. The body of literature has not come to an accord on the relative performance of the various project delivery methods, however, there a general consensus was found about the higher performance of CMAR and DB compared to DBB in terms of schedule, and a less significant consensus in terms of cost. Many studies are in favor of more Collaborative methods, wherein project parties are incentivized to coordinate and innovate, and the traditionally adversarial relationship is minimized.

Of particular note in the literature, substantial attention is paid to change orders, claims that stem from them, and the management of change. Studies report the most common sources of change under various delivery methods and for different trades, while others attempt to quantify the impact of different types of change to highlight the adverse latent ramifications of changes on project performance.

Table 2: Summary of Literature Review on Comparative Performance Studies

Parameter	Konchar 1998	Ibbs 2003	Col Debella 2006	Rojas 2008	Rojas 2008	Hale 2009	Pishdad-Bozorgi 2012	Carpenter 2016	El Asmar 2016	Franz 2017	Sullivan 2017	Shrestha 2017	Ibrahim 2018	This Study
Delivery Methods	DB CMAR DBB	DB DBB	SP MP MP-A	CMAR DBB	SP MP	DB DBB	DB DBB	CMAR DBB	IPD DB CMAR DBB	IPD DB CMAR DBB	DB CMAR DBB	DB DBB	IPD DB CMAR DBB	DB CMAR SP MP
Sample Size	351	67	94	297	1000	77	NA	137	35	204	4600	77	109	189
Owner Type	Public & private	—	Public schools	Public schools	Public projects	Federal barracks	NA	Public schools	Mostly private	Public & private	Public & private	Public university buildings	Public & private	Public projects
Cost	*	*	*	*	*	*	—	*	*	*	*	*	*	*
Schedule	*	*	*	—	—	*	—	*	*	*	*	*	*	*
Quality	*	—	*	—	—	—	—	*	*	*	*	*	*	—
Spending	*	*	*	—	—	—	—	*	*	*	*	—	*	*
Productivity	—	*	—	—	—	—	—	—	—	—	—	—	—	—
Litigation	—	—	*	—	—	—	*	*	—	—	—	—	—	—
Communication & Collaboration	—	—	*	—	—	—	—	—	*	*	—	—	*	*
Change Management	—	—	—	—	—	—	—	—	—	—	—	*	*	*
Safety	—	—	—	—	—	—	—	—	*	—	—	—	*	—
Other	—	—	—	—	—	—	—	—	*	—	—	—	—	—

2.5 Literature Gap

The extant body of literature contains numerous studies concerning project delivery methods, with various comparisons and analytical techniques performed. However, the pool of studies which focus on public projects is smaller, and that of comprehensive studies which focus on relative performance of Collaborative and Traditional delivery methods across several facility types (i.e. academic, research, detention, etc.), using a significant dataset are scarcer. No previous study of public projects has attempted to develop a multivariate unified index of this kind that can successfully capture the numerous variables (performance metrics) which affect project success and provide statistically validated results.

Further still, no study has focused solely on the Midwest as a region, or Wisconsin as a state specifically. This region and this state contain major academic institutions and high-volume infrastructure projects, yet the dearth of studies remains.

While it is true that several existing studies have examined change management in public projects, few have gone beyond evaluating change post-hoc (i.e. the overall percentage change) to analyze the likelihood of change and the reasons behind it in a comparative manner across various delivery methods.

This particular study fills the gaps herein identified by conducting a holistic quantitative assessment of the performance of Wisconsin State-funded projects. Comparative analysis of delivery methods, data analysis of multiple performance metrics, and a robust mathematical model will be used to recommend improvements to decision makers. Further two statistical techniques will be used to validate the results generated by the mathematical model.

3 Chapter 3: Mathematical Development of Project Performance Index (PPI)

This chapter will provide a thorough explanation and elaboration on the derivation process of the Project Performance Index (PPI). This model creates a comprehensive performance score using inputs from eight performance metrics over four holistic areas, as shown in Figure 15 below, serving as a dimensionality reduction technique.

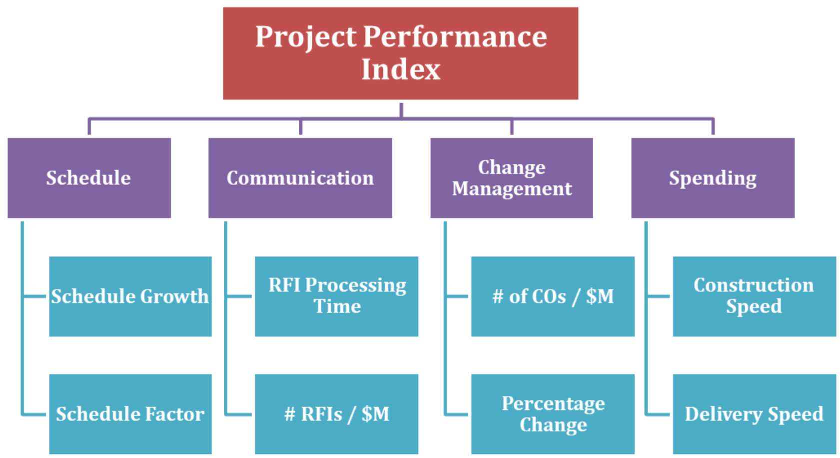


Figure 15: Project Performance Index Components

To answer the research question, two models were developed. The first uses a sum of squared correlations approach to compute weights for the performance index, such that the weights generated by this model return the maximum possible sum of squared correlations between the vector of the discrete performance metrics and that of the PPI. The second model uses an iterative approach, which assumes that the initial weights for all metrics are equal, from which an initial PPI score is computed. Multiple iterations of this model were then conducted until the computed weights converge.

While it is true that the two models have differing rationale, one utilizing correlation and the other iteration, further derivation and analytic applications on the data show that both models return the same set of weights exactly. This provides a strong case for the model's accuracy and applicability.

3.1 Correlation[2] Maximization Model

In this section we introduce a mathematical derivation of a Project Performance measure (denoted by 'g' for a given project that has the simple form:

$$g(z_1, z_2, \dots, z_n) = w_1 z_1 + w_2 z_2 + \cdots + w_n z_n, \tag{1}$$

which is a linear weighted sum of n project factors $z_1, z_2, \dots, z_n$, with weights $w_1, w_2, \dots, w_n$ satisfying the equality:

$$w_1 + w_2 + \cdots + w_n = 1. \tag{2}$$

These weights need to be extracted from raw project data comprising the factors of m projects. We let x_{ij} denote the raw value of factor j for project i, where $1 \le i \le m, 1 \le j \le n$.

For the weight w_j to have a sensible meaning indicating the importance of factor j, we need to work with standardized factors z_{ij} with mean 0 and standard deviation 1. These are calculated from the raw data as follows:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\pm s_j}, \tag{3}$$

where:

$$\bar{x}_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij} \quad and \quad s_j^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(x_{ij} - \bar{x}_j\right)^2, \tag{4}$$

and the $\pm$ sign is assigned as $+$ if factor j is a goodness factor, and – if it is a badness factor.

Thus, for given weights w_j, the Performance of project i can be described by:

$$g_i = w_1 z_{i1} + w_2 z_{i2} + \cdots + w_n z_{in}.$$

We start by using the convenience of matrix notation, and let the standardized data be organized in the matrix:

$$Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ z_{m1} & z_{m2} & \cdots & z_{mn} \end{bmatrix} = [Z_1 \quad Z_2 \quad \cdots \quad Z_n], \tag{5}$$

where Z_j denotes the j^{th} column of the matrix Z, which carries the standardized values of the j^{th} factor for all m projects. Also, the Performance measures of the projects can be organized in the column vector:

$$G = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_m \end{bmatrix} = \begin{bmatrix} w_1 z_{11} + w_2 z_{12} + \cdots + w_n z_{1n} \\ w_1 z_{21} + w_2 z_{22} + \cdots + w_n z_{2n} \\ \vdots \\ w_1 z_{m1} + w_2 z_{m2} + \cdots + w_n z_{mn} \end{bmatrix} = w_1 Z_1 + w_2 Z_2 + \cdots + w_n Z_n,$$

or simply: $G = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ z_{m1} & z_{m2} & \cdots & z_{mn} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix} = ZW$, which depends on the unknown weight vector $W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix}$.

We note that the given data does not include an objective factor that tells us (regardless of the other factors) how good or bad a certain project is. To solve this problem, we require that the performance vector $G = ZW$ is highly correlated with each of the factors Z_j. This can be made precise by stating the following:

Blanket Assumption: The weight vector W should be chosen to maximize the length of the vector whose components are the correlations between the performance vector $G = ZW$ and each of the factors Z_j. Equivalently W needs to maximize the sum of the squares of the correlations:

$$\sum_{j=1}^{n} \left[corr\left(G, Z_j\right)\right]^2,$$

subject to Condition (2), which can be phrased as:

$$W^T J = 1, \qquad where \quad J = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}. \tag{6}$$

To this end, we start by establishing the following means:

$$\bar{z}_j = \frac{1}{m}\sum_{i=1}^{m} z_{ij} = \frac{1}{m}\sum_{i=1}^{m} \frac{x_{ij} - \bar{x}_j}{\pm s_j} = \frac{1}{\pm m s_j}\left[\left(\sum_{i=1}^{m} x_{ij}\right) - m\bar{x}_j\right] = \frac{1}{\pm m s_j}\left(m\bar{x}_j - m\bar{x}_j\right) = 0,$$

$$\bar{g} = \frac{1}{m}\sum_{i=1}^{m} g_i = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n} w_j z_{ij} = \frac{1}{m}\sum_{j=1}^{n} w_j \left(\frac{1}{m}\sum_{i=1}^{m} z_{ij}\right) = \frac{1}{m}\sum_{j=1}^{n} w_j \bar{z}_j = 0,$$

as well as the variances:

$$s_{Z_j}^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(z_{ij} - \bar{z}_j\right)^2 = \frac{1}{m-1}\sum_{i=1}^{m} z_{ij}^2 = \frac{1}{m-1}\sum_{i=1}^{m} \frac{\left(x_{ij} - \bar{x}_j\right)^2}{s_j^2} = \frac{1}{s_j^2} s_j^2 = 1,$$

$$s_G^2 = \frac{1}{m-1}\sum_{i=1}^{m}(g_i - \bar{g})^2 = \frac{1}{m-1}\sum_{i=1}^{m} g_i^2 = \frac{1}{m-1} G^T G = \frac{1}{m-1}(ZW)^T ZW.$$

We can now evaluate the j^{th} correlation as follows:

$$corr(G, Z_j) = \frac{1}{(m-1)s_G s_{Z_j}} \sum_{i=1}^{m} (g_i - \bar{g})(z_{ij} - \bar{z}_j) = \frac{1}{\sqrt{(m-1)(ZW)^T ZW}} \sum_{i=1}^{m} g_i z_{ij}$$
$$= \frac{1}{\sqrt{(m-1)(ZW)^T ZW}} G^T Z_j = \frac{1}{\sqrt{(m-1)(ZW)^T ZW}} (ZW)^T Z_j$$

We now seek to maximize the sum of squares:

$$\sum_{j=1}^{n} \left[corr(G, Z_j)\right]^2 = \sum_{j=1}^{n} \frac{1}{(m-1)(ZW)^T ZW} \left[(ZW)^T Z_j\right]^2$$
$$= \frac{1}{(m-1)(ZW)^T ZW} \sum_{j=1}^{n} (ZW)^T Z_j \left[(ZW)^T Z_j\right]^T,$$

where we applied the transpose on the 1×1 matrix $(ZW)^T Z_j$.

We can thus seek to maximize the quantity:

$$F(W) = (m-1) \sum_{j=1}^{n} \left[corr(G, Z_j)\right]^2 = \frac{1}{(ZW)^T ZW} (ZW)^T \left(\sum_{j=1}^{n} Z_j Z_j^T \right) ZW.$$

Writing $Z_j = ZJ_j$, where: $J_j = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$ (with 1 appearing only in the j^{th} entry), we can simplify the last sum as follows:

$$\sum_{j=1}^{n} Z_j Z_j^T = \sum_{j=1}^{n} ZJ_j (ZJ_j)^T = \sum_{j=1}^{n} ZJ_j J_j^T Z^T = Z \left(\sum_{j=1}^{n} J_j J_j^T \right) Z^T = ZIZ^T = ZZ^T.$$

Thus, we get the simpler expression:

$$F(W) = \frac{1}{(ZW)^T ZW} (ZW)^T ZZ^T ZW = \frac{W^T Z^T ZZ^T ZW}{W^T Z^T ZW} = \frac{W^T CCW}{W^T CW}, \qquad (7)$$

Where, $C = Z^T Z$ is the so-called *correlation* matrix of the data. We note that C is symmetric, since $C^T = (Z^T Z)^T = Z^T Z = C$, and positive semidefinite, since for all $n \times 1$ vectors X, $X^T C^T X = X^T Z^T ZX = (ZX)^T (ZX) \geq 0$. Moreover, $C = Z^T Z$ will be positive definite, provided that $Rank(Z) = n$, which is a reasonable assumption, since the columns Z_j can be assumed to be independent of each other.

Since C is both symmetric and positive definite, we can find its unique Cholesky decomposition $C = U^T U$, where U is an invertible upper triangular matrix. Note that the matrices C, U are both $n \times n$ square matrices, while Z is (generally) a non-square $m \times n$ matrix. Using Cholesky decomposition we can write:

$$F(W) = \frac{W^T U^T U U^T U U W}{W^T U^T U W} = \frac{(UW)^T (UU^T)(UW)}{(UW)^T (UW)} = \frac{V^T A V}{V^T V}, \quad (8)$$

which is a Rayleigh Quotient (Horn and Johnson 2005) that is maximized when $V = UW$ is any eigenvector of the matrix $A = UU^T$ corresponding to its maximum eigenvalue. In other words, we have:

$$AV = \lambda_{max} V, \text{ i.e. } UU^T UW = \lambda_{max} UW, \text{ which is equivalent to: } U^T UW = \lambda_{\max} W$$

i.e. W must be an eigenvector of the correlation matrix $C = U^T U$ corresponding to its maximum eigenvalue λ_{max} shared[2] by the matrix A.

Finally, note that we need to choose the eigenvector W to satisfy Condition (6). Thus, given any eigenvector W' of C, we can choose another eigenvector $W = kW'$ such that:

$$1 = W^T J = (kW')^T J = k(W')^T J.$$

Thus, $k = \frac{1}{(W')^T J}$, and $W = kW' = \frac{W'}{(W')^T J}$. In other words, to satisfy Condition (6), we just need to scale the vector W' by the reciprocal of the sum of its components.

We now describe the following procedure for deriving the weight vector W from the data $[x_{ij}]$:

1. Use Equations (3-5) to calculate the $m \times n$ matrix Z of standardized values.
2. Calculate the (square) $n \times n$ correlation matrix $C = Z^T Z$.
3. Find any eigenvector W' of the matrix C corresponding to its maximum eigenvalue.
4. Calculate the vector $W = \frac{W'}{(W')^T J}$, where J is given by (6).

3.1.1 Remark about the Blanket Assumption

We now ask the question: What if we replaced the blanket assumption by an alternative assumption that maximizes the sum of the correlations:

$$\sum_{j=1}^{n} corr(G, Z_j) = \frac{1}{\sqrt{(m-1)(ZW)^T ZW}} (ZW)^T \sum_{j=1}^{n} Z_j = \frac{1}{\sqrt{(m-1)(ZW)^T ZW}} (ZW)^T ZJ.$$

[2] The previous observation shows that the two matrices $A = UU^T$ and $C = U^T U$ have exactly the same eigenvalues.

In this case, we can just maximize the quantity:

$$G(W) = (m-1)\left[\sum_{j=1}^{n} corr(G, Z_j)\right]^2 = \frac{1}{(ZW)^T ZW}[(ZW)^T ZJ]^2 = \frac{(ZW)^T ZJ[(ZW)^T ZJ]^T}{(ZW)^T ZW},$$

where (as before) we applied the transpose on the 1×1 matrix $(ZW)^T ZJ$.

Using the Cholesky decomposition of the matrix $C = Z^T Z = U^T U$, we write:

$$G(W) = \frac{W^T Z^T Z J J^T Z^T Z W}{W^T Z^T Z W} = \frac{W^T U^T U J J^T U^T U W}{W^T U^T U W} = \frac{(UW)^T U J J^T U^T (UW)}{(UW)^T (UW)} = \frac{V^T A' V}{V^T V},$$

which is very similar to Equation (8), with $A = UU^T$ replaced by $A' = UJJ^T U^T = UJ(UJ)^T$. However, unlike the positive definite matrix A, the matrix A' is singular. In fact, since the matrices U, U^T are invertible, we have:

$$Rank(A') = Rank(UJJ^T U^T) = Rank(JJ^T) = 1.$$

Therefore, A' has a zero eigenvalue of multiplicity $n-1$. Also, we can check that:

$$A'(UJ) = UJ(UJ)^T(UJ) = \lambda_{\max(UJ)}.$$

Thus, $V = UJ$ is an eigenvector of A' corresponding to its maximum eigenvalue:$\lambda_{max} = (UJ)^T(UJ) = J^T U^T U J = J^T C J = \sum_{i,j} c_{ij}$

We then get a rather "useless" weight vector $W = \frac{J}{J^T J} = \begin{bmatrix} \frac{1}{n} \\ \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{bmatrix}$, with equal weights.

3.1.2 Additional Remarks

- The developed model is very similar in derivation to the Principal Component Analysis (PCA) with subtle differences such as the utilization of the Choelsky decomposition and Rayleigh quotient.
- It should be noted that the motivation behind the model is to solve for the optimum set of weights that could generate a single index, characterized by having the maximum sum of squared correlations with each of the individual metrics, unlike PCA which aims to generating principal components that could maximize the captured variances from the data.
- Although some of the model assumptions were not met, such as the data normality and linearity among all the pairs of metrics, the results were almost identical when compared to Factor Analysis with oblique rotation that did not require those assumptions as will be explained in Chapter 5. This elaborates the negligible effect of not satisfying these assumptions on the model outputs.

3.2 Iterative Approach Model

In this section we introduce an iterative model that produces successive approximations of the weights, which converge (within a given margin of error) after a sufficient number of iterations.

We will denote the k^{th} iterate of the vector of weights by W^k and the j^{th} weight in that vector as w_j^k, for $j \in \{1, 2, \ldots . n\}$. Initially (for $k = 0$), the values for these weights are uniformly set to be: $w_j^0 = \frac{1}{n}$ for all j. When the vector of weights W^k is determined, we can calculate the k^{th} iterate of the performance vector G^k with components g_i^k, for $i \in \{1, 2, \ldots . m\}$ by the equation:

$$g_i^k = \sum_{j=1}^{n} w_j^k z_{ij}$$

For the $(k+1)^{th}$ iterate, we attain the performance vector G^k as the weighted sum of the standardized data for every project, i.e. $g_i^k = \sum_{j=1}^{n} w_j^k \cdot z_{ij}$. Next, we let the weights w_j^{k+1} be proportional to the respective correlations between the columns (factors) of the standardized matrices Z_j and G^k. The proportionality factor is taken to guarantee that the sum of the weights equals to 1, i.e. $\sum_{j=1}^{n} w_j^{k+1} = 1$. Thus, we get:

$$w_j^{k+1} = \frac{1}{\sum_{j=1}^{n} cor\left(Z_j, G^k\right)} cor\left(Z_j, G^k\right)$$

This iteration model is run until the sum of the absolute values of the differences between each weight w_j^{k+1} and the weight corresponding to the same factor in the previous iteration w_j^k is smaller than some margin of error around the limit points of these weights (Figure 16).

It is important to note that this result coincides with the result reached in the previous correlation[2] maximization model (using the same data set). This congruence is proved in the following procedure.

We know the vector of performance indices $G = ZW$, the matrix multiplication of the standardized data Z and the attained weight vector W. It is also known that the correlation of the j^{th} column factor Z_j and G is $corr\left(Z_j, G\right) = \frac{1}{m-1} \cdot \frac{1}{\sigma_{Z_j} \sigma_G} \Sigma_i \, z_{ij} \cdot g_i = \alpha \cdot Z_j^T G$, where

$\alpha = \frac{1}{(m-1)\sigma_{Z_j}\sigma_G} = \frac{1}{(m-1)\sigma_G}$, due to the fact that Z is standardized.

Define the vector of these correlations between Z_j and G for all $j \in \{1, 2, \ldots n\}$ as

$$D = \begin{bmatrix} corr(Z_1, G) \\ corr(Z_2, G) \\ \vdots \\ corr(Z_n, G) \end{bmatrix} = \alpha \cdot \begin{bmatrix} Z_1^T G \\ Z_2^T G \\ \vdots \\ Z_n^T G \end{bmatrix} = \alpha \cdot \begin{bmatrix} Z_1^T \\ Z_2^T \\ \vdots \\ Z_n^T \end{bmatrix} G = \alpha \cdot Z^T Z W = \alpha \cdot CW.$$

Now, the mathematical model produces a vector of weights W, such that $CW = \lambda_{max} \cdot W$, where W is the eigenvector of C corresponding to the maximum eigenvalue.

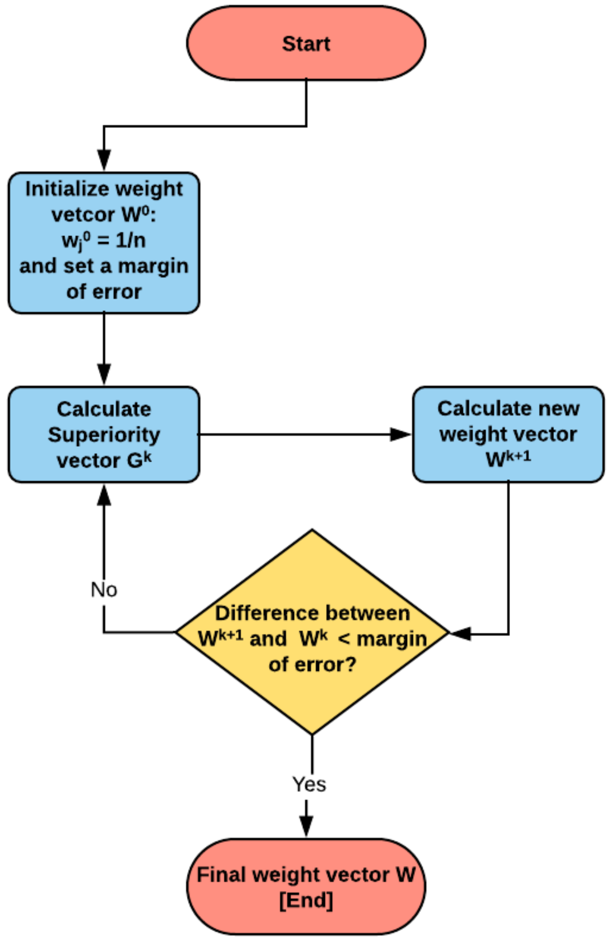


Figure 16: Flow Chart of The Iterative Model

3.3 Outputs of The Mathematical Models

The outputs of the mathematical models are optimized weights which express the relative importance of the contribution of each individual metric to the overall performance of the project, as well as the overall unified performance score which captures the eight performance metrics.

Figure 17 below presents the resultant weights. Construction speed was the highest weighted metric, at 15.72%, while RFI processing time was the lowest at 8.65%. While a higher value does simply mean a higher degree of importance, that importance should be understood within the context of correlation strength between each metric as a vector and the performance score vector.

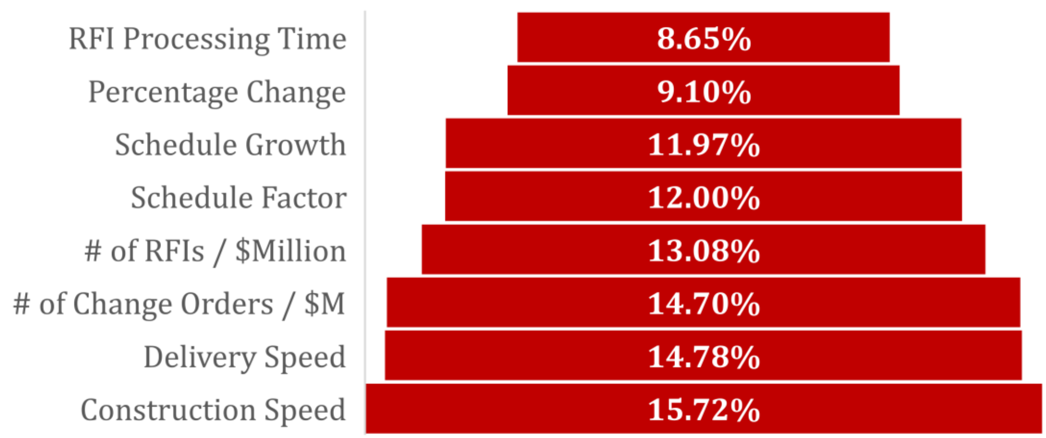


Figure 17: Weights Breakdown of PPI Metrics

Using these computed weights, the PPI score for any project can be computed using the equation below, given that all metrics are standardized. The negative sign on each coefficient represents the direction of the vector.

$$\begin{aligned} PPI\ Score = & -0.865 * RFI\ Processing\ Time - 0.91 * Percentage\ Change \\ & - 0.1197 * Schedule\ Growth - 0.12 * Schedule\ Factor \\ & - 0.1308 * \#\ of\ RFIs\ /\ \$Million - 0.147 * \#\ of\ Change\ Orders\ /\ \$Million \\ & + 0.1478 * Delivery\ Speed + 0.1572 * Construction\ Speed \end{aligned}$$

However, as there is a level of difficulty associated with standardizing each metric (as doing so requires that the whole dataset be available) a simplified equation was developed to accommodate unstandardized data. This equation, below, eliminates errors that stem from standardization, while also provides a user-friendly methodology to compute performance.

$$PPI\ Score = \sum_{i=1}^{8} \alpha_i * M_i + \beta_i$$

where M is a given performance metric.

To get values for alpha and beta, the standardization process was reversed, and the values were obtained, as shown in the table below.

Table 3: Alpha & Beta Coefficients for Each Metric

Metric	Alpha	Beta	Weights
RFI Processing Time	-1.101	0.670	-8.65%
Percentage Change	-0.219	0.670	-9.10%
Schedule Growth	-0.241	0.670	-11.97%
Schedule Factor	-0.022	0.670	-12.00%
Number of RFIs / $Million	-0.014	0.670	-13.08%
Number of Change Orders / $M	-0.019	0.670	-14.70%
Delivery Speed	0.005	0.670	14.78%
RFI Processing Time	0.011	0.670	15.72%
Total		**5.360**	

Recalling the non-simplified equation previously presented, the equation below adapts this method to calculate PPI scores for unstandardized data.

By substituting those values into the above form, the following equation could be obtained for calculating PPI scores using unstandardized data:

$$\begin{aligned} PPI\ Score = & -1.101 * RFI\ Processing\ Time - 0.219 * Percentage\ Change \\ & - 0.241 * Schedule\ Growth - 0.022 * Schedule\ Factor \\ & - 0.014 * \#\ of\ RFIs\ /\ \$Million - 0.019 * \#\ of\ Change\ Orders\ /\ \$Million \\ & + 0.005 * Delivery\ Speed + 0.011 * Construction\ Speed + 5.360 \end{aligned}$$

Finally, the following equation is used to scale the generated PPI score to the 1 to 10 range, so that it can be compared to the benchmarks and analysis results:

$$\begin{aligned} PPI\ Scaled\ Score &= \frac{[PPI\ Score - Min] * (MaxR - MinR)}{Max - Min} + MinR \\ &= \frac{[PPI\ Score + 2.45] * 9}{2.63 + 2.45} + 1 = \frac{9 * PPI\ Score + 22.02}{5.08} + 1 \\ &= 1.77 * PPI\ Score + 5.33 \end{aligned}$$

Where:

- Min is the minimum PPI score in the dataset
- Max is the maximum PPI score in the dataset
- MinR is the lower boundary of the desired range (1 in this case)
- MaxR is the upper boundary of the desired range (10 in this case)

4 Chapter 4: Data Collection & Preparation

4.1 Challenges

4.1.1 Data Collection Phase

The initial challenge of data collection was that not all performance metrics in common use by academia are tracked in a useable manner. Only limited studies and publications were found regarding the performance of Wisconsin capital projects, like the one focusing on residence halls published by the UW-System. Furthermore, there is very little publicly available information about the storage location and access requirements for this data, and the nature of the data stored.

The researcher, via direct communication with WisDOA, was able to identify some particulars of what information is tracked by the State. However, far more fruitful were conversations with the DOA personnel and former project managers of State projects. From these conversations, key performance metrics which are tracked by the state were identified. The next step was to gain access to sufficiently representative data to fulfil the research objective. The required data was collected through records requests and meetings with personnel from both the WisDOA and the University of Wisconsin department of Facilities.

4.1.2 Data Preparation Phase

Any data-driven research effort requires preparation and treatment to ensure the collected data is reliable and complete. Figure 18, below, outlines the data preparation procedure used in this research.

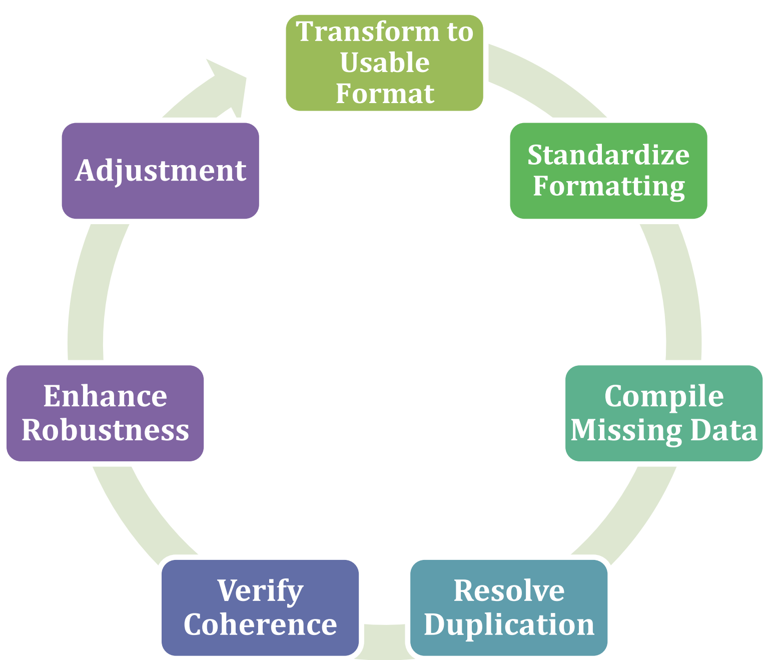


Figure 18: Data Preparation Challenges

4.1.3 Transform Data into a Usable Format

A significant portion of the data was delivered in a format that was not editable, requiring a significant data entry effort to move this data into operable form, i.e. spreadsheets. Over 400 sheets of information were transformed in this way.

4.1.4 Standardize Formatting

After ensuring all data was operable, formatting issues were identified and eliminated. Some of those were simple to adjust, such as an abbreviation of "Design-Build" as "Design Build", others were more complex in which the researcher connected with the stakeholders to provide more clarity.

4.1.5 Compile Missing Data

Every effort was made to ensure each project characteristic field and performance metric field was accurately populated. Depending on the nature of missing information, different approaches were used, as detailed below. This methodology consists of the following ordered steps:

1. Wisconsin State published documents, including Capital Budgets, State Building Program Informational Paper, State Building Commission meeting notes
2. UW-System published documents, including Capital Plans, Residence Hall Facility Study, and other studies
3. UW-System campuses Capital Planning & Development websites, publications and news
4. Online resources including project stakeholders' websites, as well as public news
5. Direct contact with project stakeholders (i.e. Architects, Contractors, etc.)
6. Individual meetings with project stakeholders' representatives for additional feedback and data completeness

4.1.6 Resolve Duplicate Conflicting Sources

In the process of acquiring complete data, it was occasionally found that the same data point was reported from multiple sources with sometimes conflicting values. Using the same methodology laid out in 4.1.5, these conflicting data points were clarified, and the accurate figures acquired. An example of this type includes dates where the months and days were flipped, an inconsistency generated from date formatting (i.e. 6th of September vs 9th of June).

4.1.7 Verify Coherence of Information

This step was a second-tier verification to ensure values were consistent with other project characteristics of each particular project, as well as those of similar projects. Whenever an input value was found to be out of alignment (did not follow the chronological sequence for instance), that value was omitted and subject to a second round of data compiling, duplicate resolution and coherence verification.

4.1.8 Enhance Robustness of the Dataset

After these steps were performed, it could be held with a high level of confidence that the data was of high quality. However, the researcher then sought to expand the dataset such that it could be more representative. The 'cleaned' dataset was compared to the research statement and objectives to identify weaknesses and areas of improvements.

One such weakness was the underrepresentation of DB and CMAR projects. This is altogether unsurprising, as these methods are rarely allowed in the state of Wisconsin. Thus, it was challenging to expand the sample size of projects within the other bounds of data to capture additional DB or CMAR projects. As such, the research scope was expanded to include projects completed as far back as 2000. All projects under the four delivery methods (DB, CMAR, SP and MP) completed during this period were included to maintain a dataset that is representative to the population, and hence, preclude bias.

4.1.9 Data Adjustment for Time, Location and Size

As the expanded scope of this research includes projects completed over an eighteen-year period (2000-2017), cost metrics required adjustment to account for inflation. As such, the RS Means 2018 – Historical Cost Index was used. As shown below, the benchmark year index (2018 for this research) is divided by the index from the year the subject project was completed (2005 in the example below). The adjusted cost is then equal to the time multiplier multiplied by the original project's cost in its year of completion.

$$Time\ Multiplier = \frac{Adjustment\ Index\ for\ 2018}{Adjustment\ Index\ for\ 2005}$$

$$Project\ Cost\ in\ 2018 = \ Time\ Multiplier * Project\ Cost\ in\ 2005$$

The dataset also encompassed the entire state of Wisconsin, which includes numerous different markets for labor and materials. As such, the prices of these project factors were standardized. While variance in these figures within the same state is not significant, standardization was performed to provide the most possible accurate results. RS Means 2018 – Location Factors was used, with Madison as the benchmark city. The below formula provides a location adjustment example of a project that was executed in Milwaukee:

$$Location\ Multiplier = \frac{Madison\ Adjustment\ Index}{Milwaukee\ Adjustment\ Index}$$

$$Project\ Cost\ in\ Madison = Location\ Multiplier * Project\ Cost\ in\ Milwaukee$$

Finally, when considering project size, it is crucial to understand that the unit cost of a project is not linearly associated with its size, such that as project size increases, unit cost nonlinearly decreases (Figure 19). Therefore, projects of different sizes were adjusted to a typical project size per RS Means 2018 – Square Foot Project Size Modifier. By dividing a project's area in gross square feet (GSF) by the typical area for that specific facility type, a Size Factor is computed. Then, the factor is used to obtain a Cost Multiplier using the following chart, developed by RS Means.

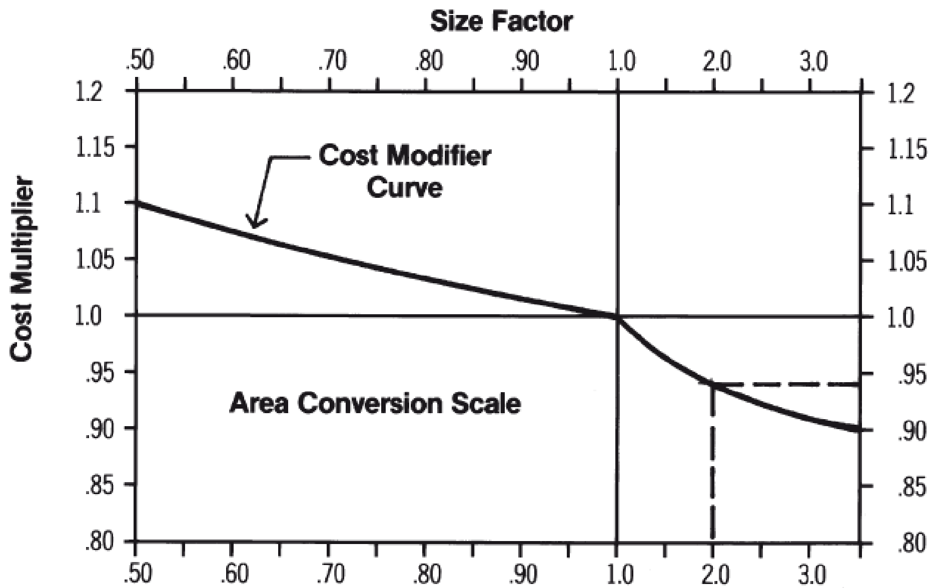


Figure 19: RS Means 2018 Size Adjustment Chart

The below formula provides an example for size adjustment of a 100,900 GSF research facility project (typical size of this facility type is 39,800 GSF).

$$Size\ Factor\ = \frac{100{,}900\ GSF}{39{,}800\ GSF} = 2.53$$

Referring to the previous Figure 19, a Size Multiplier of ~0.92 is derived and used to adjust the project cost according to the below formula.

$$Size\ Adjusted\ Project\ Cost = Original\ Project\ Cost * Size\ Multiplier$$

In an effort to eliminate any human error and subjectivity in the process of deriving the Cost Multiplier off the chart, a set of equations were developed which numerically depict the chart and provide more accurate Size Multipliers, as shown below:

$$f(x) = \begin{cases} 1.1 & x < 0.5 \\ 0.2x + 1.2 & 0.5 \leq x < 1.0 \\ -0.0008x^3 + 0.0174x^2 - 0.1052x + 1.0892 & 1.0 \leq x < 3.5 \\ 0.9 & x \geq 3.5 \end{cases}$$

where $f(x)$ is the Size Multiplier and x is the Size Factor

All of these steps were applied to every project in the dataset to standardize project costs for time, location, and size. However, it should be noted here that these adjustments are inapplicable to unitless metrics (e.g. Schedule Growth). The following collective equation was developed for a streamlined cost adjustment.

$$\begin{aligned} &Adjusted\ Project\ Cost \\ &\quad = Unadjsuted\ Project\ Cost * Time\ Multiplier * Location\ Multiplier \\ &\quad * Size\ Multiplier \end{aligned}$$

4.2 Characteristics & Overview

This section discusses the characteristics of the data set. Understanding these characteristics is essential, ensuring that the sampled projects are representative of the overall population of projects. Figure 20 below presents the magnitude of the data as a whole. 189 projects were collected, most of which had only 1 design contract (average = 1.06). On average, there were 170 change orders and 250 RFIs per project.

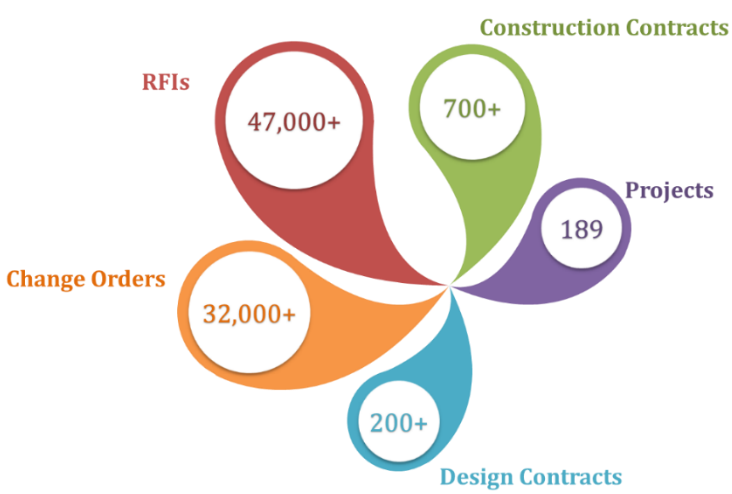


Figure 20: Order of Magnitude of the Collected Data

Figure 21 below details the data ranges. As can be seen, the delivery cost ranged from $4 to $198 million, and sums to $5.2 billion in 2018 dollars. This sum also represents over 50% of the apportioned funds for construction projects over the study period, reinforcing its representation. The projects ranged in duration from 2 to 13 years, with a total duration of 980 years of design and construction.

As has been previously discussed in this thesis, SP and MP are dominant delivery methods for State projects, but this research also encapsulates DB and CMAR projects. Finally, the collected projects were composed of more than 10 facility types, which provides deep insights later in the analysis as to which facility types perform better than the others in comparison to the national average.

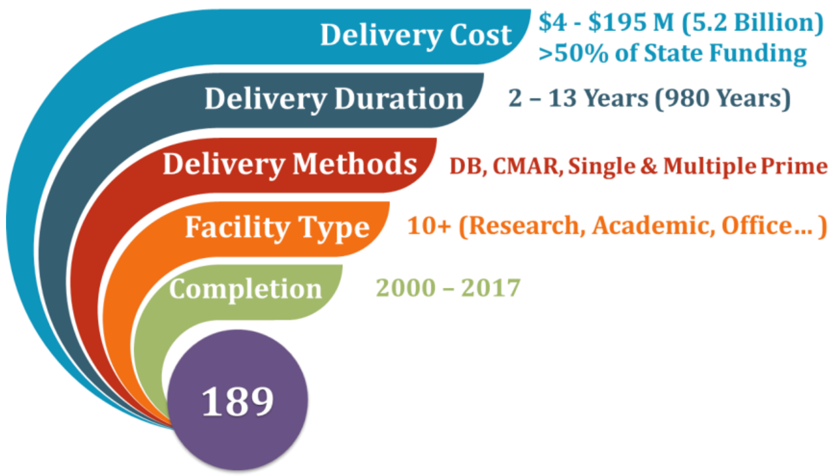


Figure 21: Key Characteristics of Collected Projects

When projects are broken down by state agency affiliation, the majority were UW-System, while others were associated with the departments of Military, Corrections, or Administrative. Still other agencies had one or two projects in the data – these are grouped together as 'Others.'

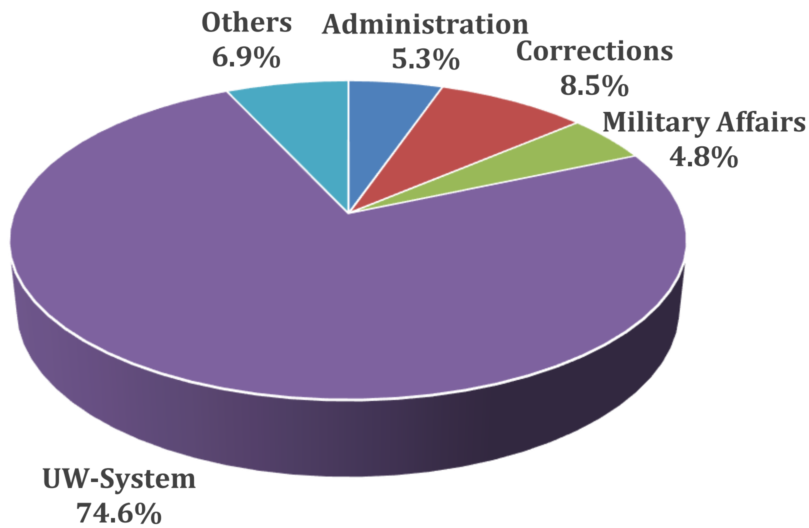


Figure 22: Distribution of Projects by Agency

A second check of representation was performed by comparing the distribution chart pictured above (of the data set) to the one below (Figure 23), which depicts the allocation of State projects from 2001 to 2017, a period which closely approximated the period of the study. It can be clearly seen that these charts are very similar, providing strong evidence of representation.

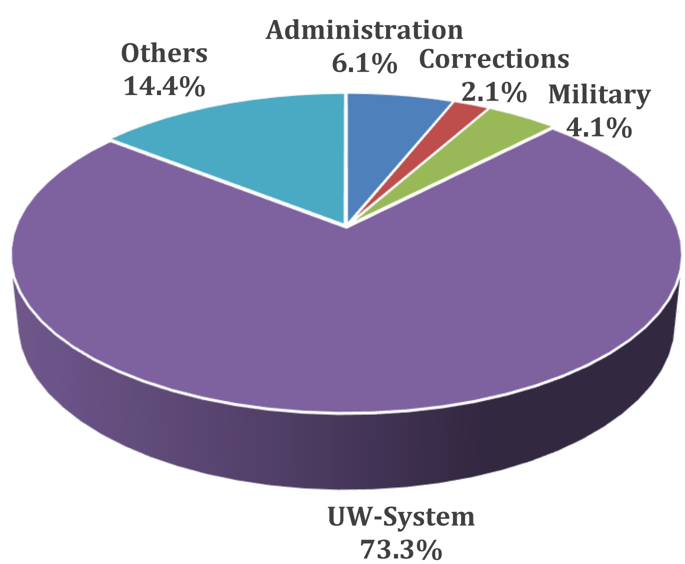


Figure 23: Distribution of Wisconsin State Funding (2001-17)

The figure below breaks down projects affiliated with the UW-System by campus. Only the top 8 campuses by volume are shown. 40% of projects were executed for UW-Madison, while the next 7 most voluminous campuses comprised 45% and the remaining 5 only 15% (listed as 'Others').

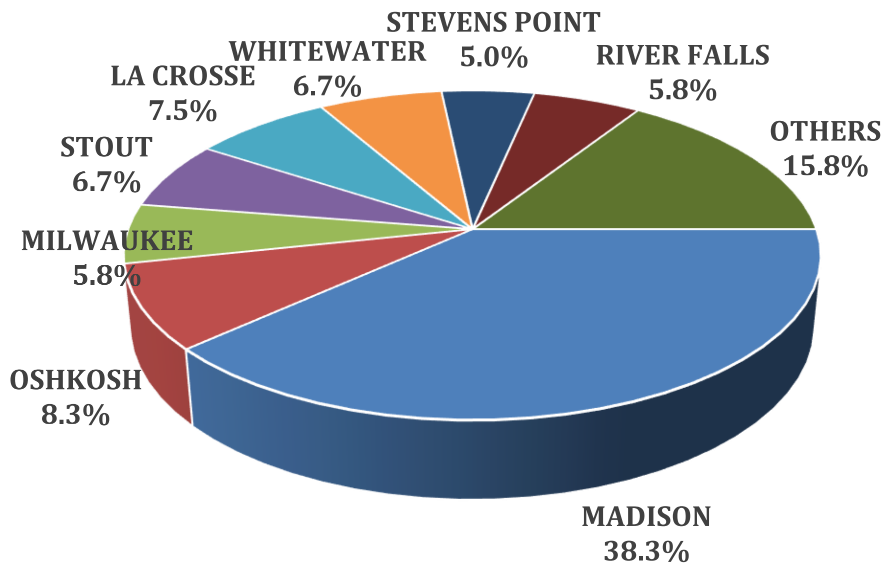


Figure 24: Distribution of Projects by UW-System Campuses

As 25% of the projects in this study were not associated with the UW-System. Therefore, all projects are broken down by location in Figure 25 below. A similar phenomenon occurs: Madison leads in volume by 38%, followed by Oshkosh (15%) and Milwaukee (9%). A notable difference is Eau Claire, with (8%) of projects, while UW-Eau Claire was not among the top eight campuses.

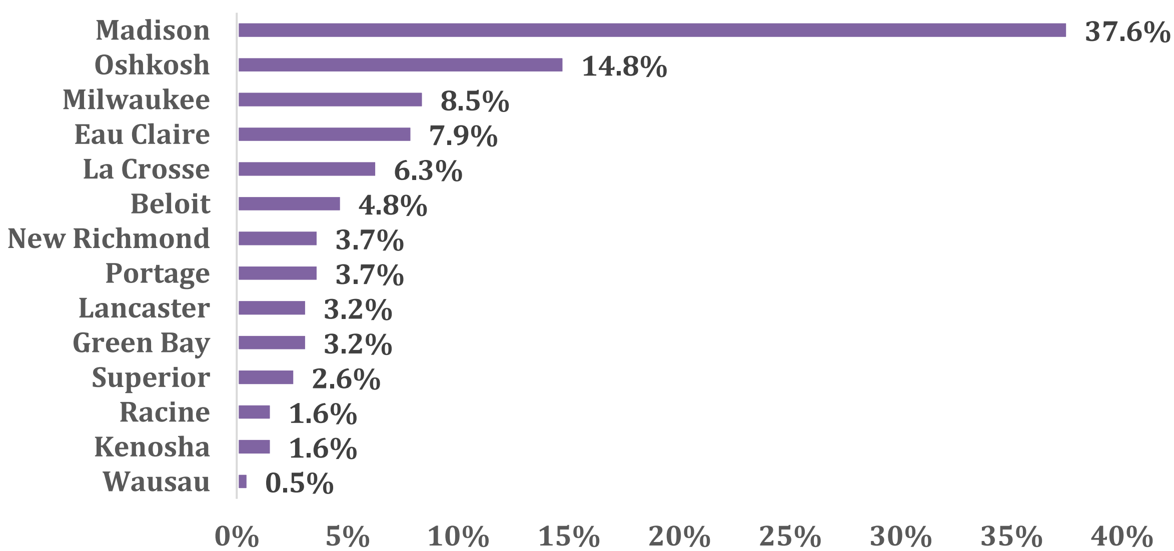


Figure 25: Distribution of Projects by Location

The figure below includes all projects in the data set, broken down by construction type. The majority, as evident, were new construction.

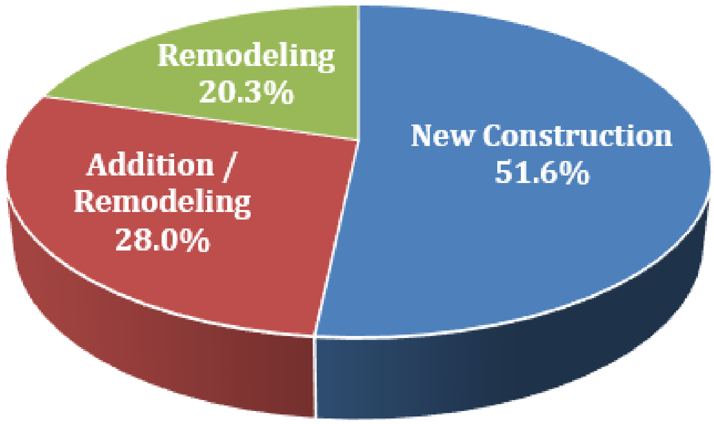


Figure 26: Distribution by Construction Type

Figure 27 below breaks down projects by facility type. Only the 10 most frequent types are pictured. Academic facilities (18%) were the most frequent. While visually, 'Other' facilities encompass 23% of the data, this figure is comprised of numerous small percentages which represent museums, libraries, fish hatcheries, parks, etc. As such, these projects were not included in analysis which pertained to square footage, as they are specialized and infrequently executed.

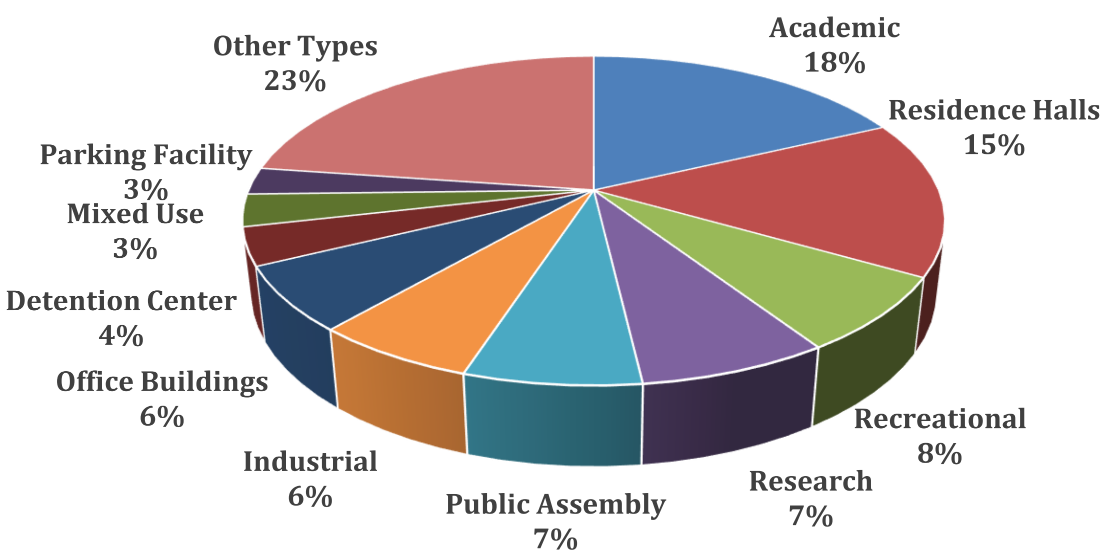


Figure 27: Distribution by Facility Type

Figure 28 below is central to this research. As has been repeatedly stated in this thesis, DBB dominates State projects. Indeed, 72% of the projects by value were DBB. While the figures for CMAR (16%) and DB (12%) might be low, the significant magnitude of the dataset, concludes that this distribution is workably representative of the current state of affairs.

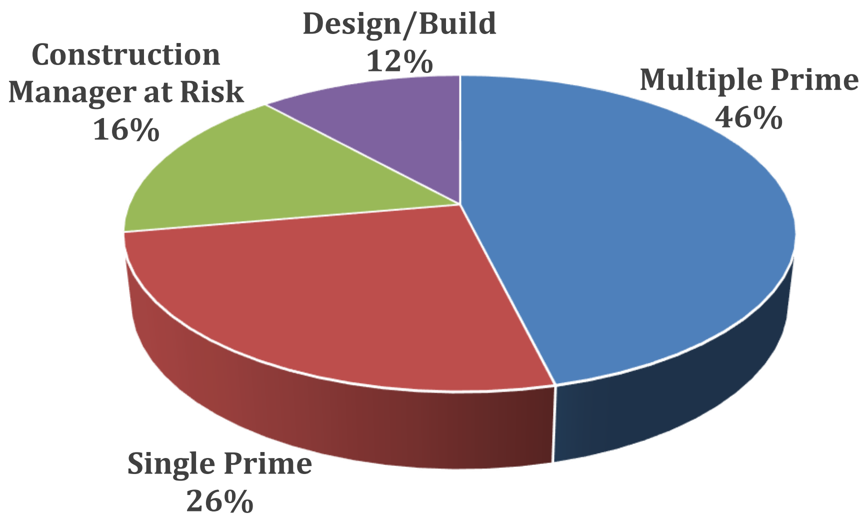


Figure 28: Distribution of Delivery Methods by Dollar Value

Since the study does encompass a large period of time, Figure 29 below was prepared to demonstrate the distribution of projects by completion year. The distribution is fairly consistent, with the exception of 2000-2002.

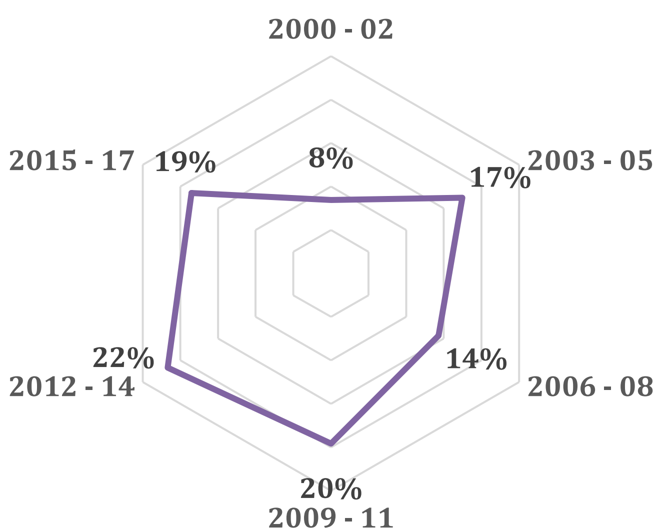


Figure 29: Distribution of Projects by Completion Year

As can be seen below (Figure 30), DB peaked between 2000 and 2002, while CMAR peaked between 2012 and 2014. Peak is a relative term, as these delivery methods reached only 25% and 10% market share, respectively. Since the 2000s, DBB delivery has held a monopoly on projects, with MP reaching 75% of executed projects in 2011, before being replaced by SP, which currently holds 100% of the market share.

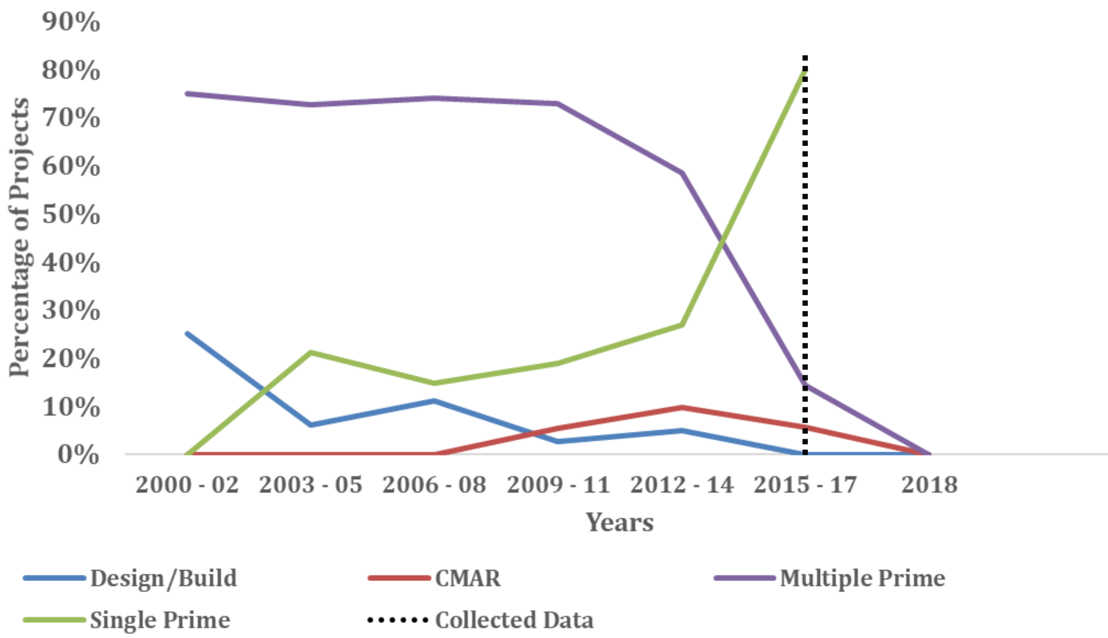


Figure 30: Projects Completion Year by Delivery Method

Figure 31 below details the duration of projects. 50% were executed over a period between 4 and 6 years, while 80% over a period between 3 and 7 years per project.

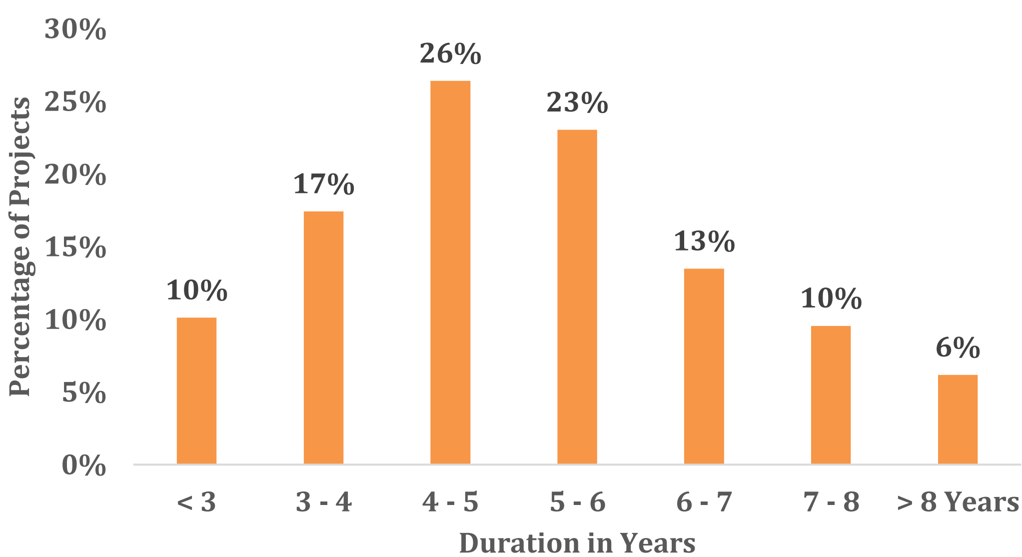


Figure 31: Distribution by Total Project Delivery Duration

Seen below in Figure 32, 80% of projects had a budgeted cost of $5 and $50 million.

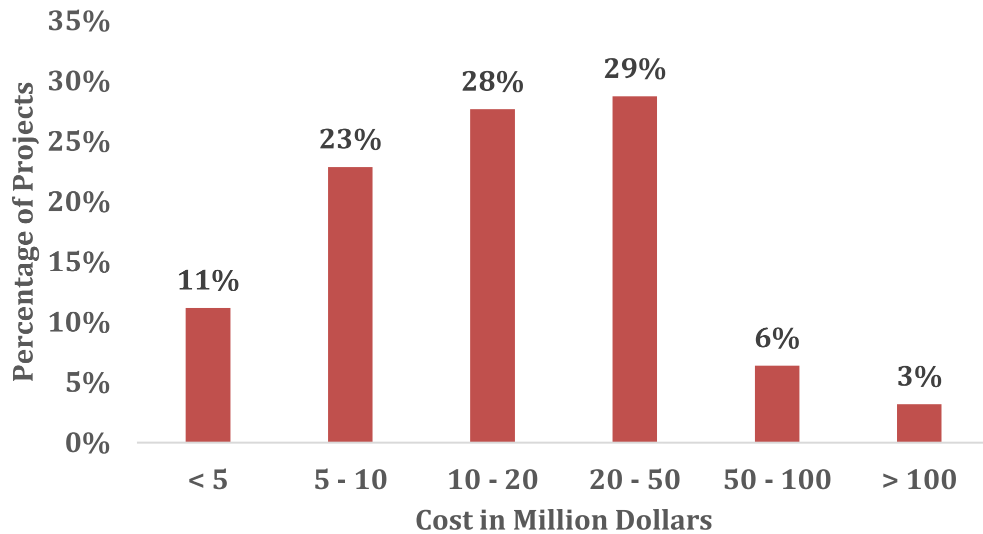


Figure 32: Distribution by Total Project Delivery Cost

By combining the two immediately preceding charts, the following Figure 33 reveals the average project duration in years of each budget range.

Per this chart, a project with cost between $10 and $20 million consumes 5.3 years on average for design and construction. An abnormality was also found in this graphical analysis: projects between $50 and $100 million consume less time than less expensive projects. Further investigation revealed that 70% of the projects in this range used more Collaborative delivery methods, featuring quicker execution as will be explained in Chapter 5 of this document.

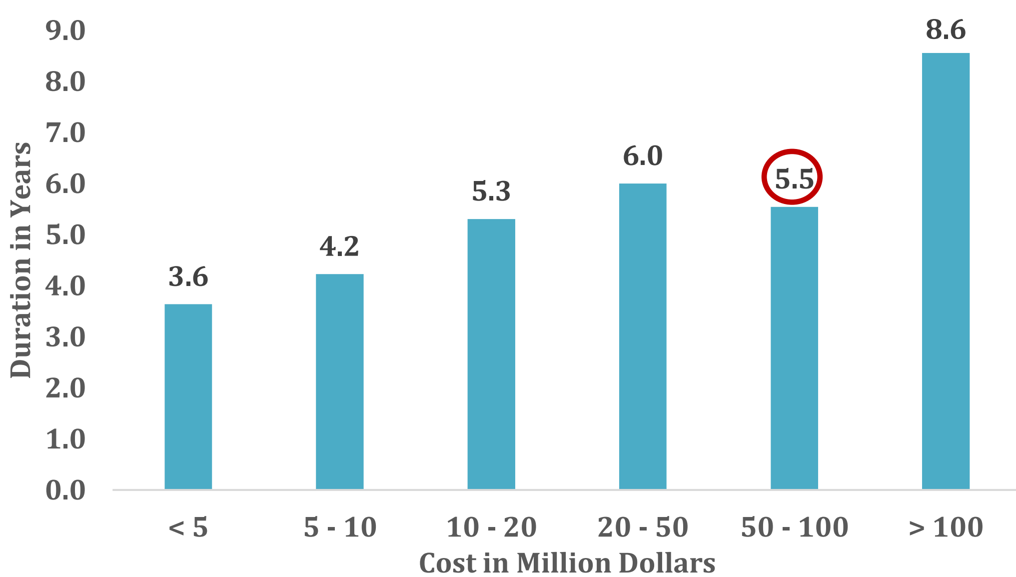


Figure 33: Average Project Delivery Duration Per Project Delivery Cost

Below is a classification of projects by adjusted gross square feet (Figure 34). 50% of the data fell between 50 and 150 thousand GSF, while 80% fell between 25 and 200 thousand GSF.

Figure 34: Distribution of Project Gross Footage Area

Once again combining charts, Figure 35 below contrasts duration with GSF. On average, projects with GSF between 25,000 and 50,000 consumed 4.7 years, while projects of 3 to 4 times the size (150,000 to 200,000) consumed on average 5.8 years. This chart exhibits a roughly linear increase in duration.

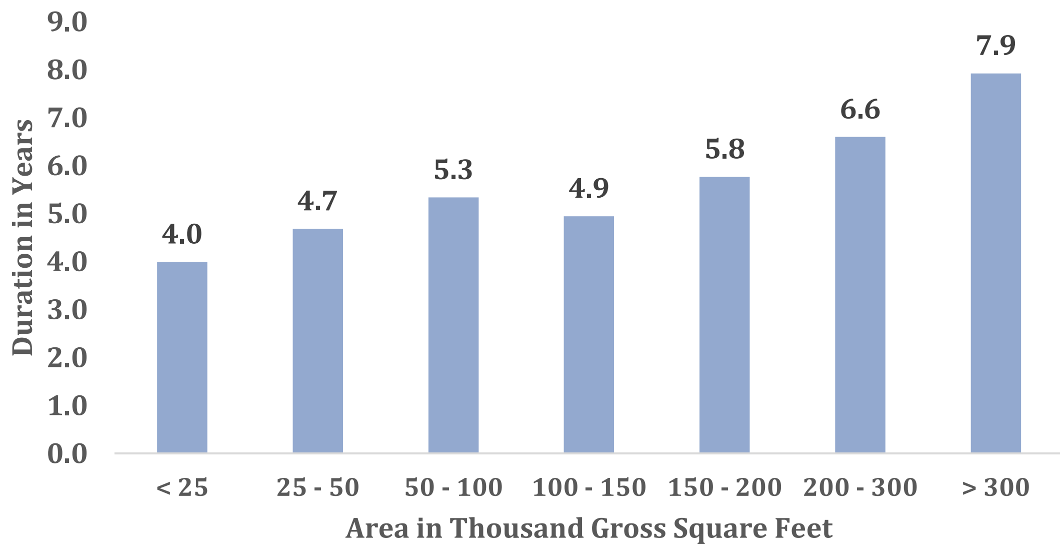


Figure 35: Average Project Delivery Duration Per Gross Footage Area

However, another nonlinear trend was found in the combined chart comparing GSF to delivery cost (Figure 36). Projects with GSF between 25,000 and 50,000 cost on average $7.8 million, while those of 3 to 4 times the size (150,000 to 200,000) cost on average $38.3 million. The primary takeaway from this chart is the rapid scaling of cost with size, which contrasts with the slower buildup of duration presented in the previous chart.

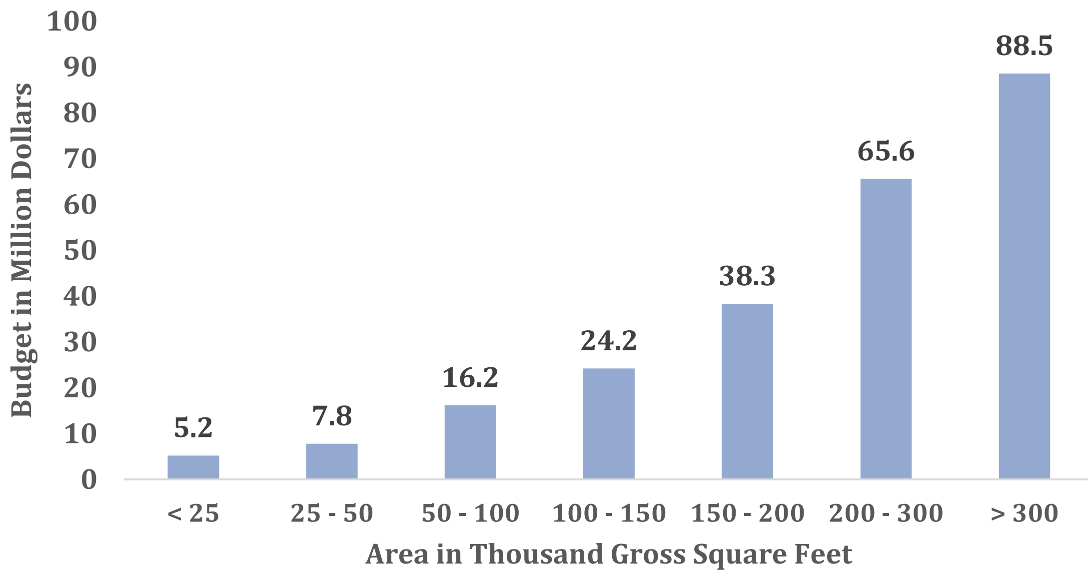


Figure 36: Average Project Budget Per Gross Footage Area

5 Chapter 5: Computation & Analysis

In this chapter, the central research question will be discussed: do the various project delivery methods investigated score significantly differently from each other in terms of project performance. Furthermore, in the case of projects executed at the various UW campuses, it was investigated if the campus on which a project was built has any impact on that project's performance.

This chapter is divided into five sections.

1) The first section discusses the likelihood of occurrence of each of the 6 categories for issuing change orders allowed by the State. The likelihood is broken down into both cost-related and schedule-related categories, which are then averaged to assess overall change. This analysis determines which type of change order is more likely to occur on projects delivered under different methods, thus also providing a comparison between different project delivery methods by volume of change orders. For example, if a delivery method were to have a higher instance of change orders due to Design Oversight, that would indicate a lower overall design review or quality, which usually predicates higher levels of design-related changes during construction.
2) The second section consists of an individual comparative performance assessment, which highlights any material differences between the various delivery methods. Nine individual performance metrics across five performance areas were assessed, with statistically significant differences highlighted. Extended analysis was further performed to distinguish differences in performance between Collaborative and Traditional delivery types (see Chapter 1 – Definition of Terms). The researcher would like to take this opportunity to reiterate that the selection of these nine performance metrics was the intersection of State-tracked data and reviewed literature.
3) The third section performs comparative analysis on the PPI scores returned by the mathematical model defined in previous chapters. For more information on the PPI score, refer to Chapter 4. Within this section, regression plots which highlight the Spearman's Rho correlation coefficient between the individual component performance metrics and the overall PPI score are also presented, which serve as a method to visualize the performance of different project types relative to each performance metric, and further highlight performance trends between Collaborative and Traditional delivery methods.
4) The fourth section applies robust statistical validation techniques to confirm the results of the model applied in Chapter 4. These techniques, as applied here, are like the model in function, but unlike the innovative model are accepted as standard practice. Their use in this case thus lends credence to the model, as the Traditional

methods confirm the innovative methods. Those two tools are Confirmatory Factor Analysis and Partitioning Clustering Analysis.

5) The fifth section explores an additional variable that is more localized to the state of Wisconsin. The performance of projects across the numerous UW-System campuses was examined. This methodology compares the performance of campuses using their average PPI scores, highlights strengths and weaknesses in the performance of each campus, in light of the eight performance metrics. The aim of this section is to direct the focus of each campus towards areas of strengths and potential improvements in its projects. However, it should be noted that the performance score assigned to each campus is a factor of multiple other factors including project complexity, performance of project participants, delivery method, to name a few. Therefore, further research is required to closely and accurately represent the performance of each campus/agency, taking into account all those factors.

5.1 Overall Performance Assessment

When project performance is under study, two key metrics are traditionally considered: cost and schedule growths. While other factors should be considered, an overall assessment identifies performance drivers for further investigation.

5.1.1 Construction Cost Growth

In this subsection, construction cost growth is examined. In this context, it is computed as the differences between the Base Bid and the Final Construction Cost. In some Collaborative delivery methods cost growth is computed as the difference between actual cost at completion and initial Guaranteed Maximum Price (GMP).

As depicted below, only 4% of projects studied experienced negative construction cost growth, 1% had zero cost growth and 95% of projects with positive cost growth. On average, Base Bid values increased by 9%.

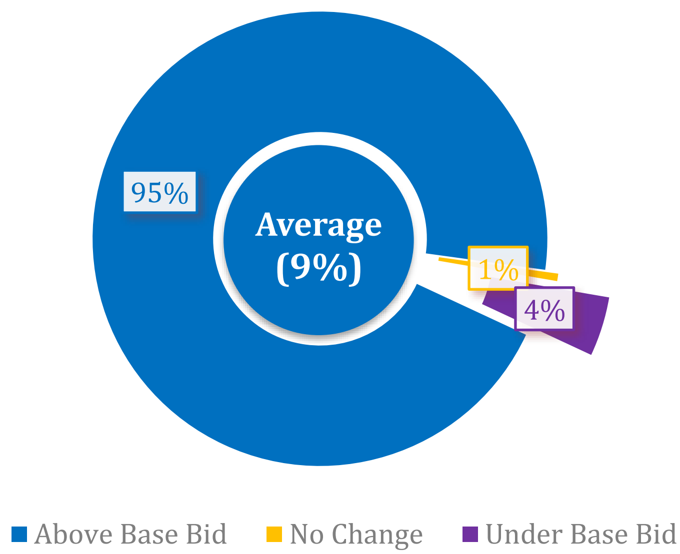


Figure 37: Distribution of Projects by Construction Cost Growth

Furthermore, the 4% of projects that finish below their base bid values save on average only 1.9% of the contractual cost. In other words, when a project comes in below base bid, it does so only to a small degree, whereas when a project exceeds its base bid, it does so to a more significant degree – the median value of above base bid projects was 6.8% which exceeds below base bid projects (1%).

Above base bid projects also tended to exhibit higher variability than below base bid projects, with standard deviations of 7.3% and 3.2%, respectively. This equates to a higher risk of exceeding base bid by significantly more than the 9% average value.

The table below presents the descriptive statistical analysis of cost performance.

Table 4: Descriptive Statistics for Construction Cost Growth Per Project Status

Statistical Metric	Above Base Bid	No Change	Below Base Bid
Number of Projects	180	1	8
Total Adjusted ($)	$345,856,286	$0.00	-$4,489,141
Total Unadjusted ($)	$275,818,719	$0.00	-$3,580,069
Mean	8.9%	0.0%	-1.9%
Maximum	47.9%	0.0%	-0.1%
75% Percentile	11.0%	0.0%	-0.1%
Median	6.8%	0.0%	-1.0%
25% Percentile	4.0%	0.0%	-1.6%
Minimum	0.9%	0.0%	-9.0%
Standard Deviation	7.3%	0.0%	3.2%

5.1.1.1 Comparative Analysis by Delivery Methods

However, the analysis in the preceding section does not account for delivery method. This is an essential piece of information to understand, as insight into which delivery methods perform best could be very useful to decision-makers. Referring to the plot below, it is evident that MP delivery is responsible for the clear majority of projects that experienced overruns (62%), while SP accounts for only 31%. Collaborative delivery methods, contrarily, amounted to only 7% of those projects (4% DB and 3% CMAR).

This suggests that more Collaborative projects tend to have better performance over their construction costs. However, it should be reiterated that DBB in both SP and MP cases does not traditionally allocate a contingency fund, and a significant portion of the cost growth reported herein may result from normal use of contingency. Also, the data for DB and CMAR contain contingency and as such are perhaps better equipped to respond to and absorb cost overruns.

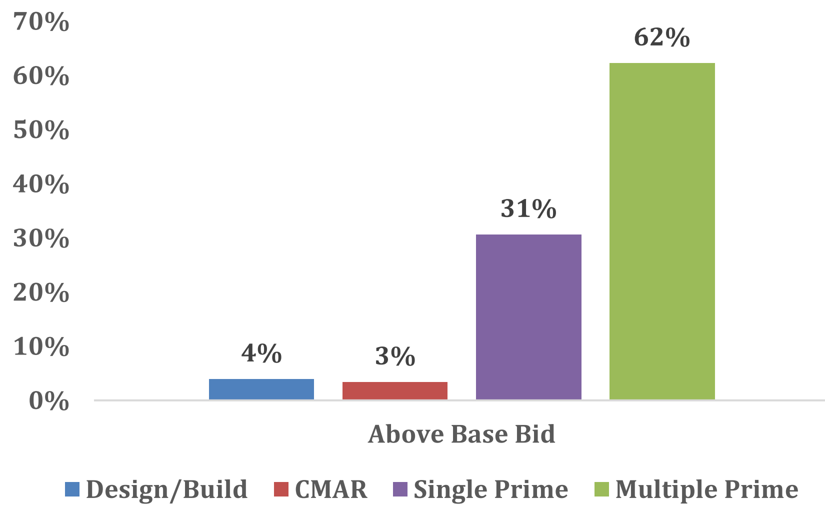


Figure 38: Distribution of Projects with Construction Cost Overrun by Delivery Method

Conclusions based solely on the visual plot here pictured might be misleading. While it is clear that the percentage of projects above base bid that were collaboratively delivered is much lower, this figure does not adjust for the original distribution of projects by delivery method present in the data. As such, an additional statistical test was required to validate these observations and determine if there is a statistically significant association between project delivery method and construction cost overrun.

The contingency table below separates the data by delivery method and cost compliance status as a preparatory step for statistical testing. While visually there may appear to be an asymmetry in the sample, it is important to understand that even casting the widest net possible, there are simply so few projects completed within the study period using DB or CMAR that the sample obtained represents the best possible data collection. Furthermore, statistical controls were used to account for sample size.

Table 5: Contingency Table of Projects with Construction Cost Overrun & PDMs

Above Base Bid		*Project Delivery Method*				*Row Total*
		Design/Build	**CMAR**	**Single-Prime**	**Multiple-Prime**	
	Yes	**7**	**6**	**55**	**112**	**180**
	No	**5**	**2**	**2**	**0**	**9**
	Column Total	**12**	**8**	**57**	**112**	**189**

As the data segregation reveals an observation of zero, Aylmer's test is conducted and validated by Fischer's Exact Test accompanied by a Monte Carlo Simulation with bootstrapping of 2000 replications. Both Aylmer and Fischer in this case result in a *p*-value of 0.000, which is sufficient to reject the null hypothesis of no association between delivery method and construction cost growth performance, with a statistically significant difference found at the 95% confidence level. This suggests that CMAR and/or DB projects achieve better cost performance than SP and/or MP.

Table 6: Significance of Association Between Construction Cost Overrun & PDMs

Statistical Test	*p*-value	Significant at 95% Level	Significant at 90% Level
Aylmer	0.000	Significant	--
Fisher's Exact	0.000	Significant	--

5.1.1.2 Comparative Analysis by Delivery Types

In addition to comparison by delivery method, consolidating projects into Collaborative and Traditional delivery provides both more robust results (by consolidating observations and increasing the count in each pool) and also confirms the previous findings in a summarized manner. As shown below, Traditional delivery accounts for 93% of projects with cost overrun, while as previously stated only 7% of these projects were Collaborative. This provides additional evidence for the higher performance of Collaborative projects in terms of cost performance.

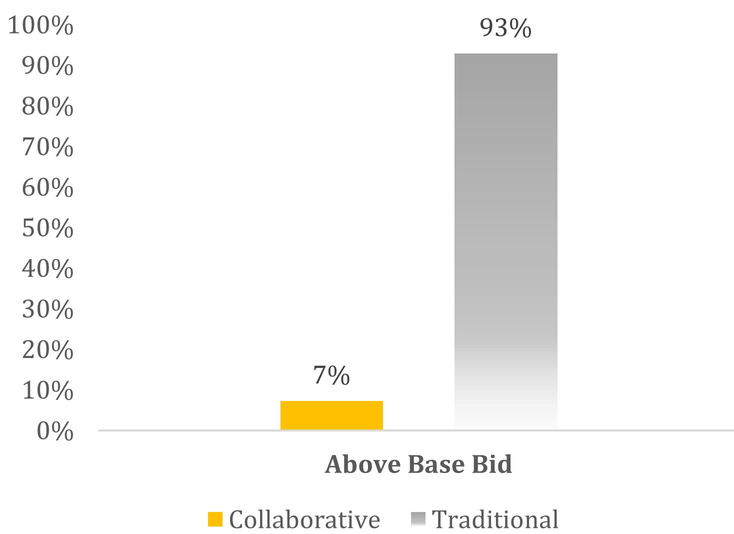


Figure 39: Distribution of Projects Construction Cost Overrun by Delivery Type

However, as was seen in subsection 5.1.1.1, these conclusions are not adjusted for the relative distribution of projects by delivery method. Therefore, statistical tests were conducted to validate the previous observations. The test examines if there is a statistically significant association between delivery type and cost overrun. As with the previous procedure, the table below classifies data by delivery type and cost compliance status as a preparatory step.

Table 7: Contingency Table of Projects with Construction Cost Overrun & PDTs

Above Base Bid		*Project Delivery Type*		
		Collaborative	**Traditional**	***Row Total***
	Yes	**13**	**167**	**180**
	No	**7**	**2**	**9**
	Column Total	**20**	**169**	**189**

In this case, no values are zero, but some expected values are less than 5. Therefore, Fischer's exact test is conducted, accompanied by the Chi-Square test with Yates continuity correction and a Monte Carlo Simulation (bootstrapping) of 2000 replicates. As shown below, both tests return *p*-values of 0.000, sufficient to reject the null hypothesis that there is no association between project delivery type and construction cost overrun, with a statistically significant difference at the 95% confidence level. This is yet additional evidence that Collaborative projects are better able to maintain their initial costs as compared to Traditional ones.

Table 8: Significance of Association Between Construction Cost Overrun & PDTs

Statistical Test	*p*-value	Significant at 95% Level	Significant at 90% Level
Chi-Square	0.000	Significant	--
Fisher's Exact	0.000	Significant	--

5.1.2 Design Cost Growth

This subsection examines the overall change in design initial cost. Similar to the definition of construction cost growth previously, this refers to the Percent Difference between the initial design cost and the final design cost.

As shown below, only 2% of projects' design was completed below their initial cost, 4% with no change, and 94% above their initial cost. On average, design cost increased by 30%.

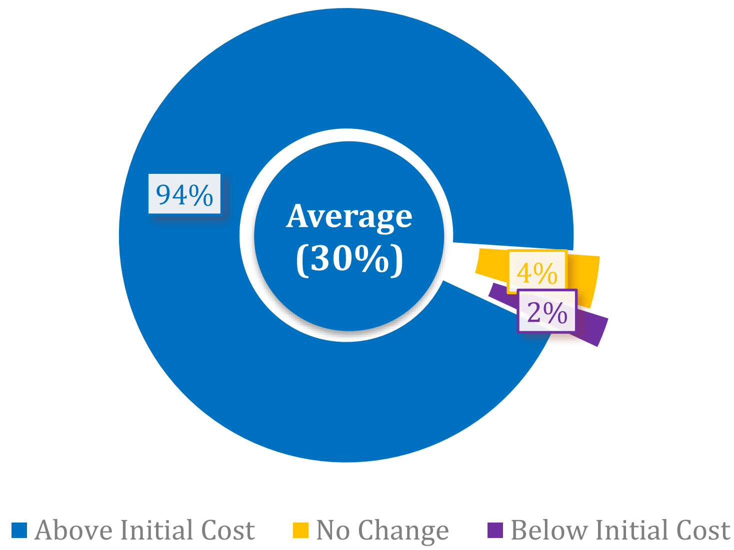


Figure 40: Distribution of Projects by Design Cost Growth

Not only do only 2% of projects finish below their initial cost, these projects on average save only 6.7% of cost. The same proportion exists in the median values of these projects: 5.3% and 17.9%, respectively.

It was once again found that cost overrun projects exhibited higher variability, with a standard deviation of 47.9%. Below initial cost projects had a standard deviation of 7.6%. This translates to a significantly higher risk of exceeding cost by a value greater than 30%.

Table 9: Descriptive Statistics for Design Cost Growth Per Project Status

Statistical Metric	Above Initial Cost	No Change	Below Initial Cost
Number of Projects	173	7	4
Total Adjusted ($)	$129,735,441	$0.00	-$343,119
Total Unadjusted ($)	$92,521,130	$0.00	-$244,696
Mean	32.4%	0.0%	-6.9%
Maximum	300%	0.0%	-0.7%
75% Percentile	37.8%	0.0%	-0.8%
Median	17.9%	0.0%	-5.3%
25% Percentile	6.5%	0.0%	-11.4%
Minimum	0.1%	0.0%	-16.3%
Standard Deviation	47.9%	0.0%	7.6%

5.1.2.1 Comparative Analysis by Delivery Methods

Projects which experienced design cost overruns were then classified by delivery method to discover which methods exhibited a higher percentage of overruns. This comparison is plotted below. As shown, 59% of overrun projects were MP delivery, while 32% were SP. Collaborative methods (DB and CMAR) each were responsible for 5% of those projects. This suggests that Collaborative delivery methods are better able to maintain initial design cost.

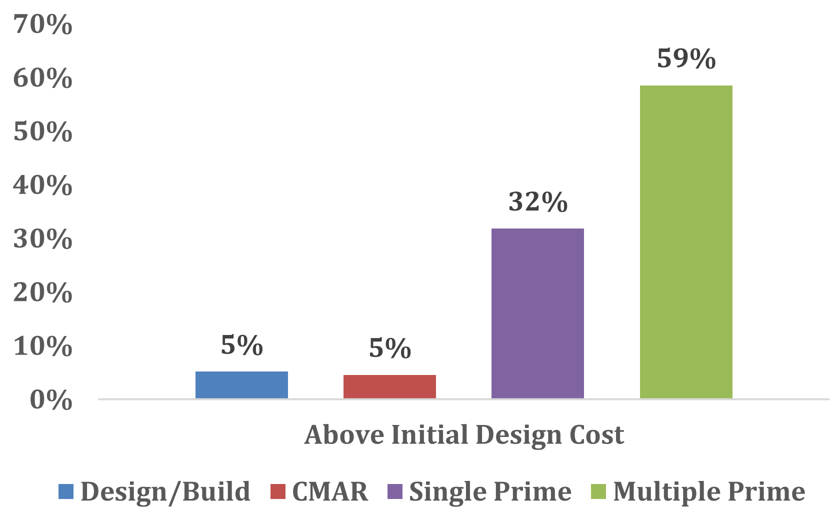


Figure 41: Distribution of Projects Design Cost Overrun by Delivery Method

While these results seem to indicate better design cost performance in more Collaborative projects, as previously, this figure does not consider distribution by delivery methods. Once again, additional testing was performed to assess if there was a statistically significant difference in performance between the various delivery methods. Below is the contingency table for this analysis.

Table 10: Contingency Table of Projects with Design Cost Overrun & PDMs

Above Initial Design Cost		*Project Delivery Method*				*Row Total*
		Design/Build	**CMAR**	**Single-Prime**	**Multiple-Prime**	
	Yes	**9**	**8**	**56**	**103**	**176**
	No	**2**	**0**	**1**	**8**	**11**
	Column Total	**11**	**8**	**57**	**111**	**187**

In this instance, there is an observation equal to zero. Aylmer's test is thus conducted and validated by Fischer's exact test with a 2000 replicate Monte Carlo Simulation (bootstrapping). Aylmer's test results in a *p*-value of 0.074, sufficient to reject the null hypothesis at the 90% confidence level. Fischer's exact test, however, returns a *p*-value of 0.115 which is insufficient to reject the null hypothesis at the 95 or 90% confidence levels.

These results suggest that DB and CMAR projects have higher performance over their design cost, per Aylmer's test. However, there is no corresponding validation by Fischer's exact test, which suggests instead that all delivery methods have similar performance. From this inconsistency, no substantial conclusions can be drawn, however it is likely from data trends that Collaborative projects have slightly better performance, but not to a significant degree.

Table 11: Significance of Association Between Design Cost Overrun & PDMs

Statistical Test	*p*-value	Significant at 95% Level	Significant at 90% Level
Aylmer	0.074	Not Significant	Significant
Fisher's Exact	0.115	Not Significant	Not Significant

5.1.2.2 Comparative Analysis by Delivery Types

By grouping projects by delivery type, the observation count is increased, and the results of the previous section are confirmed in a summarized manner. As plotted below, 90% of projects with cost overrun used Traditional delivery, while only 10% were Collaborative. This again suggests intuitively that Collaborative projects have more performance over their design cost than Traditional ones.

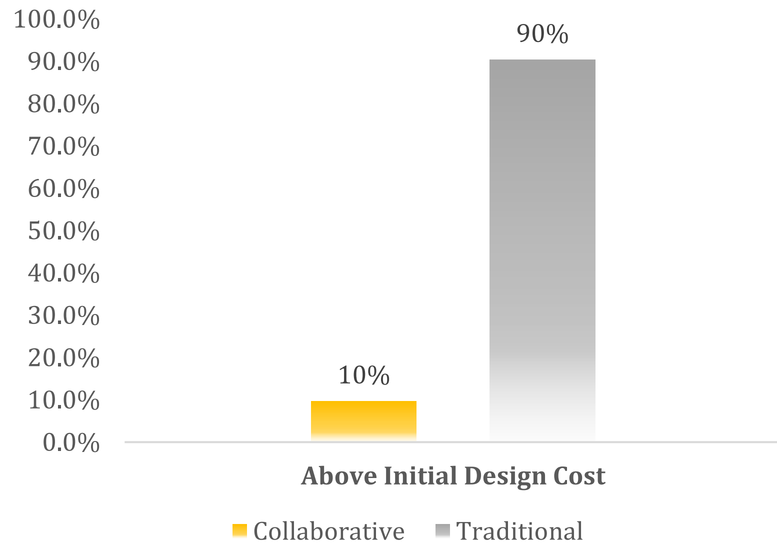


Figure 42: Distribution of Projects Design Cost Overrun by Delivery Type

Once again, while this presentation is a useful comparison, it should not inspire conclusions. Therefore, statistical tests were performed to validate the observations. The contingency table for this analysis is below.

Table 12: Contingency Table of Projects with Design Cost Overrun & PDTs

		Project Delivery Type		
		Collaborative	**Traditional**	***Row Total***
Above Initial Design Cost	**Yes**	**17**	**159**	**176**
	No	**2**	**9**	**11**
	Column Total	**19**	**168**	**187**

In this instance no values are equal to zero, but some expected values are less than 5. As such, Fischer's exact test is conducted and subsequently validated by a Chi-Square test with Yates continuity correction and a 2000-replicate Monte Carlo simulation (bootstrapping).

The table below presents the findings. *p*-values of 0.616 and 0.310 were returned by the Chi-Square test and Fischer's Exact Test, respectively. As such, there is insufficient evidence in both cases to reject the null hypothesis at both the 95 and 90% confidence levels. Therefore, there is no evidence to conclude that both delivery types perform differently in their ability to maintain initial design costs.

Table 13: Significance of Association Between Design Cost Overrun & PDTs

Statistical Test	*p*-value	Significant at 95% Level	Significant at 90% Level
Chi-Square	0.616	Not Significant	Not Significant
Fisher's Exact	0.310	Not Significant	Not Significant

5.1.3 Delivery Cost Growth

The next metric considered was Delivery Cost Growth. Similar to previous metrics, this examines the difference between Initial Delivery Cost and Final Delivery Cost over the whole delivery period. As such it includes design costs as well as construction costs. The figure below shows that 97% of projects experienced Delivery Cost Growth, and those projects increased on average by 10.3%.

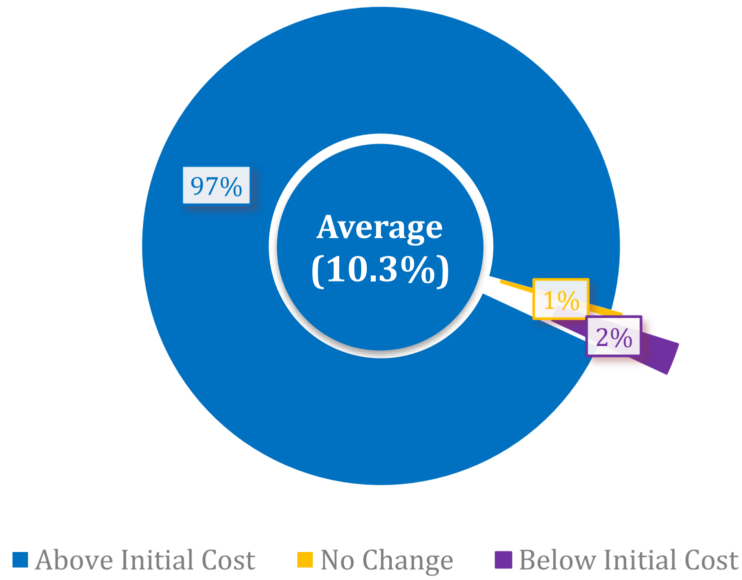


Figure 43: Distribution of Projects by Delivery Cost Growth

Of the 2% which delivered below initial cost, the average savings was 3.6%. The same observation is present when comparing the medians of both types of projects: 1.2% and 8.4%, respectively. Projects with cost overrun also exhibited higher variability, with a standard deviation of 7.6%, compared to the 4.3% demonstrated by projects with negative cost growth. Thus, there is an increased risk of projects that experience cost overruns, doing so by several degrees of variance more than the 10.3% average.

The results are not uncommon. A study in 2015 reported that 98% of mega projects suffer cost overruns greater than 30% (McKinsey & Co. 2015). Another study estimated that 9 out of 10 mega projects (90%) experience cost overruns (Flyvbjerg 2017).

Table 14: Descriptive Statistics for Delivery Cost Growth Per Project Status

Statistical Metric	Above Initial Cost	No Change	Below Initial Cost
Number of Projects	184	1	3
Total Adjusted ($)	$450,245,898	$0.00	-$2,466,257
Total Unadjusted ($)	$356,506,887	$0	-$1,952,795
Mean	10.3%	0.0%	-3.6%
Maximum	46.5%	0.0%	-1.0%
75% Percentile	13.3%	0.0%	-1.1%
Median	8.4%	0.0%	-1.2%
25% Percentile	5.1%	0.0%	-4.9%
Minimum	0.3%	0.0%	-8.6%
Standard Deviation	7.6%	0.0%	4.3%

5.1.3.1 Comparative Analysis by Delivery Methods

When delivery cost growth performance is broken out by delivery method, the figure below is produced. MP accounted for 61% of projects, while SP was responsible for 31%. Collaborative delivery methods, however, each amounted for only 4%, indicating that Collaborative methods are more able to deliver projects where the final cost is close to the initial contractual delivery cost.

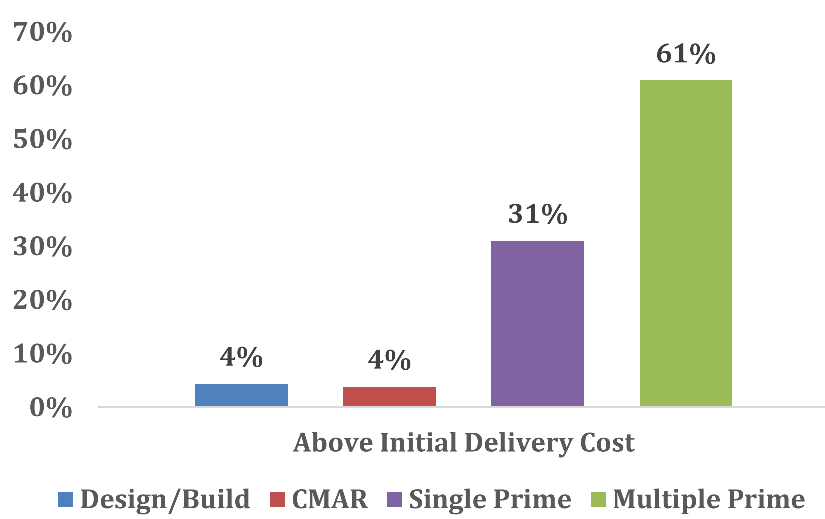


Figure 44: Distribution of Projects Delivery Cost Overrun by Delivery Method

While these results are visually evident, a statistical confirmation is still required as the overall distribution of projects by delivery method is not accounted for. This test, based on the contingency table below, determines if statistically significant correlations exist between delivery method and projects above their delivery cost.

Table 15: Contingency Table of Projects with Delivery Cost Overrun & PDMs

Above Initial Delivery Cost		*Project Delivery Method* Design/Build	CMAR	Single-Prime	Multiple-Prime	*Row Total*
	Yes	**8**	**7**	**57**	**112**	**184**
	No	**4**	**1**	**0**	**0**	**5**
	Column Total	**12**	**8**	**57**	**112**	**189**

In this case, some observations are zero. As such, Aylmer's test was performed and subsequently validated by Fischer's exact test with a Monte Carlo simulation (bootstrapping) of 2000 replicates. As shown in the table below, both Aylmer's and Fischer's tests return a *p*-value of 0.000, sufficient to reject the null hypothesis at the 95% significance level. This suggests that DB and CMAR projects perform better in terms of delivery cost than Single or Multiple prime.

Table 16: Significance of Association Between Delivery Cost Overrun & PDMs

Statistical Test	*p*-value	Significant at 95% Level	Significant at 90% Level
Aylmer	0.000	Significant	--
Fisher's Exact	0.000	Significant	--

5.1.3.2 Comparative Analysis by Delivery Types

Beyond comparing the impact of each delivery method singularly, by consolidating the delivery methods into two pools, Collaborative and Traditional, further conclusions may be drawn. This consolidation provides two benefits: it increases the observation count, providing more robust statistical results; and it further confirms by summarization the findings of the preceding section.

Plotted below are these results. It is visibly evident that Traditional delivery has a significantly higher percentage of cost overrun projects (92%), while only 8% of cost overrun projects were Collaborative. This suggests that Collaborative delivery projects have better performance over delivery cost.

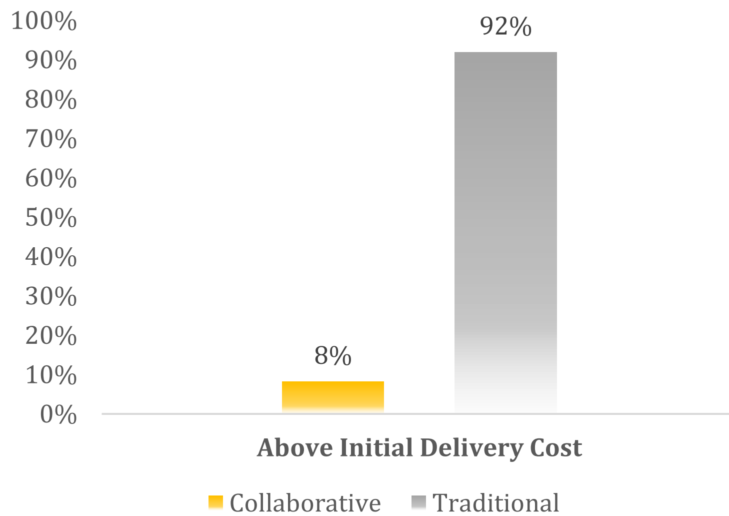


Figure 45: Distribution of Projects Delivery Cost Overrun by Delivery Type

As was highlighted earlier, drawing conclusions simply from this figure without considering overall distribution of projects by type could be deceiving. Therefore, a secondary test was conducted to validate the previous observation. This test will determine if there are any associations between delivery type and delivery cost overrun. The contingency table below classifies projects by delivery type and cost compliance status.

Table 17: Contingency Table of Projects with Delivery Cost Overrun & PDTs

		Project Delivery Type		
		Collaborative	**Traditional**	***Row Total***
Above Initial Delivery Cost	**Yes**	**0**	**169**	**184**
	No	**15**	**5**	**5**
	Column Total	**15**	**174**	**189**

As in this scenario one observation is equal to zero but the degrees of freedom do not meet Aylmer's assumption, Aylmer's test could not be performed. Therefore, Fischer's Exact test was conducted with a 2000-replicate Monte Carlo Simulation (bootstrapping) in addition to a Chi-Square test with Yates continuity correction and a 2000-replicate Monte Carlo Simulation (bootstrapping).

Shown below, both tests return a *p*-value of 0.000, sufficient to reject the null hypothesis that no association exists between delivery type and delivery cost performance at the 95% confidence level. Therefore, it is inferred that Collaborative delivery has a higher ability to maintain its initial delivery cost as compared to Traditional projects.

Table 18: Significance of Association Between Delivery Cost Overrun & PDTs

Statistical Test	*p*-value	Significant at 95% Level	Significant at 90% Level
Aylmer*	NA	NA	NA
Chi-Square	0.000	Significant	--
Fisher's Exact	0.000	Significant	--

*assumption of DF > 0 was not met

5.1.4 Change in Construction Schedule (excluding extensions)

Within this subsection, overall change in construction schedule (excluding approved extensions) was examined as a Percent Difference between initial contractual duration and actual final duration (excluding approved duration extensions).

Pictured below, 12% of studied projects were completed ahead of schedule, 8% finished on schedule, and 80% finished behind schedule. Those projects which finished behind schedule on average increased duration by 44%.

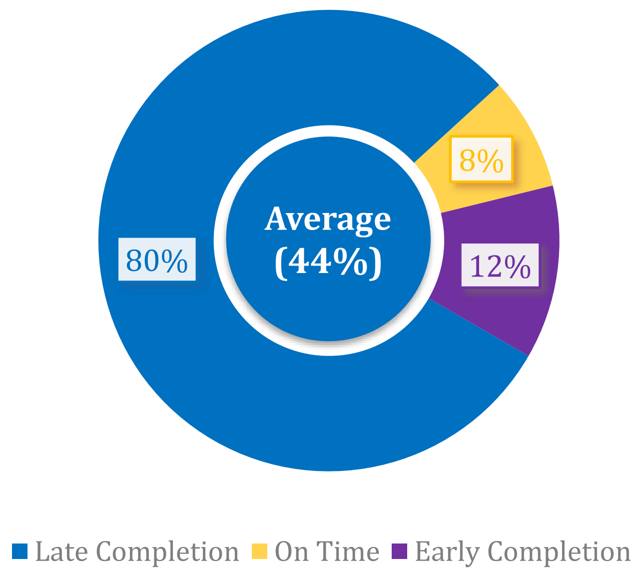


Figure 46: Construction Schedule Change (excluding extensions)

Of those projects which were completed ahead of schedule (12%), they realized only 7.3% in schedule savings. While this is still a considerable saving, and indeed any saving of schedule on a construction project is admirable, it pales in comparison to the 44% of overrun time. Comparison of the medians of behind- and ahead-of-schedule projects reveals that behind-schedule projects slip by 17.7%, while ahead-of-schedule projects make only 6.6% gain. Thus, when a project slips behind schedule, it does so by a much more significant degree.

Behind-schedule projects have significantly higher variability when compared to ahead-of-schedule projects, with standard deviations of 57% and 4.9%, respectively. This translates to a significantly higher risk of finishing construction behind schedule, by a value far higher than the 44% average figure. On the other hand, those projects which finish ahead-of-schedule are consistently closer to their 7.3% average value.

Refer to the table below for descriptive statistical analysis of construction schedule change.

Table 19: Descriptive Statistics for Construction Schedule Change (excl. extensions) Per Project Status

Statistical Metric	Behind Schedule	On Schedule	Ahead of Schedule
Number of Projects	151	15	23
Total Duration (Years)	85	0	-3
Mean	44.3%	0.0%	-7.3%
Maximum	461.7%	0.0%	-0.5%
75% Percentile	40.6%	0.0%	-2.9%
Median	17.7%	0.0%	-6.6%
25% Percentile	10.3%	3.3%	-11.0%
Minimum	0.2%	0.0%	-17.7%
Standard Deviation	57.0%	0.0%	4.9%

5.1.4.1 Comparative Analysis by Delivery Methods

When schedule change (excl. extensions) was broken out by delivery method, further insight can be gained. Plotted below are these results. Clearly evident is the significantly higher percentage of MP delivery (68%), and SP (28%). The Collaborative delivery methods, DB and CMAR, each accounted for only 2% of the behind-schedule projects. This suggests once again that Collaborative delivery methods have better performance over construction schedule.

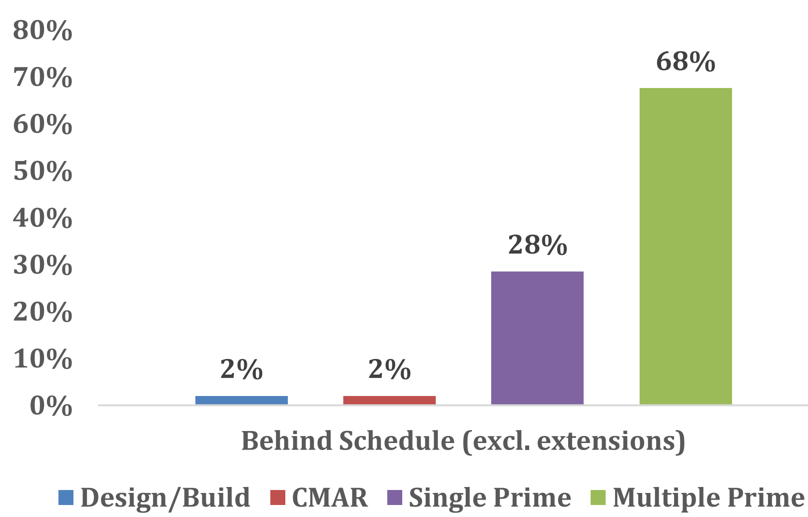


Figure 47: Distribution of Projects Behind Schedule (excl. extensions) by Delivery Method

Once again, the distribution of projects by delivery method must be controlled for. The methodology for this follow-up test has been elaborated on already and as such will not be repeated. See the contingency table below for the first step.

Table 20: Contingency Table of Behind Schedule (excl. extensions) Projects & PDMs

Behind Schedule	*Project Delivery Method*				
	Design/Build	**CMAR**	**Single-Prime**	**Multiple-Prime**	***Row Total***
Yes	**3**	**3**	**43**	**102**	**151**
No	**9**	**5**	**14**	**10**	**38**
Column Total	**12**	**8**	**57**	**112**	**189**

No observations in this case are equal to zero, but some expected values are less than five. As such, Fischer's Exact test is performed and validated by the Chi-Square test (with Yates continuity correlation) and a Monte Carlo Simulation (bootstrapping) of 2000 replicates. Both tests return p-values of 0.000, sufficient to reject the null hypothesis of no association between the delivery method and schedule overrun at the 95% confidence level. Once again, this suggests better project performance (in terms of construction schedule) by Collaborative projects.

Table 21: Significance of Association Between Behind Schedule (excl. extensions) Projects & PDMs

Statistical Test	***p*-value**	**Significant at 95% Level**	**Significant at 90% Level**
Chi-Square	0.000	Significant	--
Fisher's Exact	0.000	Significant	--

5.1.4.2 Comparative Analysis by Delivery Types

Projects were grouped by delivery type, for the aforementioned reasons of increasing observation count and summary confirmation. Traditional delivery projects comprised 96% of projects which were behind schedule, while only 4% were Collaborative. This is yet more evidence that Collaborative projects have better performance of schedule as compared to Traditional projects.

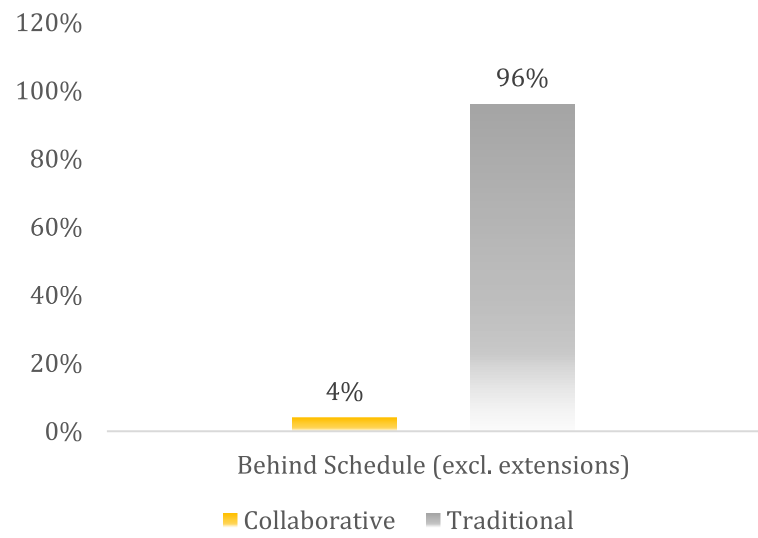


Figure 48: Distribution of Projects Behind Schedule (excl. extensions) by Delivery Type

As before, further testing is needed to validate these findings taking into account the distribution of projects by delivery type. Refer to the contingency table below.

Table 22: Contingency Table of Behind Schedule (excl. extensions) Projects & PDTs

		Project Delivery Type		
		Collaborative	**Traditional**	***Row Total***
Behind Schedule	**Yes**	**6**	**145**	**151**
	No	**14**	**24**	**38**
	Column Total	**20**	**169**	**189**

In this case, no observations are equal to zero, but some expected values are less than five (note that Table 22 shows observations only, not expected frequencies). Fischer's Exact test is thus performed with validation from the Chi-Square test (with Yates correction. The resultant *p*-values of 0.000 are sufficient to reject the null hypothesis of no association between project delivery type and late construction completion. This is further evidence of the higher performance of Collaborative projects as compared to Traditional ones.

Table 23: Significance of Association Between Behind Schedule (excl. extensions) Projects & PDTs

Statistical Test	***p*-value**	**Significant at 95% Level**	**Significant at 90% Level**
Chi-Square	0.000	Significant	--
Fisher's Exact	0.000	Significant	--

5.1.5 Change in Construction Schedule (including extensions)

This section considers construction schedule including approved time extensions. Once again, this refers to the Percent Difference between Initial Contractual Duration and Approved Extended Duration against Actual Final Duration. Pictured below, 23% of projects were completed ahead of schedule, 8% on schedule, and 69% behind schedule. Of those projects which finished behind schedule, they suffered average increase of 39%. Note that this metric was calculated in accordance with the widely acceptable definition of Schedule Factor, as listed under Chapter 1 – Definition of Terms.

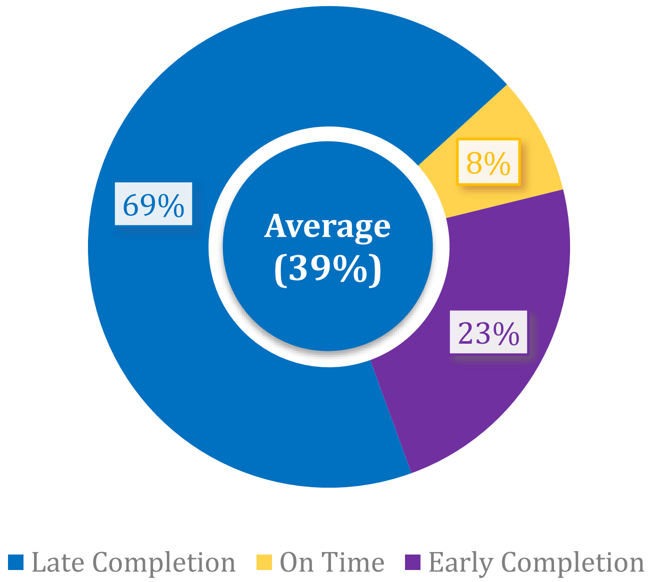


Figure 49: Construction Schedule Change (including extensions)

Of the projects which finished ahead of schedule (23%), they realized only 13%, of schedule savings. This is relatively low compared to the 39% of unapproved time extensions in the case of late projects. As with the previous iteration, the projects which were behind schedule had a higher standard deviation (0.53) compared to only 0.11 for projects completed early. As such, when a project falls behind schedule, it is much more unpredictable and may fall even further behind schedule than its 39% average.

The results are not uncommon. A study in 2015 reported that 77% of mega projects suffer schedule overruns greater than 40% (McKinsey & Co. 2015).

Table 24: Descriptive Statistics for Construction Schedule Change (incl. extensions) Per Project Status

Statistical Metric	Behind Schedule	On Schedule	Ahead of Schedule
Number of Projects	130	15	44
Total Duration (Years)	61	0	-17
Mean	1.39	1.00	0.87
Mean (%)	39%	0%	13%
Maximum	5.21	1.00	0.99
75% Percentile	1.29	1.00	0.94
Median	1.10	1.00	0.89
25% Percentile	1.09	1.00	0.84
Minimum	1.00	1.00	0.48
Standard Deviation	0.53	0.00	0.11

5.1.5.1 Comparative Analysis by Delivery Methods

When projects that experienced schedule overruns are broken out by delivery method, it is evident that MP had the highest percentage of behind-schedule projects (70%), while SP amounted to only 28%. Collaborative delivery methods performed much better, however. DB amounted for 2% of projects, while CMAR amounted to 0%.

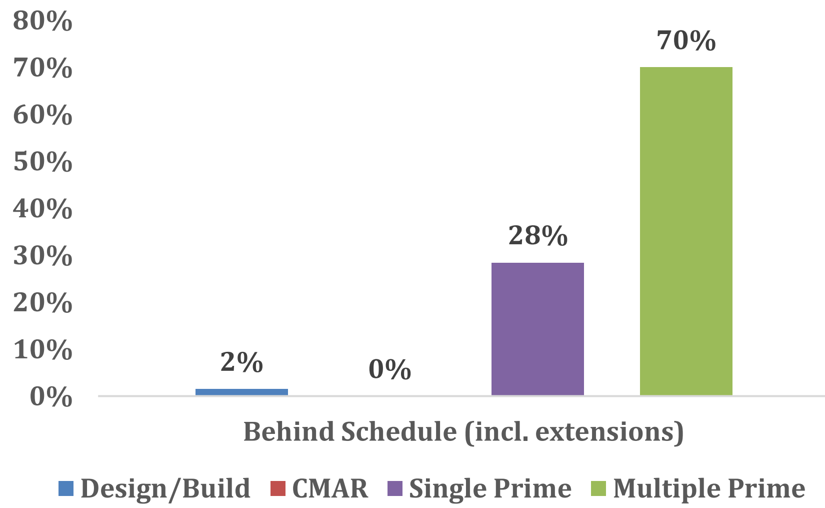


Figure 50: Distribution of Behind Schedule (incl. extensions) Projects by Delivery Method

Once again, while the results are suggestive, they must be adjusted to control for distribution of projects by delivery method. The same methodology as before will be repeated. See the contingency table below.

Table 25: Contingency Table of Behind Schedule (incl. extensions) Projects & PDMs

Behind Schedule	*Project Delivery Method*				*Row Total*
	Design/Build	**CMAR**	**Single-Prime**	**Multiple-Prime**	
Yes	**2**	**0**	**37**	**91**	**130**
No	**10**	**8**	**20**	**21**	**59**
Column Total	**12**	**8**	**57**	**112**	**189**

As there is an observation which is equal to zero, Aylmer's test can be conducted and validated by Fischer's exact test with a Monte Carlo Simulation (bootstrapping) of 2000 replicates. The resultant *p*-values were 0.000 in both cases, sufficient to reject the null hypothesis of no association between delivery method and schedule overrun at the 95% confidence level.

Table 26: Significance of Association Between Behind Schedule (incl. extensions) Projects & PDMs

Statistical Test	*p*-value	Significant at 95% Level	Significant at 90% Level
Aylmer	0.000	Significant	--
Fisher's Exact	0.000	Significant	--

5.1.5.2 Comparative Analysis by Delivery Types

When the projects were consolidated into Collaborative and Traditional, as previously, more robust statistical results can be achieved and further confirmation by summary. Traditional projects accounted for 98% of behind-schedule projects (incl. extensions), while only 2% were Collaborative. Once more, this suggests that Collaborative projects have better performance of schedule than Traditional projects.

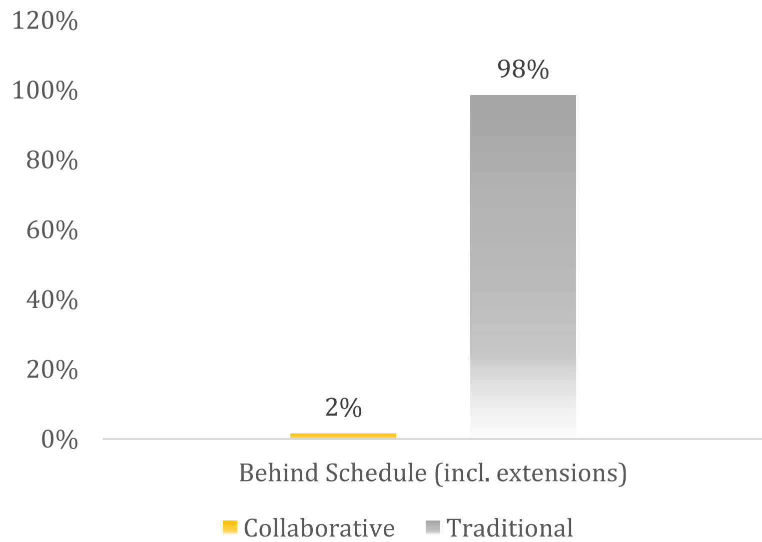


Figure 51: Distribution of Behind Schedule (incl. extensions) Projects by Delivery Type

As with all previous analyses in this section, the conclusions were validated by controlling for delivery type distribution. Per the contingency table below, projects were prepared for additional analysis.

Table 27: Contingency Table of Behind Schedule (incl. extensions) Projects & PDTs

Behind Schedule	*Project Delivery Type*: Collaborative	Traditional	*Row Total*
Yes	**2**	**128**	**130**
No	**18**	**41**	**59**
Column Total	**20**	**169**	**189**

In this case, no observation is equal to zero, and no expected values are less than five. The Chi-Square test is conducted and then validated by Fischer's Exact test, with no bootstrapping. The resultant *p*-values in both cases were 0.000, sufficient at the 95% confidence level to reject the null hypothesis of no correlation between project delivery type and late construction completion.

Table 28: Significance of Association Between Behind Schedule (incl. extensions) Projects & PDTs

Statistical Test	*p*-value	Significant at 95% Level	Significant at 90% Level
Chi-Square	0.000	Significant	--
Fisher's Exact	0.000	Significant	--

5.1.6 Section Summary

The analysis in this section has shown that 95% of projects studied experienced an increase in Base Bid, on average by 9%; 94% of studied project experienced an increase in initial design cost, on average by 30%; and 97% of the projects studied experienced an increase in initial delivery cost, on average by 10.3%.

Statistical testing revealed that Traditional delivery methods had a significantly higher number of projects with cost overruns in both construction and delivery projects at the 95% confidence level, while design cost overrun was not statistically significant.

Further, 80% of projects experienced an increase in construction duration, on average by 44% (excluding approved time extension), while including time extensions 69% increased duration, on average by 39%.

Statistical testing revealed that at the 95% confidence level that Traditional delivery had a higher number of behind-schedule projects when both excluding and including approved time extensions.

5.2 Change Orders Analysis

Change orders are changes to the original cost or schedule. The impact of change orders can be complex and cascading, depending on project-specific factors. There are a number of reasons that change orders can be initiated. In Wisconsin State, there are six such reasons, depicted in Figure 52 below.

However, no specific official definitions for each of those reasons were found – were left to interpretation by project managers. Interviews were conducted by the researcher with a number of project managers to standardize the definitions, which are listed under Chapter 1 – Definition of Terms.

The goal of this section is not to compare delivery methods by the magnitude of cost- and schedule-related changes, but rather the likelihood of occurrence of each change type for each delivery method. This provides a different perspective about the dynamics and driving forces of each method.

Figure 52: Change Order Reasons Allowed by The State of Wisconsin

Change orders once issued result in either cost adjustment (addition or subtraction) or schedule extension (addition) or some combination of both. Therefore, analysis was conducted to determine which, if any change order class was most likely to occur under each delivery method for cost-related change, schedule-related change, and a combination of both.

Quality/Value reevaluation, Program Work Reinstated, and Other were deemed too generic, ambiguous, and/or represented too small a portion of the data as compared to other reasons for change orders. Thus, these reasons were assessed statistically only for completeness. The researcher was unable to draw conclusions. Instead, the analysis was performed in full, with the goal of initiating state-wide discussion.

5.2.1 Cost-Related Change Orders

Figure 53 shows the distribution of reasons for cost-related change orders. Three reasons amounted to 76% of cost-related change orders: Scope change (38%), Unforeseen Conditions (21%) and Design Oversight (17%). Accordingly, these reasons were weighted more heavily in the analysis.

While change orders generally have negative ramifications on construction projects, as change disrupts existing operations and planning and can lead to rework or out-of-sequence work (Hanna 2004; Riley et al. 2005). Scope Changes, of the three considered, is the most impactful as it can add value, while Unforeseen Conditions and Design Oversight tend not to add value if they are justified.

Design Oversight is critical to understand, as it is generally the product of errors or poor coordination. These, when they occur, require that the contractor be compensated by the State, but are often not deducted from the design firm (typically the source of the design error).

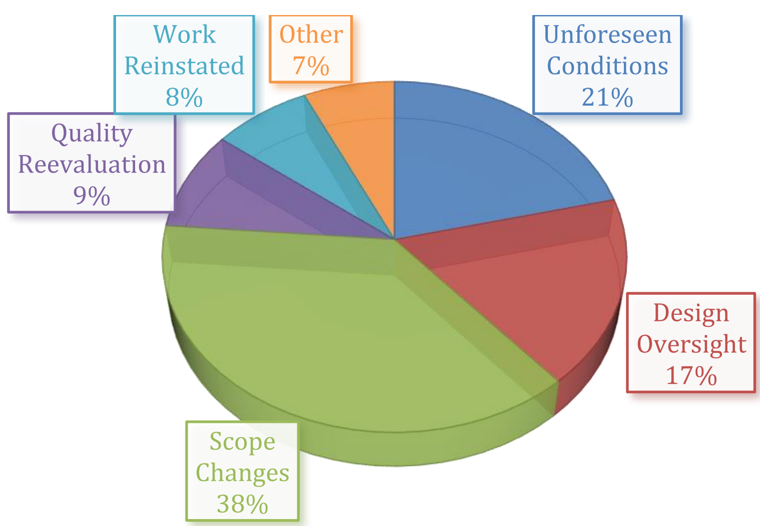
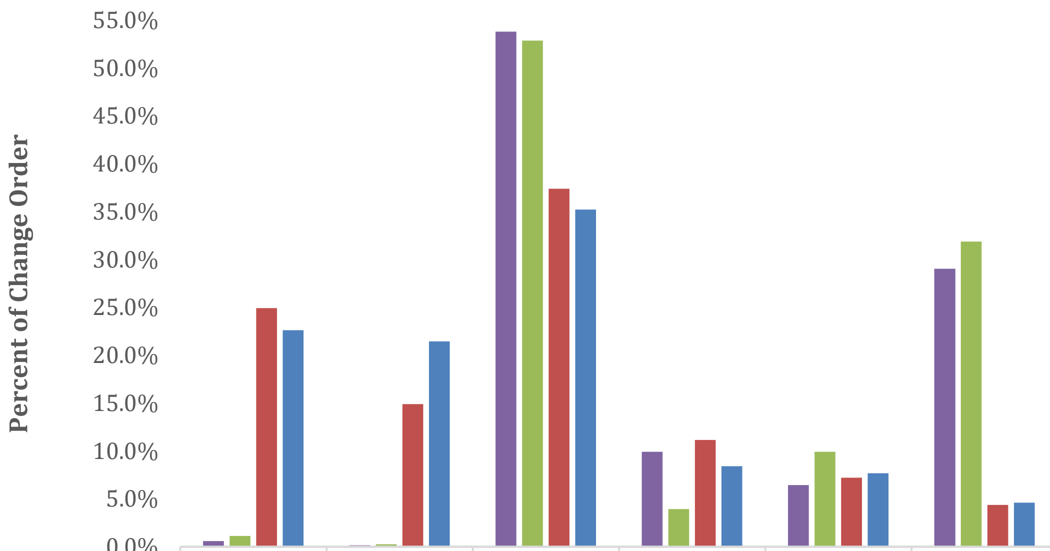


Figure 53: Distribution of Cost-Related Change Orders by Reasons

As the focus of this research is delivery methods, the first assessment made was the impact of each delivery method on overall change order volume. The figure below presents this analysis.



	Unforeseen Conditions	Design Oversight	Scope Changes	Quality Reevaluation	Work Reinstated	Other
Design/Build	0.6%	0.2%	53.8%	9.9%	6.4%	29.0%
CMAR	1.1%	0.3%	52.9%	3.9%	9.9%	31.9%
Single prime	24.9%	14.9%	37.4%	11.2%	7.2%	4.4%
Multiple Prime	22.6%	21.5%	35.2%	8.4%	7.7%	4.6%

Design/Build CMAR Single prime Multiple Prime

Figure 54: Average Likelihood of Cost-Related Change Orders by Delivery Method

As evident in Figure 54, DB and CMAR experienced much lower percentages in terms of Unforeseen Conditions (0.6 and 1.1%) when compared to SP and MP (24.9 and 22.6%). Similar to the case of Design Oversight (0.2 and 0.3% vs 14.9 and 21.5%, respectively). Conversely, Traditional delivery methods experienced less Scope Changes and Other (represented by lower average percentages).

The performance of delivery methods in terms of Quality reevaluation and Work Reinstated was largely not different among Collaborative and Traditional delivery methods. The forthcoming sections elaborate and support with statistical tests the significant differences between the delivery methods and the change order causes.

5.2.1.1 Unforeseen Conditions

5.2.1.1.1 <u>Comparative Analysis of Delivery Methods</u>

The first step in conducting this analysis was validating the assumption of normality to determine which significance tests are appropriate to use. The Q-Q plot below shows a clear departure from normality at the overall dataset, and also at the level of each delivery method.

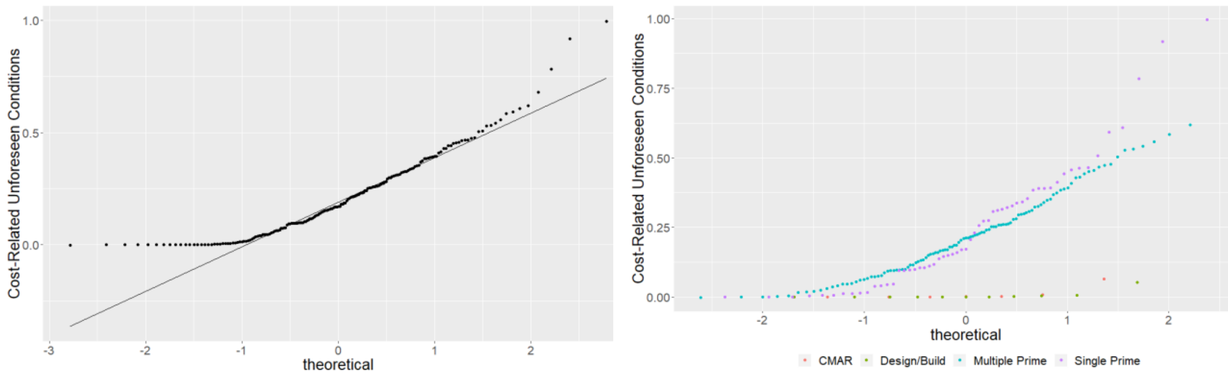


Figure 55: Q-Q Plot of Cost-Related Changes Due to Unforeseen Conditions

Confirming the visual normality of the above Q-Q plots, the table below presents the results of a Shapiro-Wilk test, with a resultant *p*-value of 0.000 for each delivery method, sufficient to reject the null hypothesis of normality at the 95% confidence level.

Table 29: Significance of Normality for Cost-Related Changes Due to Unforeseen Conditions

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.000	Significant	--
	CM at Risk	0.000	Significant	--
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

By using box-and-whisker plots, a clear difference is identified between the Collaborative delivery methods and Traditional Methods. However, within the subgroups there were no visually evident differences between DB and CMAR, while a minor difference was evident between SP and MP. Further, it is visually evident that Traditional delivery methods have higher variability than Collaborative ones from the level of compaction of the boxplot.

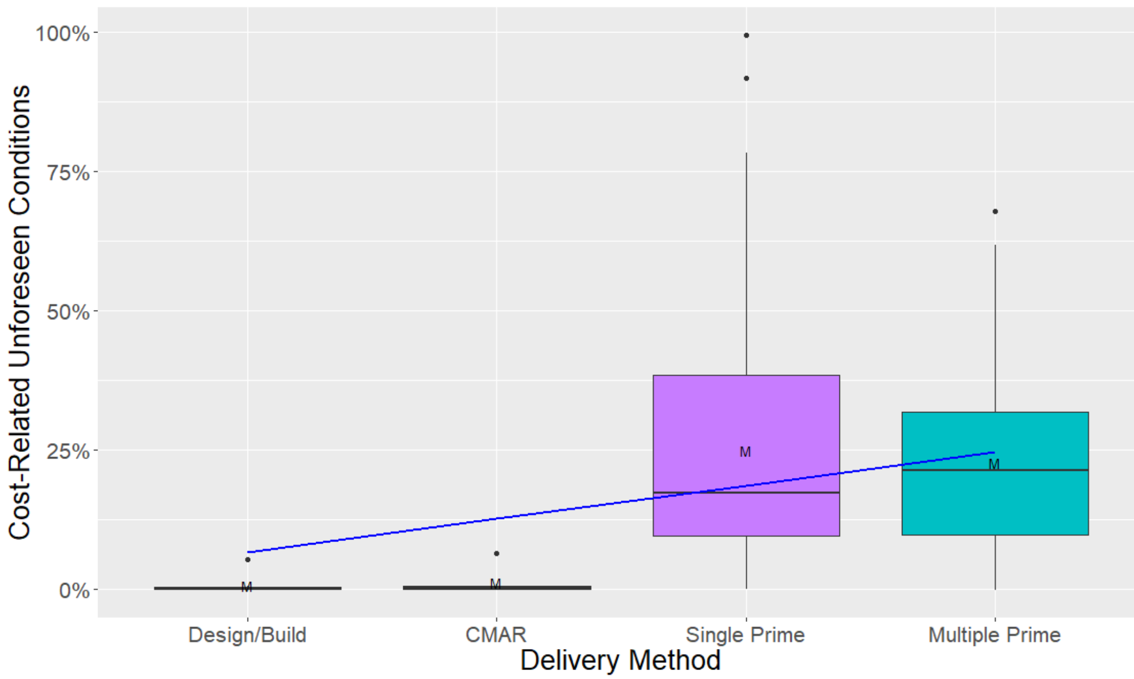


Figure 56: Boxplot of Cost-Related Changes Due to Unforeseen Conditions

This boxplot is broken out into a descriptive statistical table, below. This table suggests a difference in median performance between DB and CMAR (0 and 0.1%) and SP and MP (17.2 and 21.3% on the other hand). A further difference is highlighted between SP and MP. Variability was much the same as it has been in other tests in this research: Collaborative delivery experienced low standard deviations (1.6% and 2.4%) while Traditional was much higher (22.8 and 15.8%). Further tests were performed to assess if these deviations were influenced by outliers.

Table 30: Descriptive Statistics for Cost-Related Changes Due to Unforeseen Conditions Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	0.6%	1.1%	24.9%	22.6%
Maximum	5.3%	6.4%	99.5%	67.8%
75% Percentile	0.5%	0.9%	38.7%	32.4%
Median	0.0%	0.1%	17.2%	21.3%
25% Percentile	0.0%	0.0%	7.1%	9.7%
Minimum	0.0%	0.0%	0.0%	-0.2%
Standard Deviation	1.6%	2.4%	22.8%	15.8%

Given the previously stated normality testing, appropriate statistical tests were performed. Fligner-Killeen test resulted in a *p*-value of 0.000, sufficient at the 95% confidence level to reject the null hypothesis of equal variance among delivery methods, confirming the previous suggestion of higher variability among Traditional methods.

Additionally, Kruskal-Wallis H-Test resulted in $p = 0.000$, sufficient to reject the null hypothesis of equal population medians among delivery methods. Post-hoc, Conover-Iman test was performed to evaluate pairwise statistical differences between the individual delivery methods. Statistically significant differences were found between DB and SP/MP and CMAR and SP/MP, but not between DB and CMAR nor SP and MP.

Table 31: Significance of Variance & Distribution for Cost Changes Due to Unforeseen Conditions

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Kruskal-Wallis	Overall	0.000	Significant	--
Conover-Iman	DB vs Single	0.000	Significant	--
	DB vs Multiple	0.000	Significant	--
	CMAR vs Single	0.000	Significant	--
	CMAR vs Multiple	0.000	Significant	--
	DB vs CMAR	0.957	Not Significant	Not Significant
	Single vs Multiple	0.975	Not Significant	Not Significant

These findings confirm the previous inference about higher performance of DB compared to SP and MP, as well as CMAR relative to SP and MP in terms of cost-related change orders due to unforeseen conditions. While logic might dictate that unforeseen conditions, given their unforeseen nature, should be independent of delivery method, results suggest that Collaborative delivery methods were better able to handle unforeseen change and minimize project impact.

5.2.1.1.2 Comparative Analysis of Delivery Type

A second quantitative assessment was performed to assess the impact of each delivery type on change orders initiated due to unforeseen conditions. Therefore, a similar strategy as in the preceding section was performed to validate normality and select significance tests. The Q-Q plot shows a clear departure from normality in each delivery type.

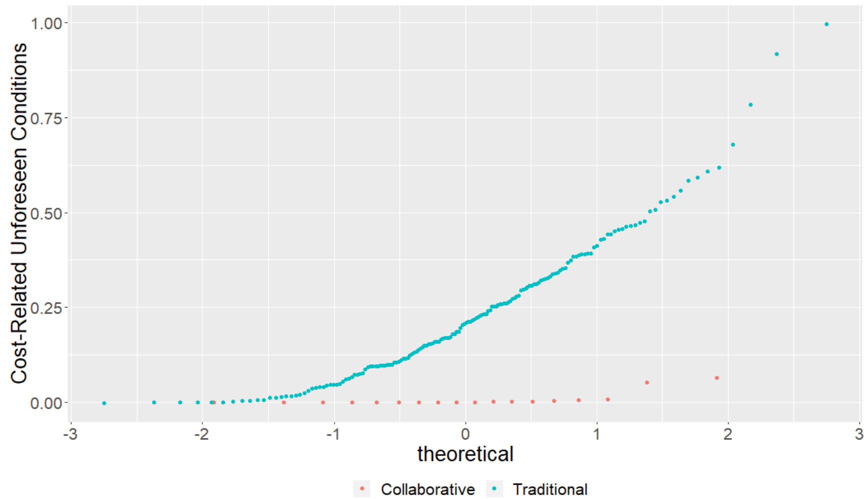


Figure 57: Q-Q Plot of Cost-Related Changes Due to Unforeseen Conditions

As previously, *p*-values of 0.000 from Shapiro-Wilk are sufficient to reject the null hypothesis of normality at the 95% confidence level.

Table 32: Significance of Normality for Cost-Related Changes Due to Unforeseen Conditions

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.000	Significant	--
	Traditional	0.000	Significant	--

A comparative boxplot identified visual differences between Collaborative and Traditional delivery types. Lower variability among Collaborative delivery was also evident.

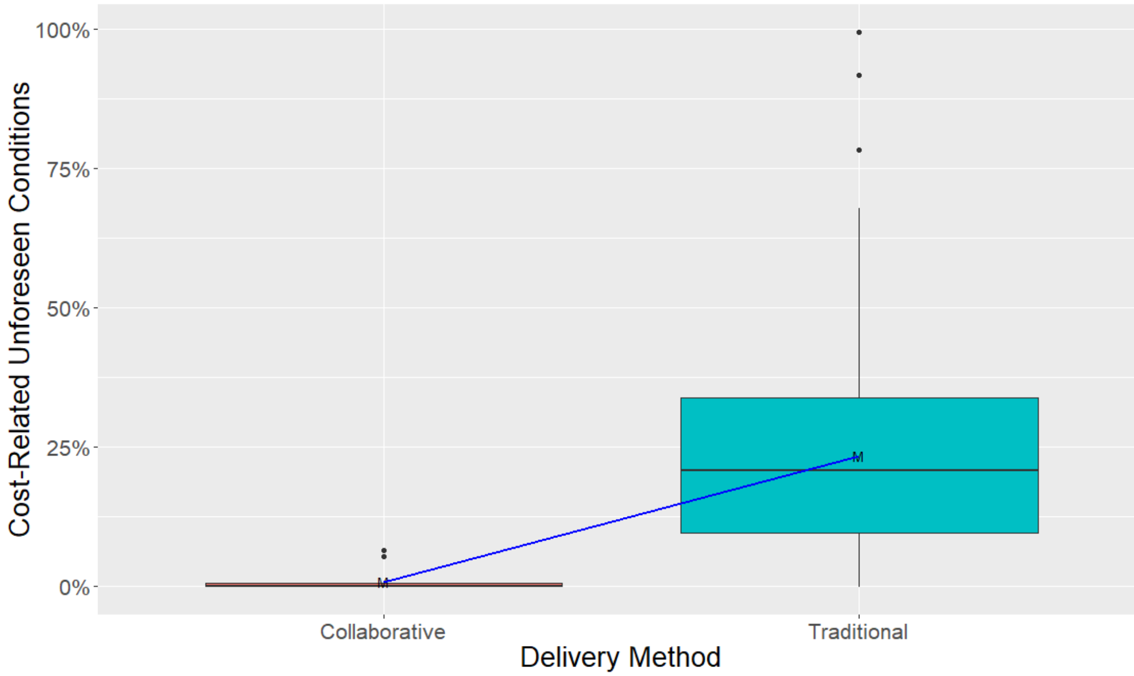


Figure 58: Boxplot of Cost-Related Changes Due to Unforeseen Conditions

The table of descriptive statistics below presents a clear difference in median and mean for Collaborative (0/0.8%) and Traditional (20.8/23.4%) respectively. Collaborative methods also exhibited lower standard deviation than Traditional (1.9% v. 18.5%).

Table 33: Descriptive Statistics for Cost-Related Changes Due to Unforeseen Conditions Per PDT

Statistical Test	Collaborative	Traditional
Mean	0.8%	23.4%
Maximum	6.4%	99.5%
75% Percentile	0.5%	33.9%
Median	0.0%	20.8%
25% Percentile	0.0%	9.5%
Minimum	0.0%	-0.2%
Standard Deviation	1.9%	18.5%

Based once again on the normality testing, statistical tests were performed. Fligner-Killeen and Mann-Whitney U tests both returned *p*-values of 0.000, sufficient to reject the appropriate null hypotheses at the 95% confidence level. Thus, it was concluded the inequal variance among delivery types, and that there were not equal medians among the two delivery types.

Table 34: Significance of Variance & Distribution for Cost Changes Due to Unforeseen Conditions

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Mann–Whitney	Overall	0.000	Significant	--

This is further evidence to the higher performance of Collaborative delivery over Traditional delivery in terms of cost-related change orders. This is arguably attributed to the higher levels of collaboration and integration in Collaborative versus Traditional methods (Pishdad-Bozorgi and de la Garza 2012; Franz et al 2017).

5.2.1.2 Design Oversight

5.2.1.2.1 <u>Comparative Analysis of Delivery Method</u>

As with the other analysis, the initial step was to determine normality. As evident from the Q-Q plots below, normality was not present. This was confirmed at the 95% confidence level by the Shapiro-Wilk test, as shown in the table below.

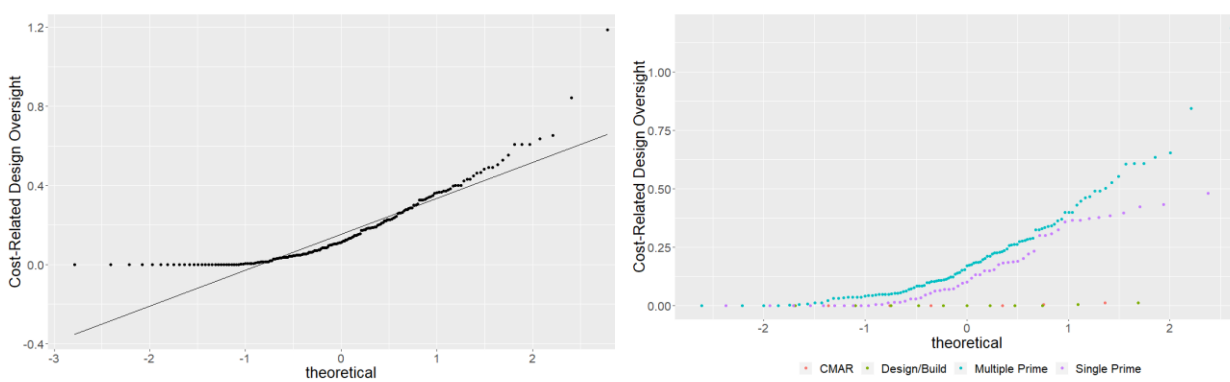


Figure 59: Q-Q Plot of Cost-Related Changes Due to Design Oversight

Table 35: Significance of Normality for Cost-Related Changes Due to Design Oversight

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.000	Significant	--
	CM at Risk	0.001	Significant	--
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

By considering the comparative boxplot below, there is a clear visual difference in performance between Collaborative and Traditional delivery methods, but there is no clear difference between DB and CMAR and only a slight one between SP and MP. Traditional delivery also exhibited higher variability.

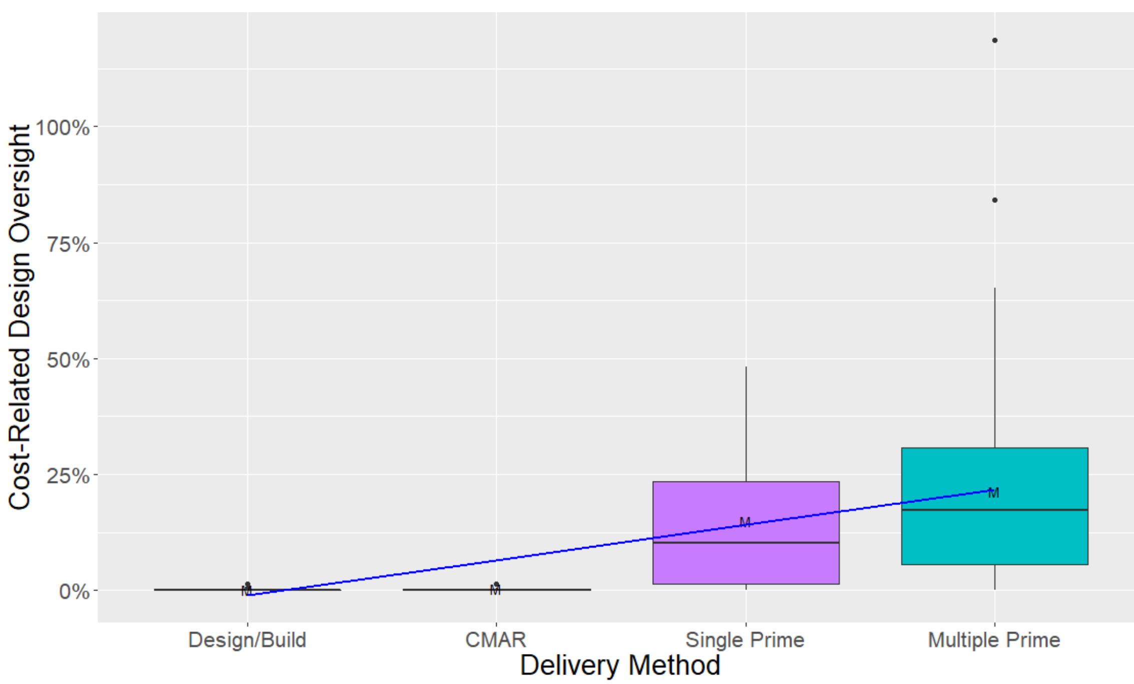


Figure 60: Boxplot of Cost-Related Changes Due to Design Oversight

The table of descriptive statistics, below, was compiled to validate the observations. A clear difference in the median of Collaborative (0/0%) and Traditional (10.2/17.2%). The medians suggest no difference between DB and CMAR, and a slight difference between SP and MP, as was previously speculated. Further testing was performed to validate this assumption. Variability was higher in SP and MP (14.6/20.2%) than DB and CMAR (0.4/0.5%). Once again, statistical tests were performed to validate these findings.

Table 36: Descriptive Statistics for Cost-Related Changes Due to Design Oversight Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	0.2%	0.3%	14.9%	21.5%
Maximum	1.3%	1.3%	48.1%	118.7%
75% Percentile	0.0%	0.5%	26.7%	32.5%
Median	0.0%	0.0%	10.2%	17.2%
25% Percentile	0.0%	0.0%	1.4%	5.4%
Minimum	0.0%	0.0%	0.0%	0.0%
Standard Deviation	0.4%	0.5%	14.6%	20.2%

Fligner-Killeen returned a *p*-value of 0.000, confirming the inequal variance among delivery methods at the 95% confidence level and suggesting a higher variability in SP and MP than in DB and CMAR.

Kruskal-Wallis further returned a *p*-value of 0.000, rejecting the assumption of equal medians at the 95% confidence level. Conover-Iman post-hoc test showed statistically significant differences between DB and SP/MP, CMAR and SP/MP and between SP and MP, with no difference shown between DB and CMAR.

Table 37: Significance of Variance & Distribution for Cost-Related Changes Due to Design Oversight

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Kruskal-Wallis	Overall	0.000	Significant	--
Conover-Iman	DB vs Single	0.000	Significant	--
	DB vs Multiple	0.000	Significant	--
	CMAR vs Single	0.002	Significant	--
	CMAR vs Multiple	0.000	Significant	--
	DB vs CMAR	0.964	Not Significant	Not Significant
	Single vs Multiple	0.037	Significant	--

The results in this section support the conclusion of a higher performance of DB and CMAR over SP and MP, this time in terms of cost-related changes due to design oversight. Ergo, DB and CMAR deliver higher design quality than DBB projects, with reduced variability. However, a set of assumptions and limitations were not covered such as: contingency, complexity, etc. Future research could perhaps be more inclusive of these variables.

5.2.1.2.2 Comparative Analysis of Delivery Type

A second quantitative assessment was performed to study the impact of each delivery type on initiated change orders due to design oversight. Once again, normality was assessed, and a lack of normality was found at the 95% confidence level, as shown in the figure and table below.

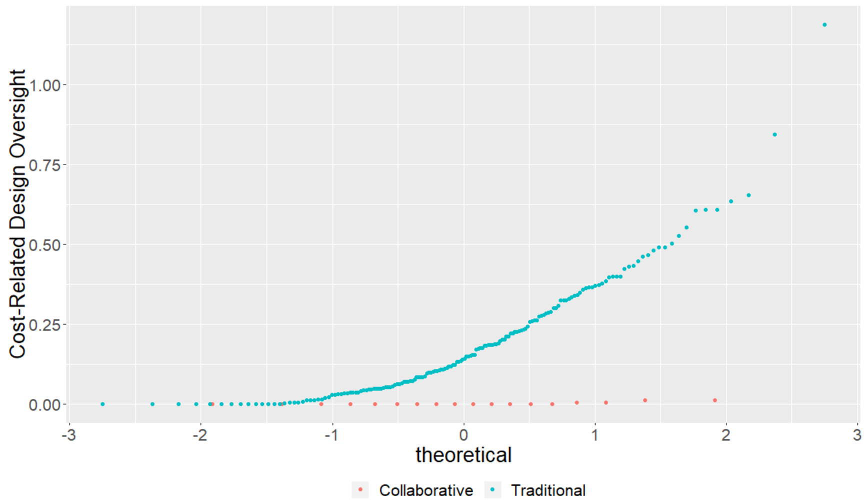


Figure 61: Q-Q Plot of Cost-Related Changes Due to Design Oversight

Table 38: Significance of Normality for Cost-Related Changes Due to Design Oversight

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.000	Significant	--
	Traditional	0.000	Significant	--

The following boxplot highlights a clear difference between Collaborative and Traditional delivery types. Higher variability was also noted in Traditional methods.

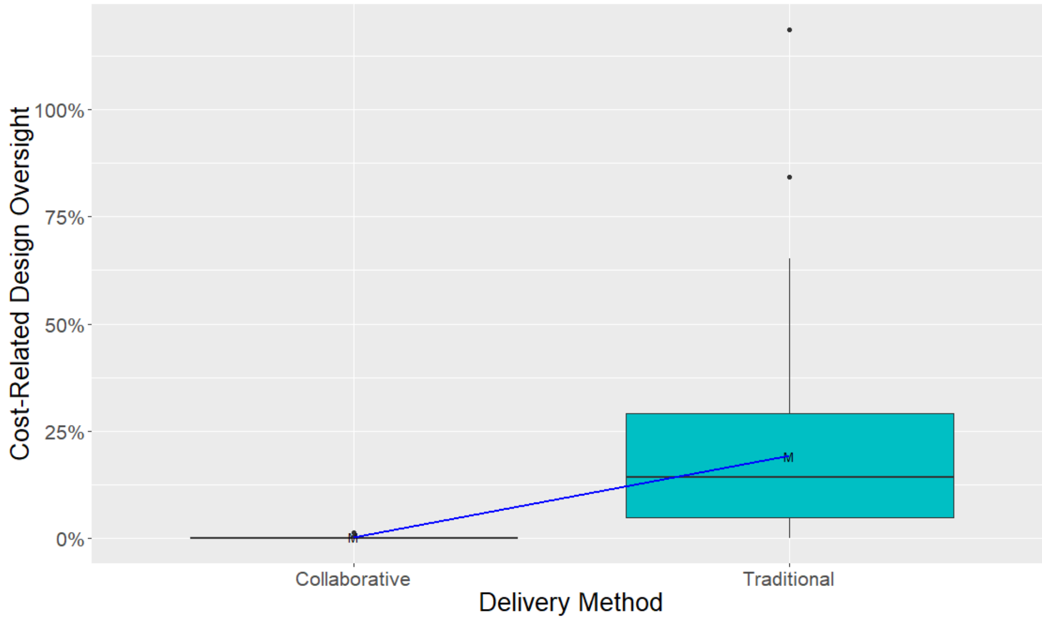


Figure 62: Boxplot of Cost-Related Changes Due to Design Oversight

Thus, the following table of descriptive statistics was created, which highlights differences between the medians of Collaborative and Traditional methods (0 and 14.1%) and their means (0.2 and 19.2%). Furthermore, Collaborative delivery methods exhibited lower variability (std. deviation) than Traditional ones.

Table 39: Descriptive Statistics for Cost-Related Changes Due to Design Oversight Per PDT

Statistical Test	Collaborative	Traditional
Mean	0.2%	19.2%
Maximum	1.3%	118.7%
75% Percentile	0.2%	29.7%
Median	0.0%	14.1%
25% Percentile	0.0%	4.7%
Minimum	0.0%	0.0%
Standard Deviation	0.4%	18.7%

To validate the preceding allegations, statistical tests were performed. Fligner-Killeen returned a *p*-value of 0.000, sufficient to reject the null hypothesis at 95% confidence level. This suggests a higher variability for Traditional delivery types than Collaborative ones.

Mann-Whitney U tests also returned a *p*-value of 0.000, sufficient to reject the null hypothesis of equal medians, once again at the 95% confidence levels.

Table 40: Significance of Variance & Distribution for Cost-Related Changes Due to Design Oversight

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Mann–Whitney	Overall	0.000	Significant	--

These findings provide additional evidence of the higher performance of Collaborative delivery over Traditional, in this case in terms of cost-related change orders due to design oversight. Largely, this can be attributed to improved design quality generated from the early involvement of the contractor during the design process, which can reduce cost-related changes. Additionally, design issues can be resolved prior to a change order needing to be issued as a portion of the contractual responsibilities of the project parties, which passes less risk along to the owners. Nevertheless, further research with specific design-related focus is needed to confirm these allegations.

5.2.1.3 Scope Changes

5.2.1.3.1 <u>Comparative Analysis of Delivery Methods</u>

The presumption of normality was once again rejected and confirmed by the Shapiro-Wilk test. All delivery methods were found significant at the 95% confidence level, with the exception of CMAR, which was significant at the 90% confidence level.

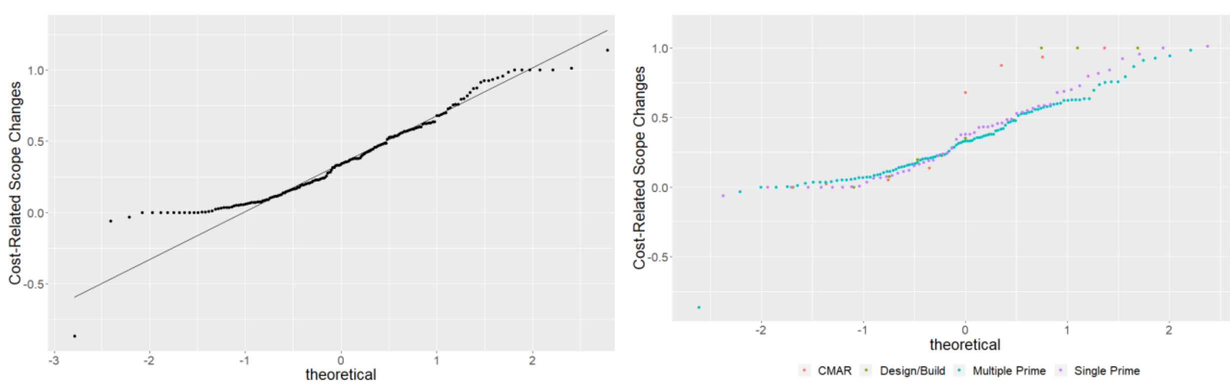


Figure 63: Q-Q Plot of Cost-Related Scope Changes

Table 41: Significance of Normality for Cost-Related Scope Changes

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.033	Significant	--
	CM at Risk	0.072	Not Significant	Significant
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

The boxplot below shows no clear difference between the delivery methods. CMAR was slightly more variable, as well as having a higher median.

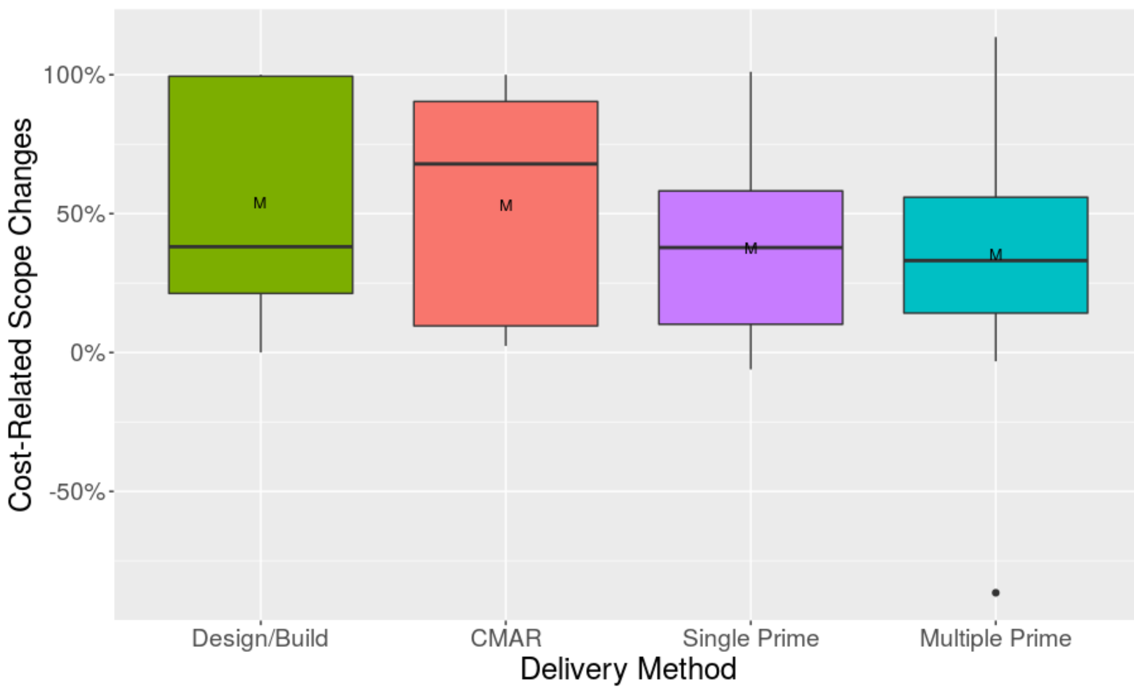


Figure 64: Boxplot of Cost-Related Scope Changes

To provide more accurate findings, the table of descriptive statistics below was drawn. These statistics suggest minor difference between DB, SP, and MP, with median values (38.1, 37.8, and 33.1%) that are similar. However, CMAR is distinct, with a median of 67.9%, which will be validated for significance with robust statistical methodologies.

SP and MP experienced similar standard deviations (30.5 and 28.6%, respectively). DB had a higher variability (40.5%) and CMAR the highest (44%).

Table 42: Descriptive Statistics for Cost-Related Scope Changes Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	53.8%	52.9%	37.4%	35.2%
Maximum	100.0%	100.0%	101.1%	113.6%
75% Percentile	100.0%	93.3%	58.4%	56.0%
Median	38.1%	67.9%	37.8%	33.1%
25% Percentile	19.9%	5.3%	9.3%	13.9%
Minimum	0.0%	2.4%	-6.1%	-86.5%
Standard Deviation	40.5%	44.0%	30.5%	28.6%

Statistical validation returned the following results. Fligner-Killeen returned a *p*-value of 0.047, which is barely sufficient to reject the null hypothesis of equal variance among delivery methods at the 95% confidence level. This suggests higher variability in DB and CMAR. Kruskal-Wallis did not return a sufficient value to reject the null hypothesis (*p*-value = 0.807) at either the 95 or 90% confidence levels. No post hoc test was thus performed.

Table 43: Significance of Variance & Distribution for Cost-Related Scope Changes

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.047	Significant	--
Kruskal-Wallis	Overall	0.807	Not Significant	Not Significant
Conover-Iman	DB vs Single	NA	NA	NA
	DB vs Multiple	NA	NA	NA
	CMAR vs Single	NA	NA	NA
	CMAR vs Multiple	NA	NA	NA
	DB vs CMAR	NA	NA	NA
	Single vs Multiple	NA	NA	NA

Rarely in this research have Traditional delivery methods outperformed Collaborative ones – DB and CMAR were found to be more prone to scope changes. Although one would expect similar results for such methods due to their nature of having a less-specified scope that is typically developed over the course of the project, further analysis found that, counterintuitively, there was insignificant difference in cost-related scope changes between Collaborative and Traditional delivery methods.

5.2.1.3.2 <u>Comparative Analysis of Delivery Type</u>

A second analysis was performed to study the impact of delivery type on initiated change orders due to scope. Once again, the assumption of normality was rejected, as shown in the Q-Q plot below and validated by Shapiro-Wilk at the 95% confidence level.

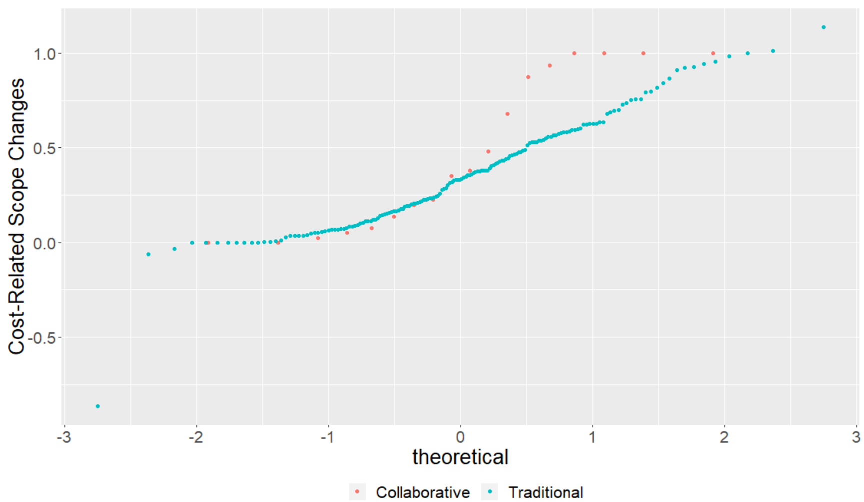


Figure 65: Q-Q Plot of Cost-Related Scope Changes

Table 44: Significance of Normality for Cost-Related Scope Changes

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.007	Significant	--
	Traditional	0.000	Significant	--

As shown in the comparative boxplot, there is no visually evident difference between Collaborative and Traditional delivery, though Collaborative delivery demonstrated higher levels of variability.

Figure 66: Boxplot of Cost-Related Scope Changes

To confirm the previous observations, the following table of descriptive statistics was drawn. Similar median values were found between Collaborative and Traditional delivery (53% and 33.9%). However, Collaborative methods exhibited larger standard deviation (40.6% vs 29.2%).

Table 45: Descriptive Statistics for Cost-Related Scope Changes Per PDT

Statistical Test	Collaborative	Traditional
Mean	53.5%	36.0%
Maximum	100.0%	113.6%
75% Percentile	99.2%	56.5%
Median	53.0%	33.9%
25% Percentile	12.3%	12.0%
Minimum	0.0%	-86.5%
Standard Deviation	40.6%	29.2%

More robust statistical tests were performed to confirm the preceding claims. Fligner-Killeen returned a *p*-value of 0.005 which is sufficient to reject the null hypothesis of equal variance among delivery types at the 95% confidence level which suggests higher variability in Collaborative delivery.

In addition, Mann-Whitney U test resulted in a *p*-value of 0.422, which is insufficient to reject the null hypothesis of equal medians at either the 95 or 90% confidence levels.

Table 46: Significance of Variance & Distribution for Cost-Related Scope Changes

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.005	Significant	--
Mann–Whitney	Overall	0.422	Not Significant	Not Significant

These findings that there is a comparable likelihood of occurrence among all the delivery methods in terms of cost-related scope changes. While Collaborative delivery methods are more prone to scope changes by nature, there was insignificant differences between the delivery types.

It is believed that the higher level of collaboration between the design and construction teams, who are able to communicate more effectively both internally and with the owner, allows the project team to refine the project scope in a smooth and clear manner. Additionally, the increased emphasis placed by the owner on experience of contracting parties can perhaps provide more proactive strategies to ensure cost compliance.

5.2.1.4 Quality/Value Reevaluation

5.2.1.4.1 <u>Comparative Analysis of Delivery Methods</u>

First, normality was assessed. The following Q-Q plots show a clear departure from normality. The subsequent table shows validation at the 95% confidence level across all delivery types.

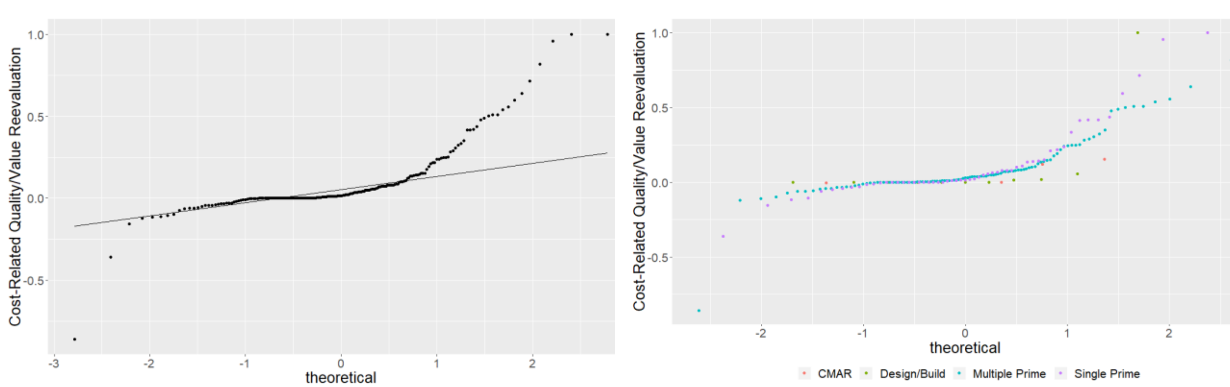


Figure 67: Q-Q Plot of Cost-Related Changes Due to Quality Reevaluation

Table 47: Significance of Normality for Cost-Related Changes Due to Quality Reevaluation

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.000	Significant	--
	CM at Risk	0.001	Significant	--
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

Per the comparative boxplot below, there is no immediately evident delivery method which is higher in performance in terms of cost related change due to quality/value reevaluation. It is, however, evident that SP has a slightly higher variability than other methods.

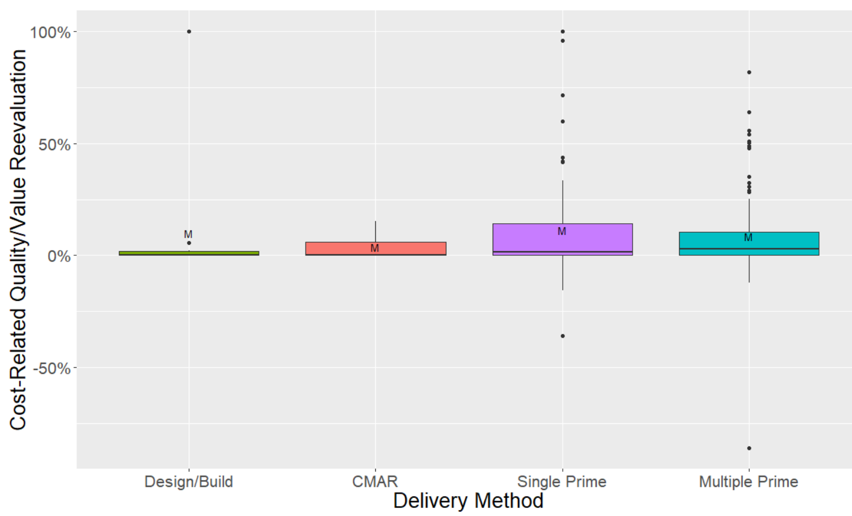


Figure 68: Boxplot of Cost-Related Changes Due to Quality Reevaluation

Per the table of descriptive statistics below, the median values of all four delivery methods are similar: 0%, 0%, 1.6% and 3%. Standard deviations were similar as well; DB, SP, and MP all had similar figures: 29.9%, 24.5%, and 19.2% respectively. CMAR had a much lower figure at 6.8%. This suggests there is no difference in performance between the different delivery methods, but more robust statistical testing is needed.

Table 48: Descriptive Statistics for Cost-Related Changes Due to Quality Reevaluation Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	9.9%	3.9%	11.2%	8.4%
Maximum	100.0%	15.4%	100.0%	81.7%
75% Percentile	2.1%	12.2%	14.2%	10.6%
Median	0.0%	0.0%	1.6%	3.0%
25% Percentile	0.0%	0.0%	0.0%	0.0%
Minimum	0.0%	-0.3%	-36.0%	-85.9%
Standard Deviation	29.9%	6.8%	24.5%	19.2%

Robust statistical tests were performed, with the following results. Fligner-Killeen returned a *p*-value of 0.032, which was sufficient to reject the null hypothesis of equal variance among delivery methods at the 95% confidence level. Kruskal-Wallis returned a *p*-value of 0.657, which failed to reject the null hypothesis of equal population of medians at the 95% confidence level. Therefore, the post-hoc test was not performed.

Table 49: Significance of Variance & Distribution for Cost Changes Due to Quality Reevaluation

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.032	Significant	--
Kruskal-Wallis	Overall	0.657	Not Significant	Not Significant
Conover-Iman	DB vs Single	NA	NA	NA
	DB vs Multiple	NA	NA	NA
	CMAR vs Single	NA	NA	NA
	CMAR vs Multiple	NA	NA	NA
	DB vs CMAR	NA	NA	NA
	Single vs Multiple	NA	NA	NA

5.2.1.4.2 Comparative Analysis of Delivery Type

A secondary analysis of delivery type was performed to assess if there was any correlation between that variable and change orders due to reevaluation. Once again, normality was tested and rejected at the 95% confidence level, per the Q-Q plot and table below.

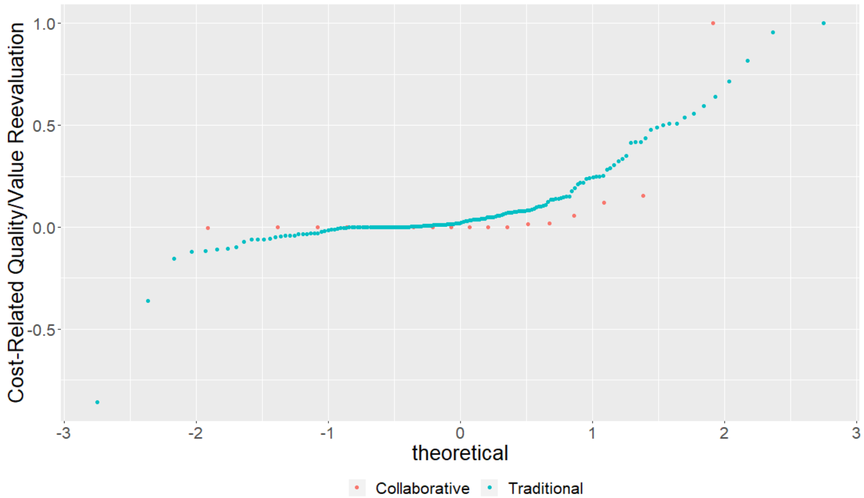


Figure 69: Q-Q Plot of Cost-Related Changes Due to Quality Reevaluation

Table 50: Significance of Normality for Cost-Related Changes Due to Quality Reevaluation

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.000	Significant	--
	Traditional	0.000	Significant	--

The comparative boxplot drawn below shows no immediately evident difference between delivery types. Traditional delivery were evidently more variable.

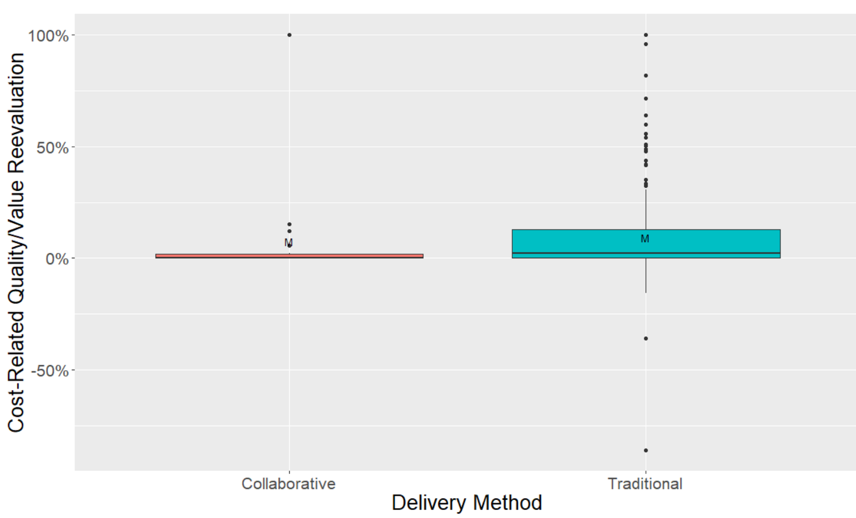


Figure 70: Boxplot of Cost-Related Changes Due to Quality Reevaluation

The following table of descriptive statistics was drawn in order to assess the suspicions brought previously in this section. The median values were similar between Collaborative and Traditional delivery (0 and 2.2%). No difference is suggested by these figures. Variability was also similar, with standard deviations of 23.5 and 21.1%. This is inconsistent with the visual observation from the boxplot. More robust testing was needed.

Table 51: Descriptive Statistics for Cost-Related Changes Due to Quality Reevaluation Per PDT

Statistical Test	Collaborative	Traditional
Mean	7.6%	9.4%
Maximum	100.0%	100.0%
75% Percentile	3.0%	13.2%
Median	0.0%	2.2%
25% Percentile	0.0%	0.0%
Minimum	-0.3%	-85.9%
Standard Deviation	23.5%	21.1%

Fligner-Killeen test returned a *p*-value of 0.003 which was sufficient to reject the null hypothesis of equal variance among delivery types at the 95% confidence level. This suggests a higher variability in Traditional delivery.

Mann-Whitney U test returned a *p*-value of 0.219, which is insufficient to reject the null hypothesis of equal medians at the 95 or 90% confidence levels.

Table 52: Significance of Variance & Distribution for Cost Changes Due to Quality Reevaluation

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.003	Significant	--
Mann–Whitney	Overall	0.219	Not Significant	Not Significant

5.2.1.5 Program Work Reinstated

5.2.1.5.1 Comparative Analysis of Delivery Methods

First, normality was assessed and rejected at the 95%, as shown in the Q-Q plots and table below.

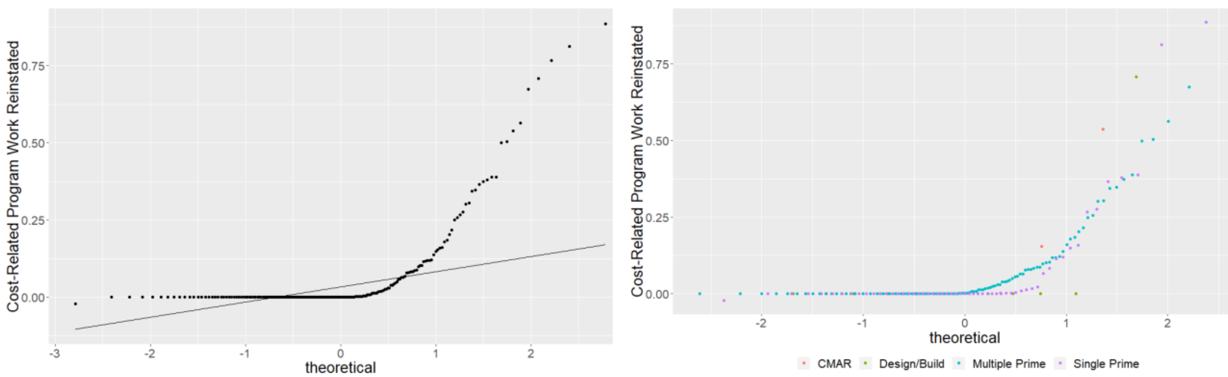


Figure 71: Q-Q Plot of Cost-Related Changes Due to Program Reinstated

Table 53: Significance of Normality for Cost-Related Changes Due to Program Reinstated

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.000	Significant	--
	CM at Risk	0.000	Significant	--
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

Per the comparative boxplot below, there is no visually evident difference between the different delivery methods. It is visually evident that there is similar variability between CMAR and DB, which is higher than either MP or SP.

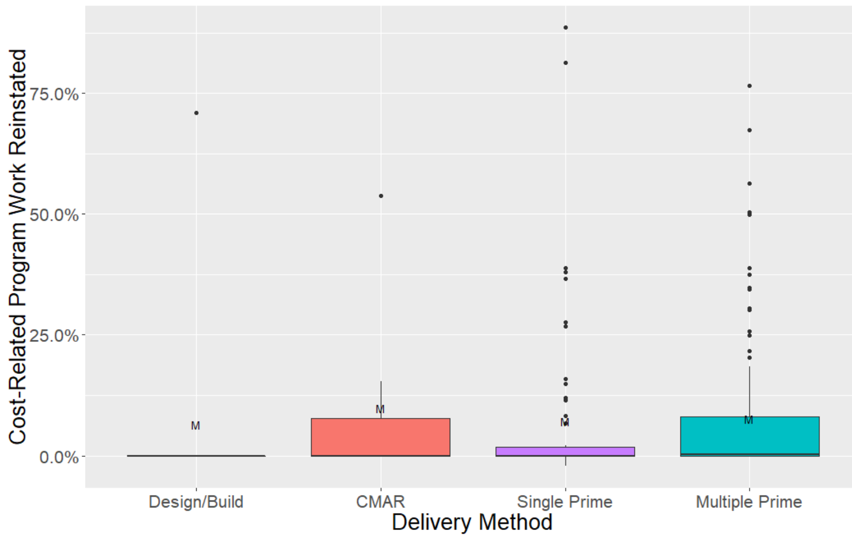


Figure 72: Boxplot of Cost-Related Changes Due to Program Reinstated

The following table of descriptive statistics was drawn to increase the robustness of the conclusions. The median values of each delivery method were: 0%, 0%, 0% and 0.2% respectively. These statistics suggest no difference between the delivery methods but will still require further statistical tests. DB and CMAR, as visually hypothesized, had similar standard deviations at 21.4 and 20.2 percent, with SP at 18% and MP at 14.8%.

Table 54: Descriptive Statistics for Cost-Related Changes Due to Program Reinstated Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	6.4%	9.9%	7.2%	7.7%
Maximum	70.9%	53.8%	88.6%	76.6%
75% Percentile	0.0%	15.4%	2.0%	8.3%
Median	0.0%	0.0%	0.0%	0.2%
25% Percentile	0.0%	0.0%	0.0%	0.0%
Minimum	0.0%	0.0%	-2.2%	0.0%
Standard Deviation	21.4%	20.2%	18.0%	14.8%

More robust tests were performed to quantitatively confirm the previous findings. Fligner-Killeen returned a *p*-value of 0.000, which was sufficient to reject the null hypothesis of equal variance at the 95% confidence level. Kruskal-Wallis, furthermore, returned a *p*-value of 0.032 – also sufficient to reject the null hypothesis of equal populations of medians among delivery methods. Accordingly, Conover-Iman post-hoc test was performed. Statistically significant difference was only found between DB and MP, but not in any other pairwise comparison.

Table 55: Significance of Variance & Distribution for Cost-Related Changes Due to Program Reinstated

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Kruskal-Wallis	Overall	0.032	Significant	--
Conover-Iman	DB vs Single	0.465	Not Significant	Not Significant
	DB vs Multiple	0.044	Significant	--
	CMAR vs Single	0.904	Not Significant	Not Significant
	CMAR vs Multiple	0.913	Not Significant	Not Significant
	DB vs CMAR	0.484	Not Significant	Not Significant
	Single vs Multiple	0.093	Not Significant	Significant

5.2.1.5.2 Comparative Analysis of Delivery Type

A second quantitative assessment was performed to study the impact of each delivery method. The Q-Q plot and table below show a clear departure from normality at the 95% confidence level.

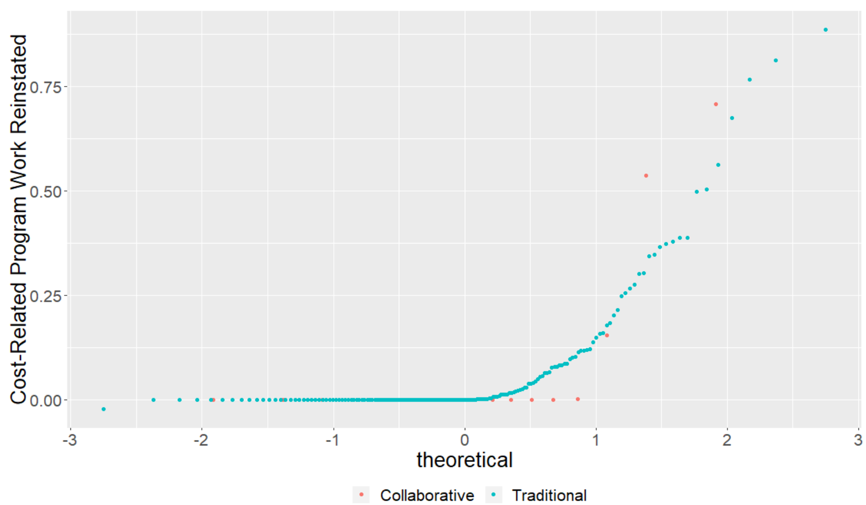


Figure 73: Q-Q Plot of Cost-Related Changes Due to Program Reinstated

Table 56: Significance of Normality for Cost-Related Changes Due to Program Reinstated

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.000	Significant	--
	Traditional	0.000	Significant	--

The comparative boxplot below shows no visually evident difference between delivery types. In terms of variability, Collaborative was much lower than Traditional.

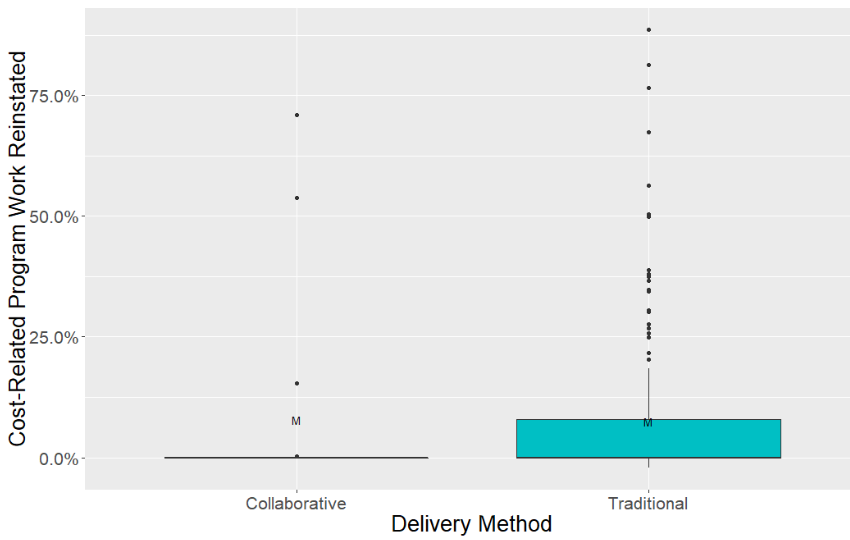


Figure 74: Boxplot of Cost-Related Changes Due to Program Reinstated

Therefore, the descriptive statistical table below was drawn to further characterize these findings. No clear difference in the median of each delivery method was found, with both methods at 0%. The mean values of the delivery methods were also highly similar, 7.8% and 7.5%, respectively. Collaborative methods also experienced similar standard deviation to Traditional (20.4 and 15.9%). Therefore, additional analysis was performed.

Table 57: Descriptive Statistics for Cost-Related Changes Due to Program Reinstated Per PDT

Statistical Test	Collaborative	Traditional
Mean	7.8%	7.5%
Maximum	70.9%	88.6%
75% Percentile	0.0%	7.9%
Median	0.0%	0.0%
25% Percentile	0.0%	0.0%
Minimum	0.0%	-2.2%
Standard Deviation	20.4%	15.9%

In this analysis, Fligner-Killeen returned a *p*-value of 0.23, which is insufficient to reject the null hypothesis of equal variance among delivery types at either the 95 or 90% confidence levels. Mann-Whitney U test returned a *p*-value of 0.096, which was sufficient to reject the null hypothesis of equal medians among delivery types at the 90% confidence level.

Table 58: Significance of Variance & Distribution for Cost Changes Due to Program Reinstated

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.230	Not Significant	Not Significant
Mann–Whitney	Overall	0.096	Not Significant	Significant

5.2.1.6 Other Reasons

5.2.1.6.1 <u>Comparative Analysis of Delivery Methods</u>

Normality was initially assessed and rejected at the 95% confidence level, as shown in the following Q-Q plots and table.

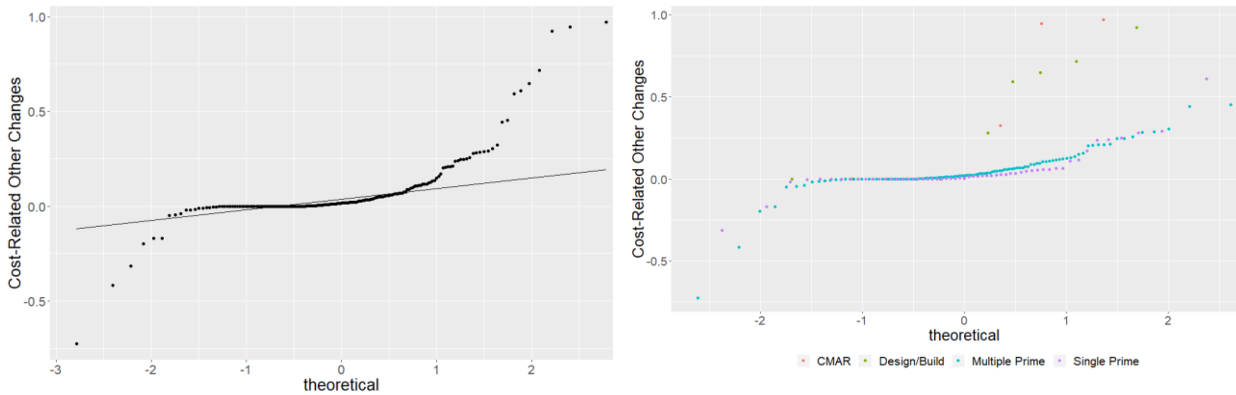


Figure 75: Q-Q Plot of Cost-Related Changes Due to Other Reasons

Table 59: Significance of Normality for Cost-Related Changes Due to Other Reasons

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.005	Significant	--
	CM at Risk	0.007	Significant	--
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

No clear difference was identified visually among delivery methods. DB and CMAR are visually more variable than SP and MP, as shown below.

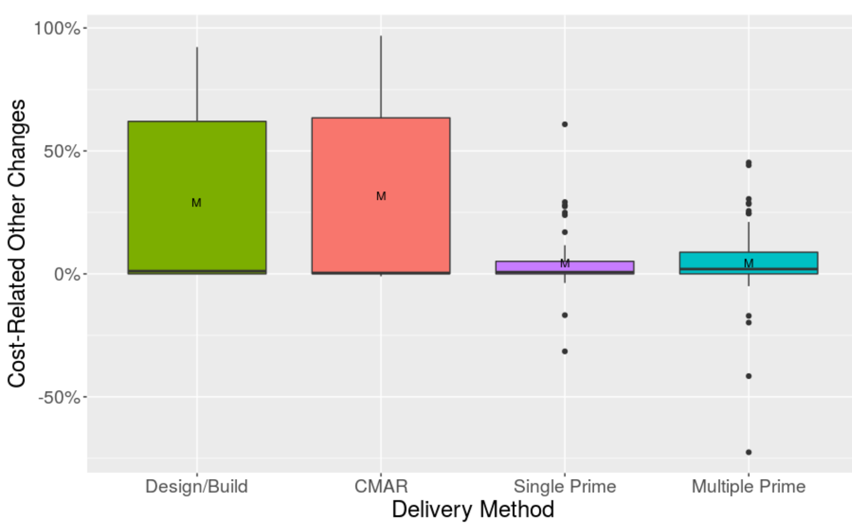


Figure 76: Boxplot of Cost-Related Changes Due to Other Reasons

The table of descriptive statistics, below, highlights the similarity of the medians, 1.1, 0.4, 0.7 and 2.0% respectively, as well as the higher standard deviations of DB and CMAR (36.1/45.2%) as compared to SP and MP (12.2/13.1%). Further tests were performed to confirm both these findings.

Table 60: Descriptive Statistics for Cost-Related Changes Due to Other Reasons Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	29.0%	31.9%	4.4%	4.6%
Maximum	92.2%	96.9%	60.8%	45.3%
75% Percentile	64.7%	94.6%	5.2%	8.9%
Median	1.1%	0.4%	0.7%	2.0%
25% Percentile	0.0%	0.0%	0.0%	0.0%
Minimum	0.0%	-1.0%	-31.5%	-72.6%
Standard Deviation	36.1%	45.2%	12.2%	13.1%

In this secondary statistical testing, Fligner-Killen returned a *p*-value of 0.020 which is sufficient to reject the null hypothesis at the 95% confidence level and suggest higher variability in DB and CMAR.

Kruskal-Wallis, however, returned a *p*-value of 0.657, which is insufficient to reject the null hypothesis of equal population medians at either the 95 or 90% confidence levels. Ergo no post hoc testing was performed.

Table 61: Significance of Variance & Distribution for Cost-Related Changes Due to Other Reasons

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.020	Significant	--
Kruskal-Wallis	Overall	0.445	Not Significant	Not Significant
Conover-Iman	DB vs Single	NA	NA	NA
	DB vs Multiple	NA	NA	NA
	CMAR vs Single	NA	NA	NA
	CMAR vs Multiple	NA	NA	NA
	DB vs CMAR	NA	NA	NA
	Single vs Multiple	NA	NA	NA

5.2.1.6.2 <u>Comparative Analysis of Delivery Type</u>

Next, a contrasted performance by delivery type was conducted. First, normality was assessed and rejected at the 95% confidence level, as shown in the Q-Q plot and table below.

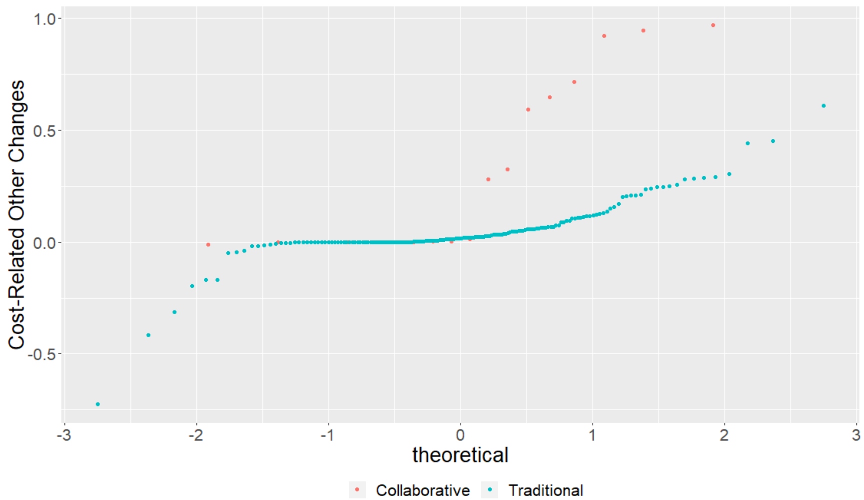


Figure 77: Q-Q Plot of Cost-Related Changes Due to Other Reasons

Table 62: Significance of Normality for Cost-Related Changes Due to Other Reasons

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.000	Significant	--
	Traditional	0.000	Significant	--

Boxplots were then generated to highlight visual differences between delivery types. Collaborative delivery methods were notably more variable than Traditional ones.

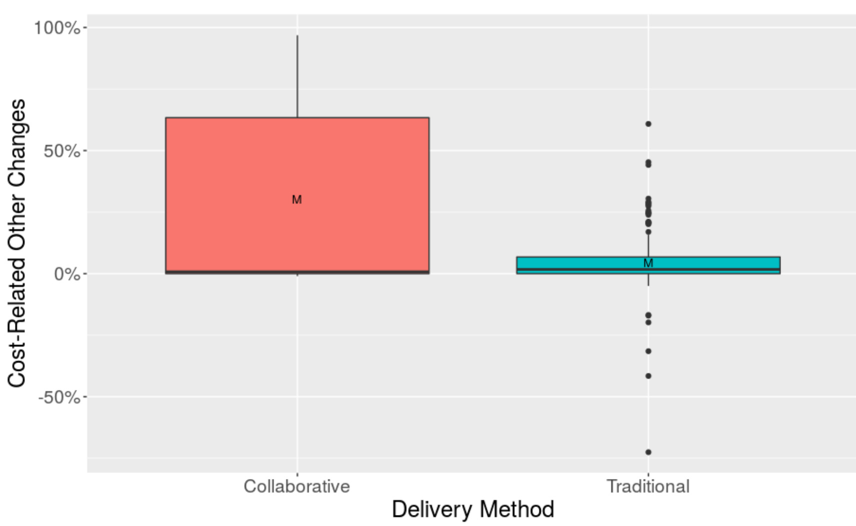


Figure 78: Boxplot of Cost-Related Changes Due to Other Reasons

To create more robust statistical observations, the following table of descriptive statistics was drawn. The median data, as evident, was similar in both delivery methods at 0.7 and 1.7%, respectively. Collaborative methods also evidently experienced a higher level of variability, with standard deviation of 38.6%, as opposed to Traditional delivery methods (12.8%).

Table 63: Descriptive Statistics for Cost-Related Changes Due to Other Reasons Per PDT

Statistical Test	Collaborative	Traditional
Mean	30.1%	4.5%
Maximum	96.9%	60.8%
75% Percentile	66.4%	6.8%
Median	0.7%	1.7%
25% Percentile	0.0%	0.0%
Minimum	-1.0%	-72.6%
Standard Deviation	38.6%	12.8%

Fligner-Killeen and Mann-Whitney U tests were performed to confirm the previous findings and claims. Fligner-Killeen returned a *p*-value of 0.000, sufficient to reject the null hypothesis of equal variance among delivery types and suggest higher variability for Collaborative delivery. Mann-Whitney U test, however, returned a *p*-value of 0.203, which is insufficient to reject the null hypothesis of equal medians among the two delivery types, with statistically insignificant differences at both the 95 and 90% confidence levels.

Table 64: Significance of Variance & Distribution for Cost-Related Changes Due to Other Reasons

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.021	Significant	--
Mann–Whitney	Overall	0.203	Not Significant	Not Significant

5.2.2 Schedule-Related Change Orders

Figure 79 shows the distribution of reasons of schedule-related change orders. The most frequent reasons were Scope Changes (37%), Unforeseen Conditions (23%) and Design Oversight (9%). These represent significant portion as in cost, showing an interesting continuity. These three most reasons represented 69% of all change orders related to schedule and will be elaborated on further in this section's analysis.

As stated earlier, Scope Changes, of the three considered, is assumed to be the most impactful as it can add value, while Unforeseen Conditions and Design Oversight tend not to add value if they are justified.

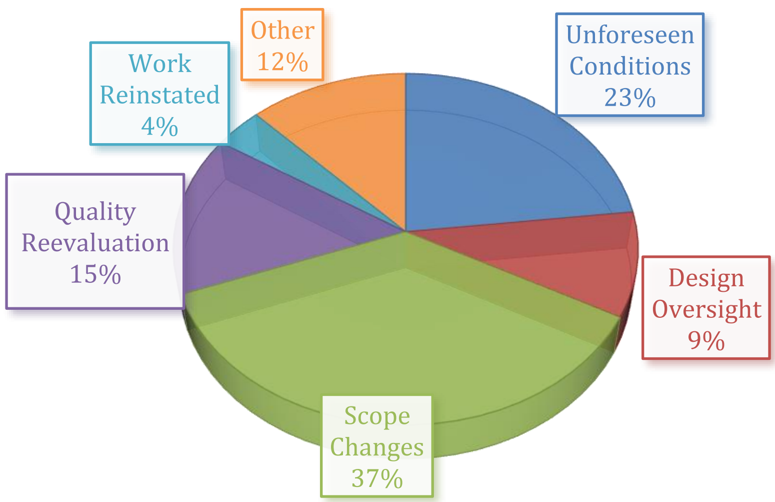
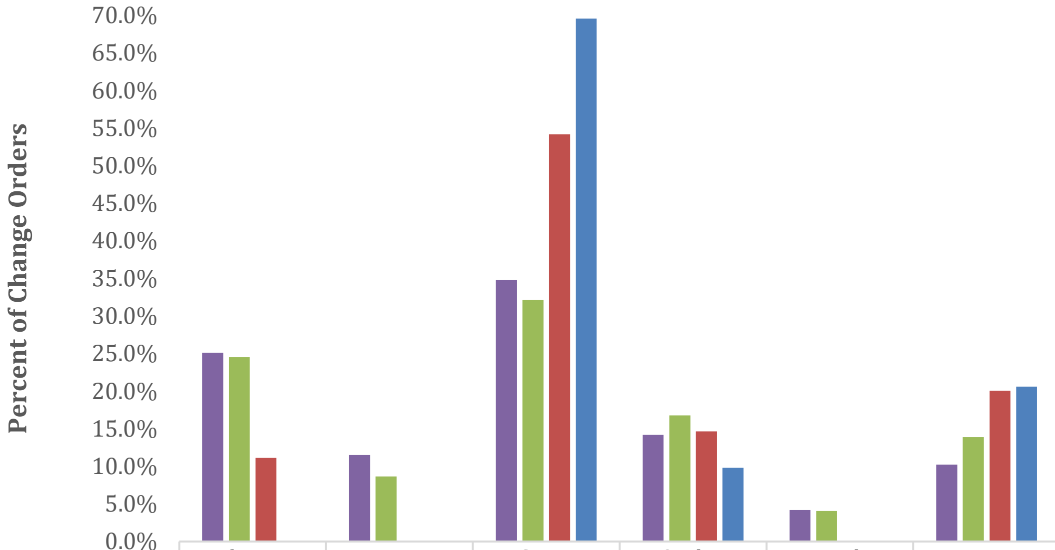


Figure 79: Distribution of Schedule-Related Change Orders by Reasons

Breaking the schedule-related change orders out by delivery method is an essential part of this research's objective. As shown in Figure 80 below, there was a much larger percentage of Unforeseen Conditions in SP and MP projects (24.5 and 25.1%), as compared to DB and CMAR (0 and 11.1%). Similar to the case of Design Oversight (11.5 and 8.6% vs 0.0 and 0.0%, respectively), which was noticed in the previous section as well. Quality Reevaluation, Work Reinstated, and Other were not as varied.



	Unforeseen Conditions	Design Oversight	Scope Changes	Quality Reevaluation	Work Reinstated	Other
Multiple Prime	25.1%	11.5%	34.8%	14.2%	4.2%	10.2%
Single prime	24.5%	8.6%	32.1%	16.8%	4.1%	13.9%
CMAR	11.1%	0.0%	54.2%	14.6%	0.0%	20.1%
Design/Build	0.0%	0.0%	69.6%	9.8%	0.0%	20.6%

Multiple Prime Single prime CMAR Design/Build

Figure 80: Average Likelihood of Schedule-Related Change Orders by Delivery Method

Each of the following sections elaborates on the impact of delivery methods and types on schedule-related change orders for each of the six reasons.

5.2.2.1 Unforeseen Conditions

5.2.2.1.1 Comparative Analysis of Delivery Methods

First, normality was assessed and dismissed at the 95% confidence level, per the Q-Q plots and table below.

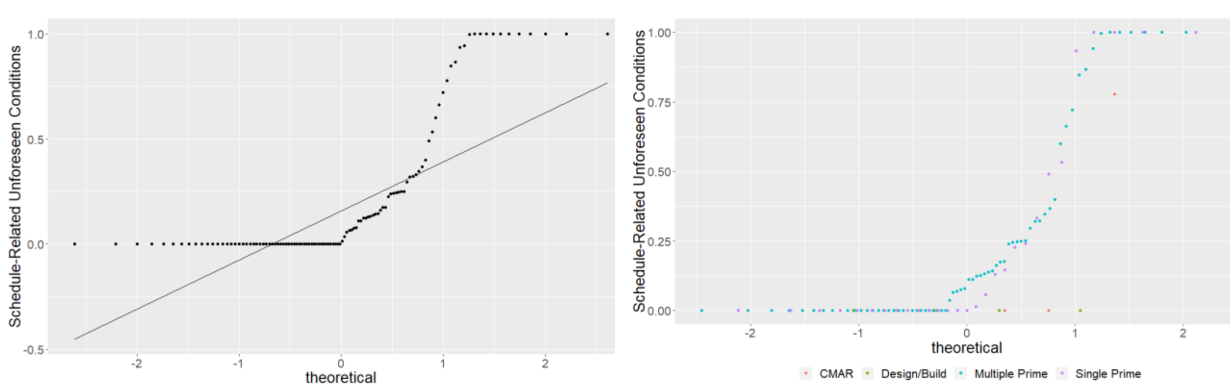


Figure 81: Q-Q Plot of Schedule-Related Changes Due to Unforeseen Conditions

Table 65: Significance of Normality for Schedule-Related Changes Due to Unforeseen Conditions

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build*	NA	NA	NA
	CM at Risk	0.000	Significant	--
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

*all the values are equivalent. Normality cannot be tested

The comparative boxplot below shows a clear difference between DB, CMAR, and SP on one hand and MP on the other. However, there were no clear differences between DB, CMAR and DBB as a group. SP and MP both exhibit higher variability than the Collaborative delivery methods.

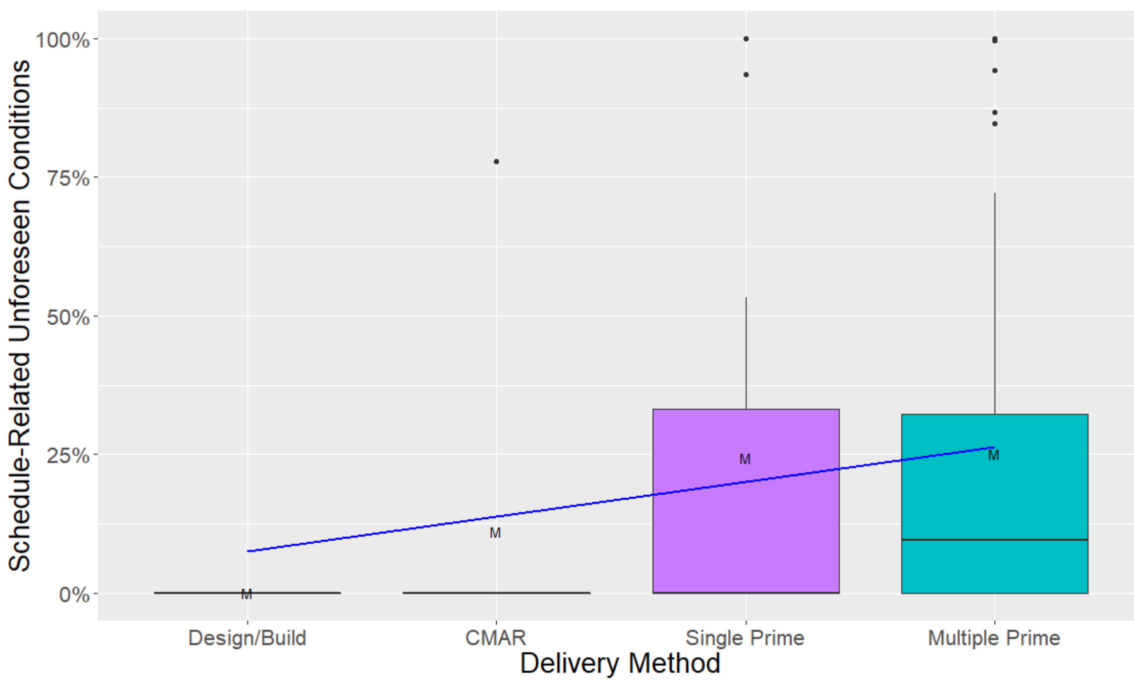


Figure 82: Boxplot of Schedule-Related Changes Due to Unforeseen Conditions

Thus, the below table of descriptive statistics was drawn, in which a clear difference in medians was identified (0, 0, 0, 9.5%). This suggests there is a considerable difference in performance of MP projects which will need to be investigated further. In terms of variability, DB had low standard deviation of 0.0%, while CMAR, SP and MP had notably higher values (29.4, 37.4 and 34.8%).

Table 66: Descriptive Statistics for Schedule-Related Changes Due to Unforeseen Conditions Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	0.0%	11.1%	24.5%	25.1%
Maximum	0.0%	77.8%	100.0%	100.0%
75% Percentile	0.0%	0.0%	41.1%	32.9%
Median	0.0%	0.0%	0.0%	9.5%
25% Percentile	0.0%	0.0%	0.0%	0.0%
Minimum	0.0%	0.0%	0.0%	0.0%
Standard Deviation	0.0%	29.4%	37.4%	34.8%

Two statistical tests were further performed to provide more robust results. Fligner-Killeen test returned a *p*-value of 0.011, sufficient to reject the null hypothesis of equal variance at the 95% confidence level. This suggests a higher variability of CMAR, SP and MP projects. Kruskal-Wallis test returned a *p*-value of 0.075, which rejects the null hypothesis at the 90% confidence level. As this test was significant, the post-hoc Conover-Iman test was also performed but did not return any significant results.

Table 67: Significance of Variance & Distribution for Schedule Changes Due to Unforeseen Conditions

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.011	Significant	--
Kruskal-Wallis	Overall	0.075	Not Significant	Significant
Conover-Iman	DB vs Single	0.224	Not Significant	Not Significant
	DB vs Multiple	0.120	Not Significant	Not Significant
	CMAR vs Single	0.372	Not Significant	Not Significant
	CMAR vs Multiple	0.184	Not Significant	Not Significant
	DB vs CMAR	0.887	Not Significant	Not Significant
	Single vs Multiple	0.871	Not Significant	Not Significant

5.2.2.1.2 Comparative Analysis of Delivery Type

As is typical of this research, a second quantitative assessment was performed which compared performance by delivery type (Collaborative v Traditional). In this analysis, normality was dismissed at the 95% confidence level, shown in the Q-Q plot and table below.

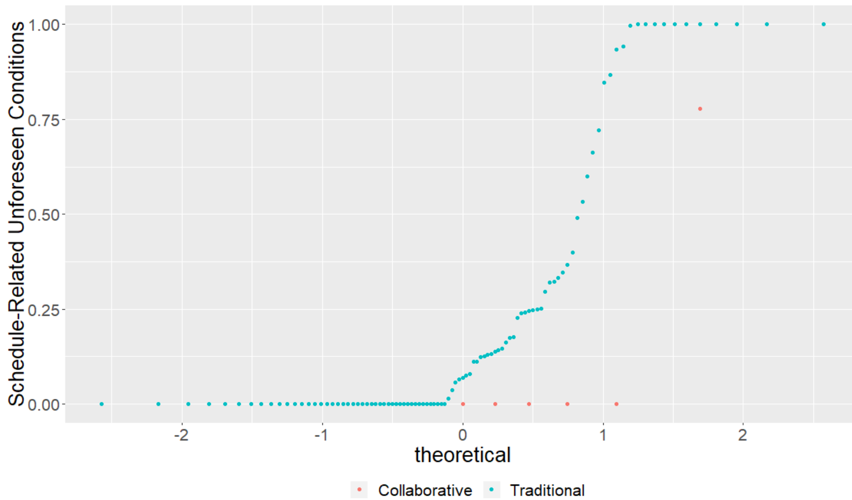


Figure 83: Q-Q Plot of Schedule-Related Changes Due to Unforeseen Conditions

Table 68: Significance of Normality for Schedule-Related Changes Due to Unforeseen Conditions

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.000	Significant	--
	Traditional	0.000	Significant	--

There is a clear differentiation in the boxplot below between Traditional and Collaborative projects. Further, there is visually lower variability in Collaborative projects as compared to Traditional ones.

Figure 84: Boxplot of Schedule-Related Changes Due to Unforeseen Conditions

The table of descriptive statistics below was drawn to provide perspective. There is an evident difference between the medians of Collaborative and Traditional methods (0 and 6.9%), as well as their means (7.1 and 25%). Collaborative methods had slightly lower standard deviation of 23.5% compared to that Traditional ones of 35.5%. As with all conclusions, statistical validation was used.

Table 69: Descriptive Statistics for Schedule-Related Changes Due to Unforeseen Conditions Per PDT

Statistical Test	Collaborative	Traditional
Mean	7.1%	25.0%
Maximum	77.8%	100.0%
75% Percentile	0.0%	33.1%
Median	0.0%	6.9%
25% Percentile	0.0%	0.0%
Minimum	0.0%	0.0%
Standard Deviation	23.5%	35.5%

To quantitatively confirm our previous findings, Fligner-Killeen test was performed, returning a *p*-value of 0.000, sufficient to reject the null hypothesis of equal variance among delivery types at the 95% confidence level. This suggests higher variability in Traditional delivery as compared to Collaborative ones. Mann-Whitney U test returned a *p*-value of 0.012, also sufficient to reject the null hypothesis of equal medians among the two delivery types.

Table 70: Significance of Variance & Distribution for Schedule Changes Due to Unforeseen Conditions

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Mann–Whitney	Overall	0.012	Significant	--

This provides statistical evidence of the higher performance of Collaborative delivery methods in terms of schedule-related changes due to unforeseen conditions. As previously iterated, the higher levels of collaboration and integration in Collaborative versus Traditional methods (Pishdad-Bozorgi and de la Garza 2012; Franz et al 2017) is likely the reason behind such differentiation.

5.2.2.2 Design Oversight

5.2.2.2.1 <u>Comparative Analysis of Delivery Methods</u>

As is typical, the assumption of normality was first tested, and subsequently rejected at the 95% confidence level as shown in the Q-Q plots and table below.

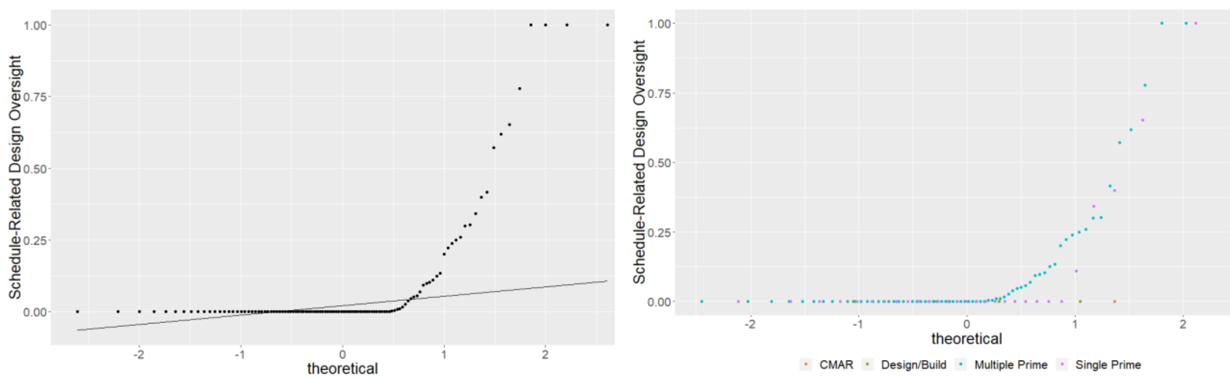


Figure 85: Q-Q Plot of Schedule-Related Changes Due to Design Oversight

Table 71: Significance of Normality for Schedule-Related Changes Due to Design Oversight

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build*	NA	NA	NA
	CM at Risk*	NA	NA	NA
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

*all the values are equivalent. Normality cannot be tested

The comparative boxplot below shows hardly any between Collaborative and Traditional deliveries. MP was also evidently more variable.

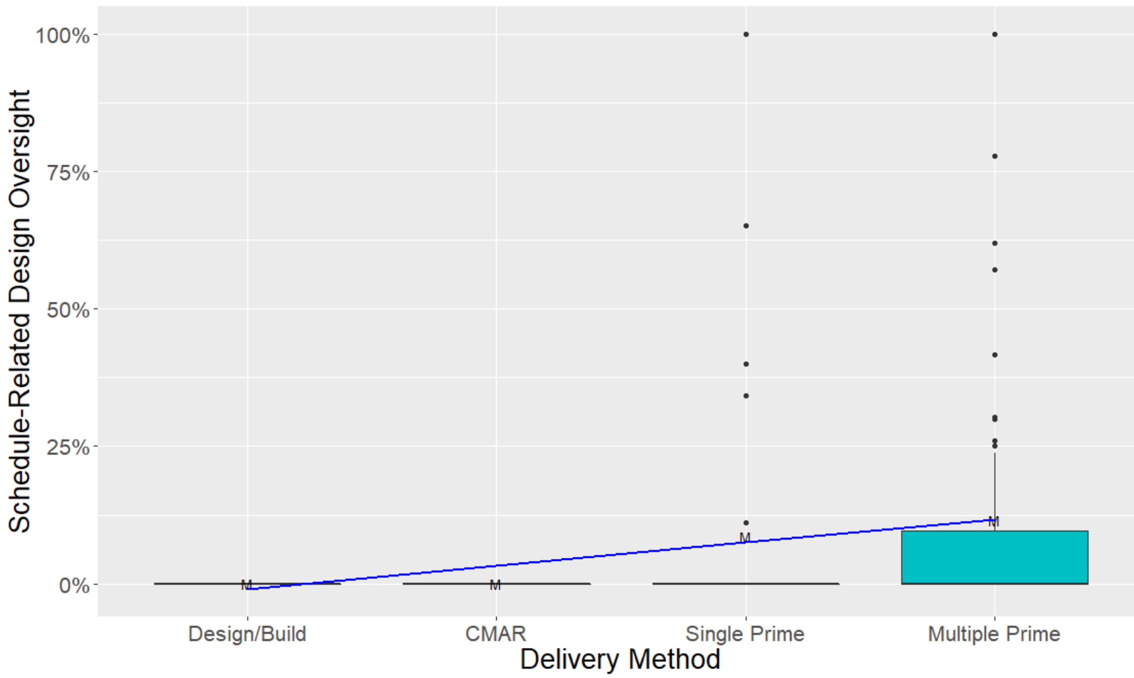


Figure 86: Boxplot of Schedule-Related Changes Due to Design Oversight

The following table of descriptive statistics was also drawn, which highlights no clear difference in the medians of Collaborative and Traditional projects (0, 0, 0, 0%). This suggests no difference between any of the delivery methods. Nevertheless, the mean values clarify a distinction between DB and CMAR (0, 0%), and SP and MP (8.6, 11.5%). DB and CMAR experienced low standard deviations (0, 0%) which is indicative of lower variability. SP and MP were much higher (23.1, 24.4%).

Table 72: Descriptive Statistics for Schedule-Related Changes Due to Design Oversight Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	0.0%	0.0%	8.6%	11.5%
Maximum	0.0%	0.0%	100.0%	100.0%
75% Percentile	0.0%	0.0%	0.0%	10.0%
Median	0.0%	0.0%	0.0%	0.0%
25% Percentile	0.0%	0.0%	0.0%	0.0%
Minimum	0.0%	0.0%	0.0%	0.0%
Standard Deviation	0.0%	0.0%	23.1%	24.4%

Fligner-Killeen test returned a *p*-value of 0.056, barely insufficient to reject the null hypothesis of equal variance among the delivery methods at the 95% confidence level. Kruskal-Wallis test returned a *p*-value of 0.021, sufficient to reject the null hypothesis at the 95% confidence level. Conover-Iman post-hoc test was performed, and returned highlighted statistically significant differences only between CMAR and MP, and SP and MP at the 90% confidence level.

Table 73: Significance of Variance & Distribution for Schedule-Related Changes Due to Design Oversight

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.056	Not Significant	Significant
Kruskal-Wallis	Overall	0.021	Significant	--
Conover-Iman	DB vs Single	0.764	Not Significant	Not Significant
	DB vs Multiple	0.223	Not Significant	Not Significant
	CMAR vs Single	0.642	Not Significant	Not Significant
	CMAR vs Multiple	0.072	Not Significant	Significant
	DB vs CMAR	0.984	Not Significant	Not Significant
	Single vs Multiple	0.095	Not Significant	Significant

5.2.2.2.2 Comparative Analysis of Delivery Type

A secondary quantitative assessment was performed to study the impact of each delivery type on change orders due to design oversight. The following Q-Q plot and table show a departure from normality significant at the 95% confidence level.

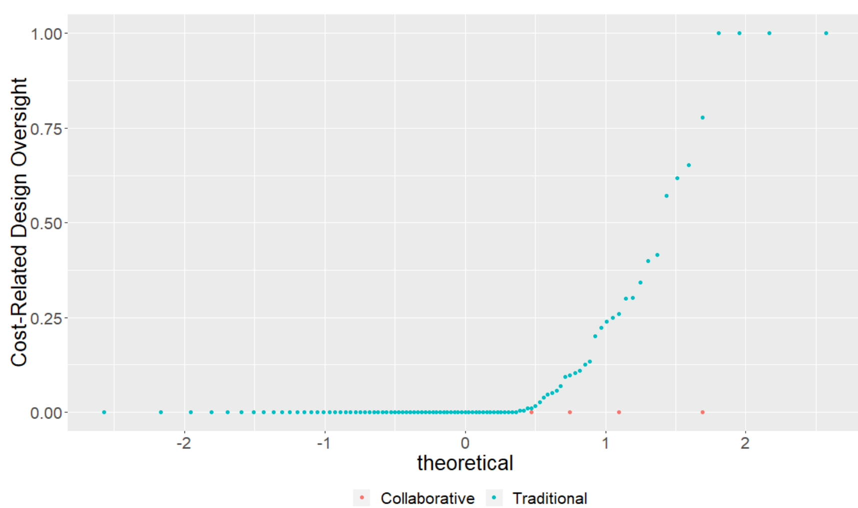


Figure 87: Q-Q Plot of Schedule-Related Changes Due to Design Oversight

Table 74: Significance of Normality for Schedule-Related Changes Due to Design Oversight

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative*	NA	NA	NA
	Traditional	0.000	Significant	--

*all the values are equivalent. Normality cannot be tested

The boxplot drawn below illustrates a difference in performance between Collaborative and Traditional delivery types. Collaborative delivery also was less variable.

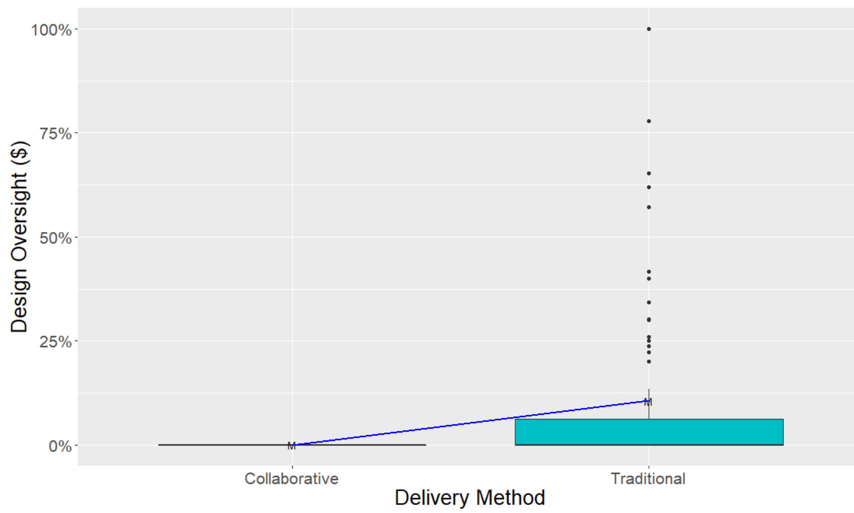


Figure 88: Boxplot of Schedule-Related Changes Due to Design Oversight

To provide further context to the analysis presented thus far, the table of descriptive statistics below was drawn. No clear difference in medians is present, as the Collaborative and Traditional delivery methods had medians of 0%. However, the mean values of the delivery methods clarify a distinction between both types at 0 and 10.7% respectively. Collaborative methods also experienced lower standard deviation (0%) as compared to Traditional (23.9%).

Table 75: Descriptive Statistics for Schedule-Related Changes Due to Design Oversight Per PDT

Statistical Test	Collaborative	Traditional
Mean	0.0%	10.7%
Maximum	0.0%	100.0%
75% Percentile	0.0%	7.0%
Median	0.0%	0.0%
25% Percentile	0.0%	0.0%
Minimum	0.0%	0.0%
Standard Deviation	0.0%	23.9%

Fligner-Killeen test returned a *p*-value of 0.030, which is sufficient to reject the null hypothesis of equal variance among delivery types at the 95% confidence level. This suggests a higher variability for Traditional delivery types. Mann-Whitney U test resulted in a *p*-value of 0.020, which rejects the null hypothesis of equal medians among the two delivery types at the 95% confidence level.

Table 76: Significance of Variance & Distribution for Schedule-Related Changes Due to Design Oversight

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.030	Significant	--
Mann–Whitney	Overall	0.020	Significant	--

These findings provide evidence of the higher performance of Collaborative delivery in terms of schedule-related changes due to design oversight. Largely, increased quality of design is responsible.

5.2.2.3 Scope Changes

5.2.2.3.1 <u>Comparative Analysis of Delivery Methods</u>

First, normality was assessed and rejected at the 95% confidence level for all delivery types except CMAR, which was rejected at the 90% confidence level, as shown in the Q-Q plots and table below.

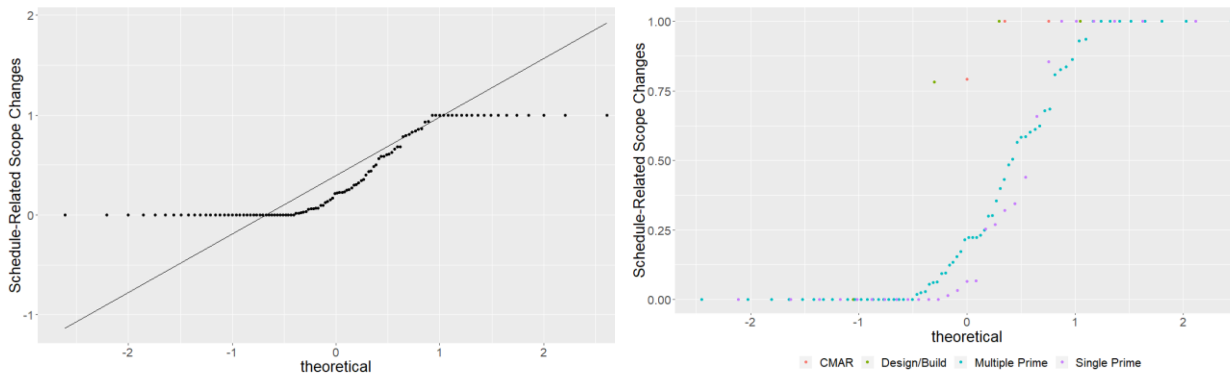


Figure 89: Q-Q Plot of Schedule-Related Scope Changes

Table 77: Significance of Normality for Schedule-Related Scope Changes

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.062	Not Significant	Significant
	CM at Risk	0.007	Significant	--
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

The comparative box-plot below shows that DB has much lower variance than any other delivery method, with DB and CMAR having higher medians that SP or MP.

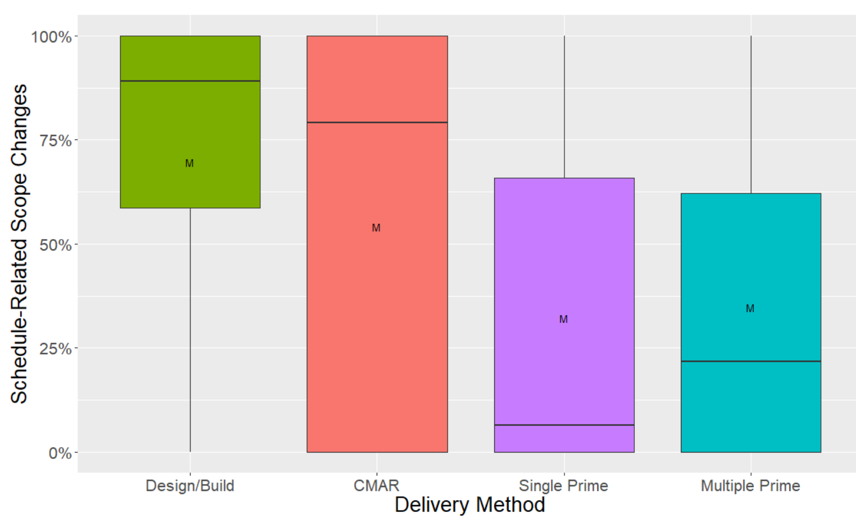


Figure 90: Boxplot of Schedule-Related Scope Changes

To provide more accurate findings, the table of descriptive statistics below was drawn. These statistics suggest obvious difference between DB/CMAR, and SP/MP, with median values (89.1, 79.2, 6.5, and 21.8%), which will be validated for significance with robust statistical methodologies.

DB, CMAR, SP and MP experienced similar standard deviations (47.5, 51.2, 41.1, and 37.5%, respectively).

Table 78: Descriptive Statistics for Schedule-Related Scope Changes Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	69.6%	54.2%	32.1%	34.7%
Maximum	100.0%	100.0%	100.0%	100.0%
75% Percentile	100.0%	100.0%	75.6%	63.9%
Median	89.1%	79.2%	6.5%	21.8%
25% Percentile	19.6%	0.0%	0.0%	0.0%
Minimum	0.0%	0.0%	0.0%	0.0%
Standard Deviation	47.5%	51.2%	41.1%	37.5%

Statistical validation returned the following results. Fligner-Killeen returned a *p*-value of 0.823, which is insufficient to reject the null hypothesis of equal variance among delivery methods at either the 95 or 90% confidence levels. This suggests no difference in variability between the four delivery methods. Kruskal-Wallis did not return a sufficient value to reject the null hypothesis (*p*-value = 0.423) at either the 95 or 90% confidence levels. No post hoc test was thus performed.

Table 79: Significance of Variance & Distribution for Schedule-Related Scope Changes

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.823	Significant	--
Kruskal-Wallis	Overall	0.472	Significant	--
Conover-Iman	DB vs Single	NA	NA	NA
	DB vs Multiple	NA	NA	NA
	CMAR vs Single	NA	NA	NA
	CMAR vs Multiple	NA	NA	NA
	DB vs CMAR	NA	NA	NA
	Single vs Multiple	NA	NA	NA

5.2.2.3.2 <u>Comparative Analysis of Delivery Type</u>

A second analysis was performed to study the impact of delivery type on initiated change orders due to scope. Once again, the assumption of normality was rejected, as shown in the Q-Q plot below and validated by Shapiro-Wilk at the 95% confidence level.

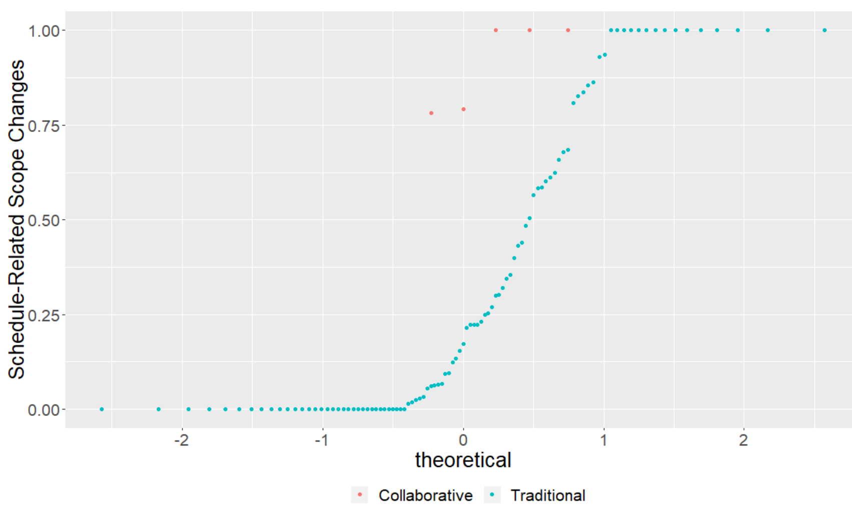


Figure 91: Q-Q Plot of Schedule-Related Scope Changes

Table 80: Significance of Normality for Schedule-Related Scope Changes

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.000	Significant	--
	Traditional	0.000	Significant	--

As shown in the comparative boxplot, there is slight difference between Collaborative and Traditional delivery, though Collaborative delivery demonstrated higher levels of variability.

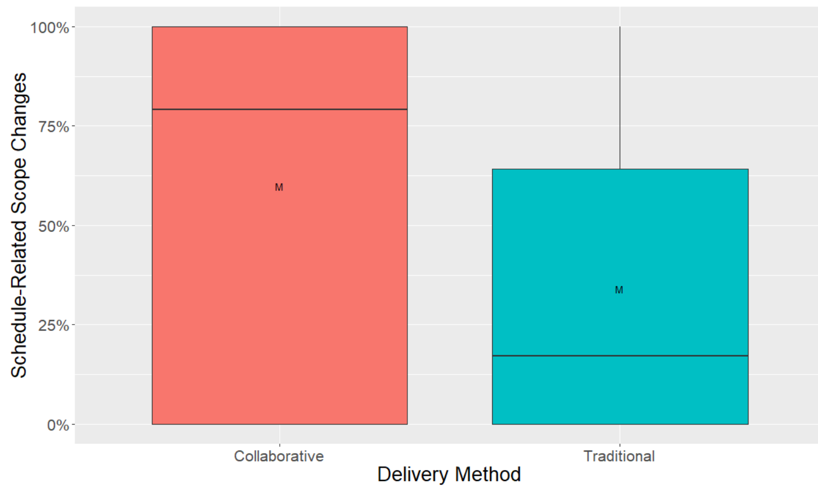


Figure 92: Boxplot of Schedule-Related Scope Changes

To confirm the previous observations, the following table of descriptive statistics was drawn. Collaborative experienced much higher median value compared to Traditional delivery (79.2 and 17.1%). Collaborative methods exhibited slightly higher standard deviation (48.1 vs 38.4%).

Table 81: Descriptive Statistics for Schedule-Related Scope Changes Per PDT

Statistical Test	Collaborative	Traditional
Mean	59.8%	34.0%
Maximum	100.0%	100.0%
75% Percentile	100.0%	65.8%
Median	79.2%	17.1%
25% Percentile	0.0%	0.0%
Minimum	0.0%	0.0%
Standard Deviation	48.1%	38.4%

More robust statistical tests were performed to confirm the preceding claims. Fligner-Killeen returned a *p*-value of 0.396 which fails to reject the null hypothesis of equal variance among delivery types at both the 95 and 90% confidence levels.

In addition, Mann-Whitney U test resulted in a *p*-value of 0.170, which is also insufficient to reject the null hypothesis of equal medians at either the 95 or 90% confidence levels.

Table 82: Significance of Variance & Distribution for Schedule-Related Scope Changes

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.396	Not Significant	Not Significant
Mann–Whitney	Overall	0.170	Not Significant	Not Significant

These findings confirm that there is a comparable likelihood of occurrence among all the delivery methods in terms of schedule-related change orders due to scope. While Collaborative delivery methods are more prone to scope changes by nature, there was insignificant differences between the delivery types.

As stated, the higher level of collaboration between the design and construction teams likely allows the project team to refine the project scope in a smooth and clear manner. Additionally, the increased emphasis placed by the owner on experience of contracting parties can perhaps provide more proactive strategies to ensure schedule compliance. Further investigation is needed to validate these claims and provide a better understanding for the results.

5.2.2.4 Quality/Value Reevaluation

5.2.2.4.1 Comparative Analysis of Delivery Methods

First, normality was assessed. The following Q-Q plots show a clear departure from normality. The subsequent table shows validation at the 95% confidence level across all delivery types.

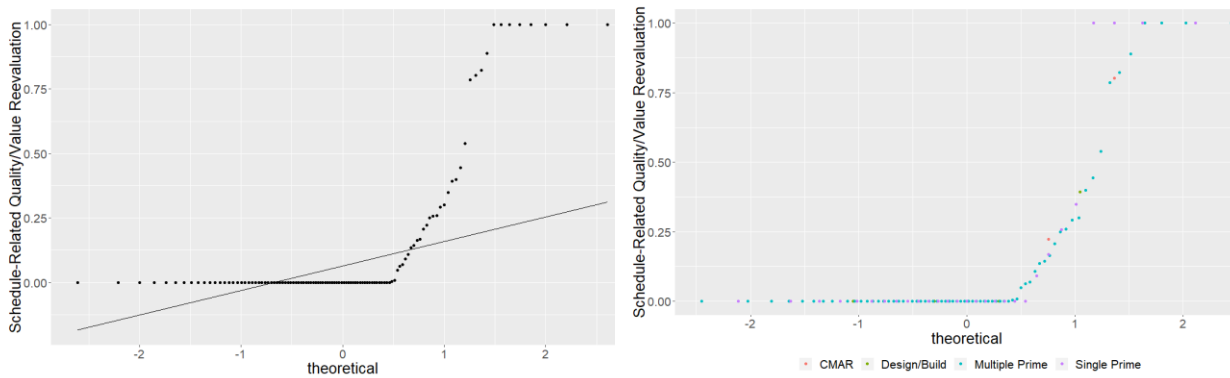


Figure 93: Q-Q Plot of Schedule-Related Changes Due to Quality Reevaluation

Table 83: Significance of Normality for Schedule-Related Changes Due to Quality Reevaluation

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.001	Significant	--
	CM at Risk	0.000	Significant	--
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

Per the comparative boxplot below, there is no immediately evident delivery method which is higher in performance in terms of schedule-related change due to quality/value reevaluation. Also, all the methods experience similar variability.

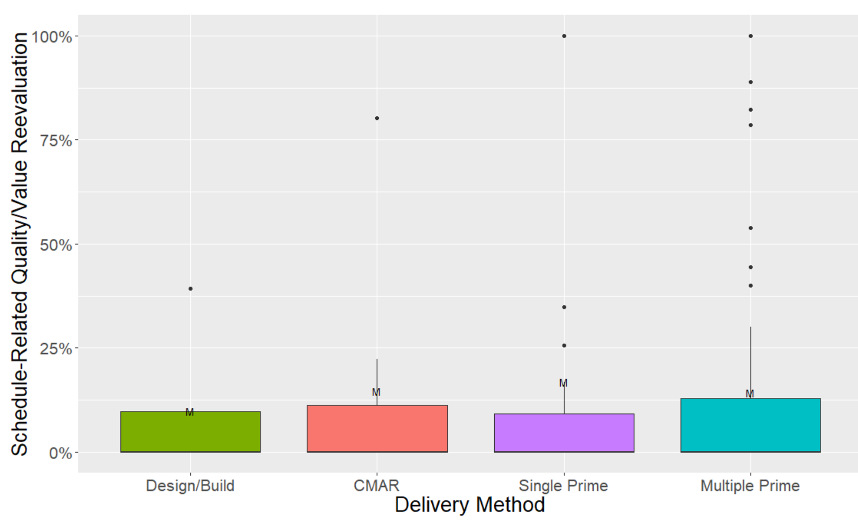


Figure 94: Boxplot of Schedule-Related Changes Due to Quality Reevaluation

Per the table of descriptive statistics below, the median values of all four delivery methods are 0%. Standard deviations were generally similar as well; DB, CMAR, SP, and MP had similar figures: 19.6, 30.1%, 34.9, and 28.9% respectively. This suggests there is no difference in performance or variability between the different delivery methods, but more robust statistical testing is needed.

Table 84: Descriptive Statistics for Schedule-Related Changes Due to Quality Reevaluation Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	9.8%	14.6%	16.8%	14.2%
Maximum	39.2%	80.3%	100.0%	100.0%
75% Percentile	29.4%	22.2%	12.9%	13.7%
Median	0.0%	0.0%	0.0%	0.0%
25% Percentile	0.0%	0.0%	0.0%	0.0%
Minimum	0.0%	0.0%	0.0%	0.0%
Standard Deviation	19.6%	30.1%	34.9%	28.9%

Robust statistical tests were performed, with the following results. Fligner-Killeen returned a *p*-value of 0.981, which was insufficient to reject the null hypothesis of equal variance among delivery methods at the 95 and 90% confidence levels. Kruskal-Wallis returned a *p*-value of 0.989, which also failed to reject the null hypothesis of equal population of medians at the 95 and 90% confidence levels. Therefore, the post-hoc test was not performed.

Table 85: Significance of Variance & Distribution for Schedule Changes Due to Quality Reevaluation

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.981	Not Significant	Not Significant
Kruskal-Wallis	Overall	0.989	Not Significant	Not Significant
Conover-Iman	DB vs Single	NA	NA	NA
	DB vs Multiple	NA	NA	NA
	CMAR vs Single	NA	NA	NA
	CMAR vs Multiple	NA	NA	NA
	DB vs CMAR	NA	NA	NA
	Single vs Multiple	NA	NA	NA

5.2.2.4.2 Comparative Analysis of Delivery Type

A secondary analysis of delivery type was performed to assess if there was any correlation between that variable and change orders due to reevaluation. Once again, normality was tested and rejected at the 95% confidence level, per the Q-Q plot and table below.

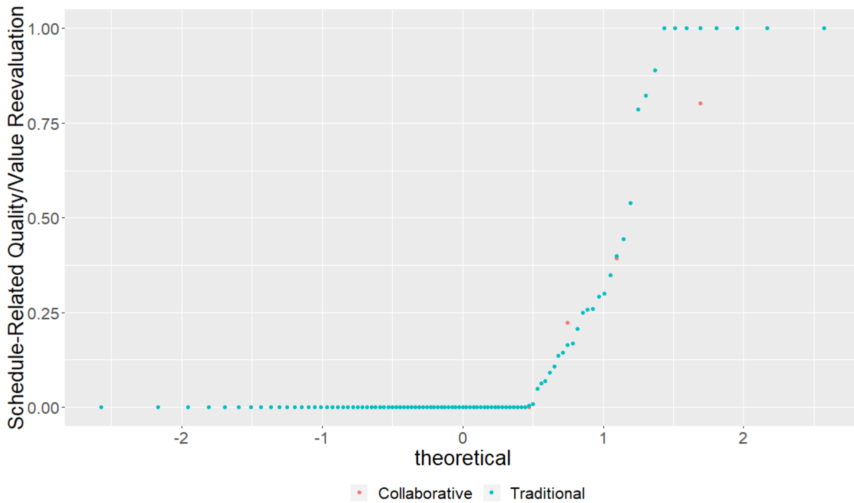


Figure 95: Q-Q Plot of Schedule-Related Changes Due to Quality Reevaluation

Table 86: Significance of Normality for Schedule-Related Changes Due to Quality Reevaluation

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.000	Significant	--
	Traditional	0.000	Significant	--

The comparative boxplot drawn below shows no immediately evident difference between delivery types, nor in terms of variability.

Figure 96: Boxplot of Schedule-Related Changes Due to Quality Reevaluation

The following table of descriptive statistics was drawn in order to assess the suspicions brought previously in this section. Median values were similar between Collaborative and Traditional delivery (0.0 and 0.0%). No difference is suggested by these figures. Variability was also similar, with standard deviations of 25.8 and 30.6%. This is consistent with the visual observation from the boxplot. More robust testing was needed.

Table 87: Descriptive Statistics for Schedule-Related Changes Due to Quality Reevaluation Per PDT

Statistical Test	Collaborative	Traditional
Mean	12.9%	14.9%
Maximum	80.3%	100.0%
75% Percentile	22.2%	13.5%
Median	0.0%	0.0%
25% Percentile	0.0%	0.0%
Minimum	0.0%	0.0%
Standard Deviation	25.8%	30.6%

Fligner-Killeen test returned a *p*-value of 0.694. This was insufficient to reject the null hypothesis of equal variance among delivery types at the 95 and 90% confidence levels. This suggests no difference in variability between both types.

Mann-Whitney U test returned a *p*-value of 0.820, which is also insufficient to reject the null hypothesis of equal medians at either the 95% or 90 confidence level.

Table 88: Significance of Variance & Distribution for Schedule Changes Due to Quality Reevaluation

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.694	Not Significant	Not Significant
Mann–Whitney	Overall	0.820	Not Significant	Not Significant

5.2.2.5 Program Work Reinstated

5.2.2.5.1 <u>Comparative Analysis of Delivery Methods</u>

First, normality was assessed and rejected at the 95%, as shown in the Q-Q plots and table below.

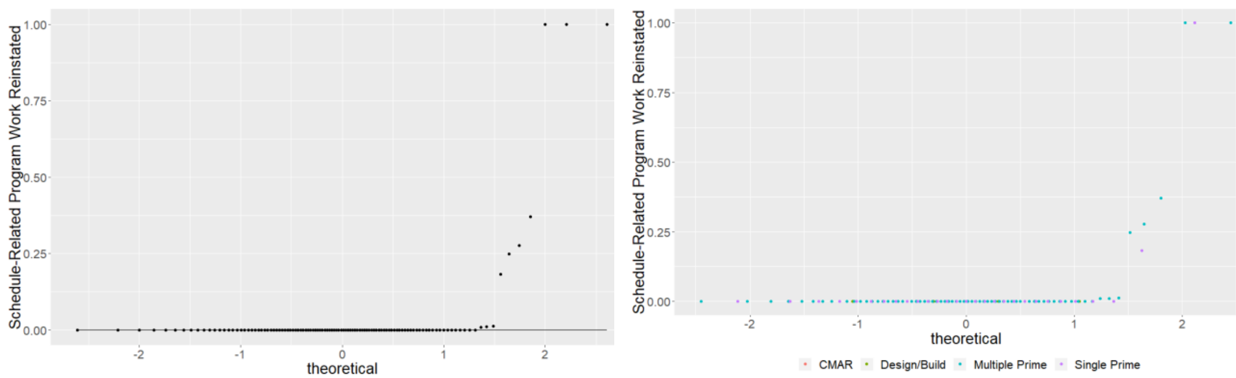


Figure 97: Q-Q Plot of Schedule-Related Changes Due to Program Reinstated

Table 89: Significance of Normality for Schedule-Related Changes Due to Program Reinstated

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build*	NA	NA	NA
	CM at Risk*	NA	NA	NA
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

*all the values are equivalent. Normality cannot be tested

Per the comparative boxplot below, there is no visually evident difference between the different delivery methods. It is visually evident that there is similar variability between the four delivery methods.

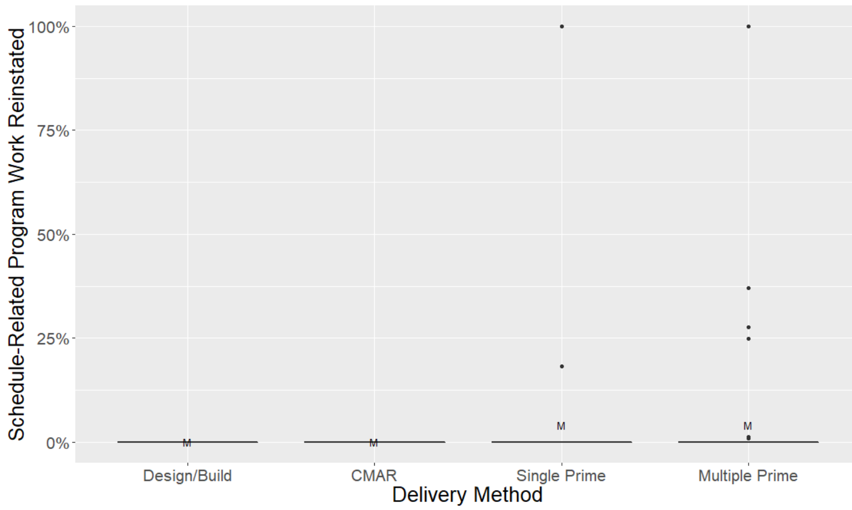


Figure 98: Boxplot of Schedule-Related Changes Due to Program Reinstated

The following table of descriptive statistics was drawn to increase the robustness of the conclusions. The median values of each delivery method were all 0.0%. These statistics suggest no difference between the delivery methods but will still require further statistical tests. DB and CMAR had similar standard deviations at 0.0 and 0.0 percent, with SP at 18.8% and MP at 17.7%.

Table 90: Descriptive Statistics for Schedule-Related Changes Due to Program Reinstated Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	0.0%	0.0%	4.1%	4.2%
Maximum	0.0%	0.0%	100.0%	100.0%
75% Percentile	0.0%	0.0%	0.0%	0.0%
Median	0.0%	0.0%	0.0%	0.0%
25% Percentile	0.0%	0.0%	0.0%	0.0%
Minimum	0.0%	0.0%	0.0%	0.0%
Standard Deviation	0.0%	0.0%	18.8%	17.7%

More robust tests were performed to quantitatively confirm the previous findings. Fligner-Killeen returned a p-value of 0.675, which was insufficient to reject the null hypothesis of equal variance at the 95 and 90% confidence levels. Kruskal-Wallis, furthermore, returned a p-value of 0.643, which was also insufficient to reject the null hypothesis of equal populations of medians among delivery methods. Accordingly, Conover-Iman post-hoc test was not performed.

Table 91: Significance of Variance & Distribution for Schedule Changes Due to Program Reinstated

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.674	Not Significant	Not Significant
Kruskal-Wallis	Overall	0.643	Not Significant	Not Significant
Conover-Iman	DB vs Single	NA	NA	NA
	DB vs Multiple	NA	NA	NA
	CMAR vs Single	NA	NA	NA
	CMAR vs Multiple	NA	NA	NA
	DB vs CMAR	NA	NA	NA
	Single vs Multiple	NA	NA	NA

5.2.2.5.2 Comparative Analysis of Delivery Type

A second quantitative assessment was performed to study the impact of each delivery method. The Q-Q plot and table below show a clear departure from normality at the 95% confidence level.

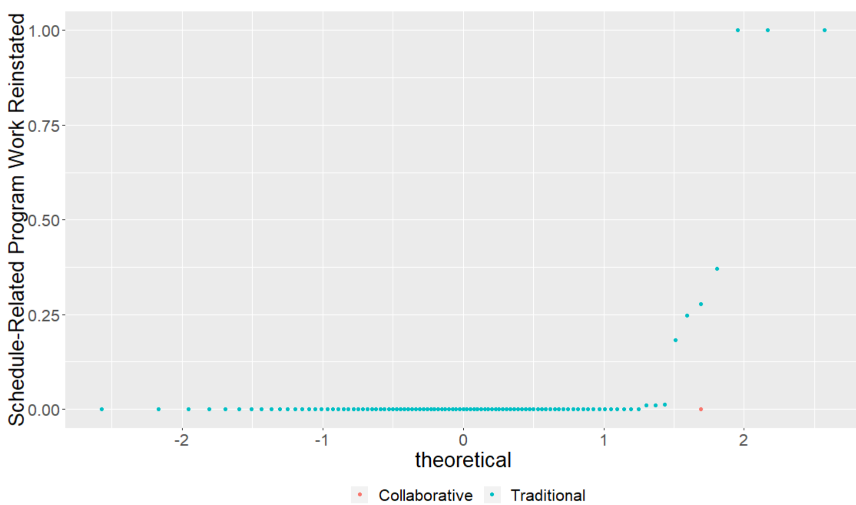


Figure 99: Q-Q Plot of Schedule-Related Changes Due to Program Reinstated

Table 92: Significance of Normality for Schedule-Related Changes Due to Program Reinstated

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative*	NA	NA	NA
	Traditional	0.000	Significant	--

*all the values are equivalent. Normality cannot be tested

The comparative boxplot below shows no visually evident difference between delivery types. In terms of variability, Collaborative was very similar to Traditional.

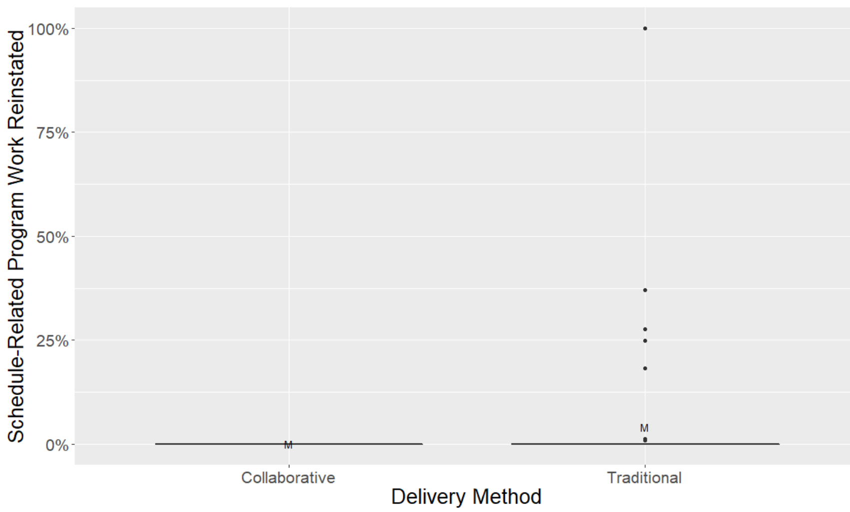


Figure 100: Boxplot of Schedule-Related Changes Due to Program Reinstated

Therefore, the descriptive statistical table below was drawn to further characterize these findings. No clear difference in the median of each delivery type was found, with both types at 0%. The mean values of the delivery types were also highly similar, 0.0% and 4.2%, respectively. Collaborative methods also experienced much lower standard deviation to Traditional (0.0 and 17.9%). Therefore, additional analysis was performed.

Table 93: Descriptive Statistics for Schedule-Related Changes Due to Program Reinstated Per PDT

Statistical Test	Collaborative	Traditional
Mean	0.0%	4.2%
Maximum	0.0%	100.0%
75% Percentile	0.0%	0.0%
Median	0.0%	0.0%
25% Percentile	0.0%	0.0%
Minimum	0.0%	0.0%
Standard Deviation	0.0%	17.9%

In this analysis, Fligner-Killeen returned a *p*-value of 0.280, which is insufficient to reject the null hypothesis of equal variance among delivery types at either the 95 or 90% confidence levels. Mann-Whitney U test returned a *p*-value of 0.276, which was also insufficient to reject the null hypothesis of equal medians among delivery types.

Table 94: Significance of Variance & Distribution for Schedule Changes Due to Program Reinstated

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.280	Not Significant	Not Significant
Mann–Whitney	Overall	0.276	Not Significant	Not Significant

5.2.2.6 Other Reasons

5.2.2.6.1 Comparative Analysis of Delivery Methods

Normality was initially assessed and rejected at the 95% confidence level, as shown in the following Q-Q plots and table.

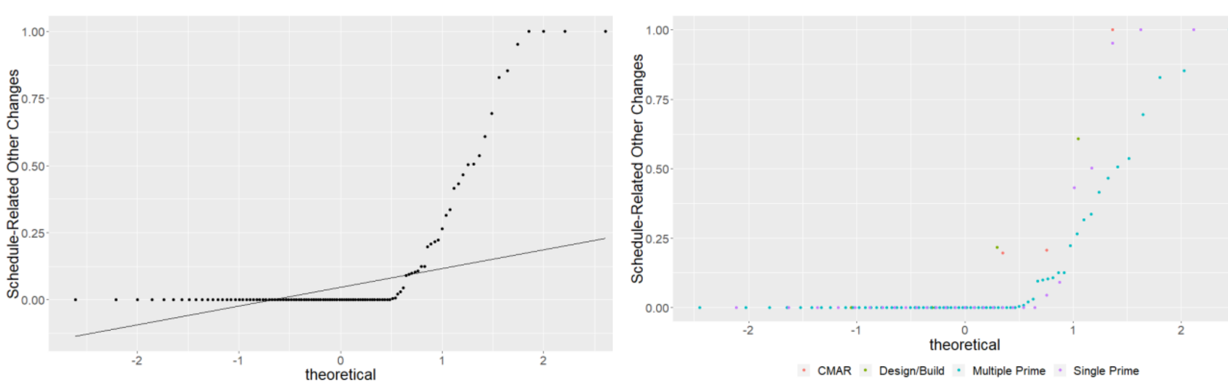


Figure 101: Q-Q Plot of Schedule-Related Changes Due to Other Reasons

Table 95: Significance of Normality for Schedule-Related Changes Due to Other Reasons

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.183	Not Significant	Not Significant
	CM at Risk	0.000	Significant	--
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

No clear difference was identified visually among the delivery methods CMAR, SP, and MP. DB shows slightly higher value. DM and CMAR also had more variability compared to SP and MP.

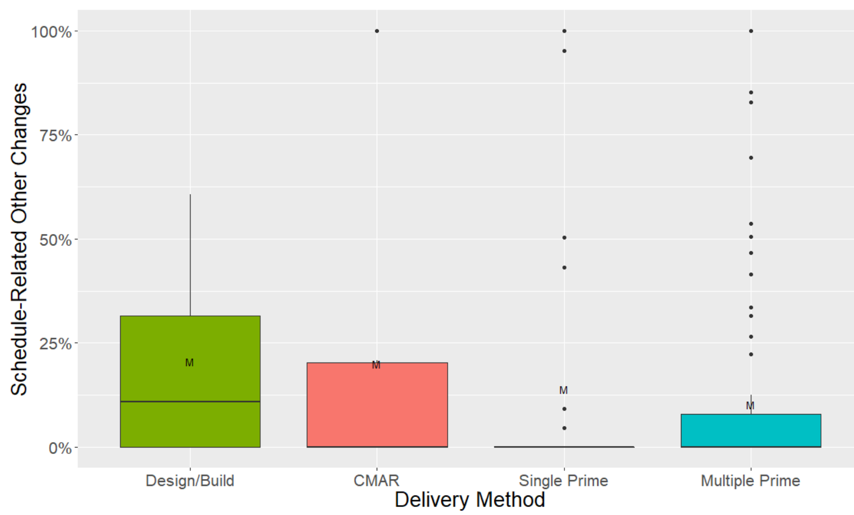


Figure 102: Boxplot of Schedule-Related Changes Due to Other Reasons

The table of descriptive statistics, below, highlights the similarity of the medians among CMAR, SP and MP at 0.0% each, with DB having the highest value at 10.9%. Similar standard deviations were experienced (28.6, 36.5, 31.6, and 22.5%, respectively). Further tests were performed to confirm both these findings.

Table 96: Descriptive Statistics for of Schedule-Related Changes Due to Other Reasons Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	20.6%	20.1%	13.9%	10.2%
Maximum	60.8%	100.0%	100.0%	100.0%
75% Percentile	51.0%	20.8%	2.3%	9.6%
Median	10.9%	0.0%	0.0%	0.0%
25% Percentile	0.0%	0.0%	0.0%	0.0%
Minimum	0.0%	0.0%	0.0%	0.0%
Standard Deviation	28.6%	36.5%	31.6%	22.5%

In this secondary statistical testing, Fligner-Killen returned a *p*-value of 0.169, which is insufficient to reject the null hypothesis at either the 95 or 90% confidence levels and suggest similar variability among all methods.

Kruskal-Wallis returned a *p*-value of 0.622, which is also insufficient to reject the null hypothesis of equal population medians at either the 95 or 90% confidence levels. Ergo no post hoc testing was performed.

Table 97: Significance of Variance & Distribution for Schedule-Related Changes Due to Other Reasons

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.169	Not Significant	Not Significant
Kruskal-Wallis	Overall	0.622	Not Significant	Not Significant
Conover-Iman	DB vs Single	NA	NA	NA
	DB vs Multiple	NA	NA	NA
	CMAR vs Single	NA	NA	NA
	CMAR vs Multiple	NA	NA	NA
	DB vs CMAR	NA	NA	NA
	Single vs Multiple	NA	NA	NA

5.2.2.6.2 <u>Comparative Analysis of Delivery Type</u>

A secondary assessment was performed that contrasted performance by delivery type. First, normality was assessed and rejected at the 95% confidence level, as shown in the Q-Q plot and table below.

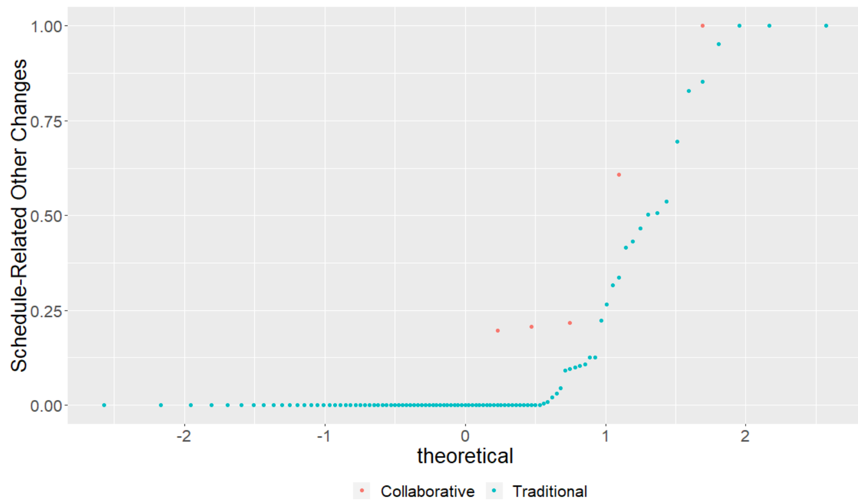


Figure 103: Q-Q Plot of Schedule-Related Changes Due to Other Reasons

Table 98: Significance of Normality for Schedule-Related Changes Due to Other Reasons

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.000	Significant	--
	Traditional	0.000	Significant	--

The boxplots were then subsequently drawn to highlight visual differences between delivery types. Collaborative delivery methods were notably more variable than Traditional ones.

Figure 104: Boxplot of Schedule-Related Changes Due to Other Reasons

To create more robust statistical observations, the following table of descriptive statistics was drawn. The median data, as evident, was similar in both delivery types at 0.0 and 0.0%, respectively. Collaborative methods also evidently experienced a higher level of variability, with standard deviation of 32.3%, as compared to Traditional (25.4%).

Table 99: Descriptive Statistics for Schedule-Related Changes Due to Other Reasons Per PDT

Statistical Test	Collaborative	Traditional
Mean	20.3%	11.3%
Maximum	100.0%	100.0%
75% Percentile	21.7%	4.5%
Median	0.0%	0.0%
25% Percentile	0.0%	0.0%
Minimum	0.0%	0.0%
Standard Deviation	32.3%	25.4%

Fligner-Killeen and Mann-Whitney U tests were performed to validate the previous findings and claims. Fligner-Killeen test returned a *p*-value of 0.208, insufficient to reject the null hypothesis of equal variance among delivery types. Mann-Whitney U test returned a *p*-value of 0.221, which is also insufficient to reject the null hypothesis of equal medians among the two delivery types, at both the 95 and 90% confidence levels.

Table 100: Significance of Variance & Distribution for Schedule-Related Changes Due to Other Reasons

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.208	Significant	--
Mann–Whitney	Overall	0.221	Significant	--

5.2.3 Overall (Cost & Schedule) Change Orders

So far, the research assessed the likelihood of various change order reasons for each of the four delivery methods, as well as each delivery type. Having this assessment performed for cost-related and schedule-related changes was essential to evaluate each specific case on its own. However, with too many variables and conclusions, it was also significantly valuable to perform a comprehensive analysis that studied both cost-related and schedule-related changes combined to highlight and summarize key findings and also understand the big-picture of change orders likelihood of occurrence for each delivery method and delivery type.

Figure 105 shows the distribution of reasons for overall change orders, where it can be noticed that Scope Changes was the most common reason for overall change orders (39%), followed by Unforeseen Conditions (22%), in addition to Design Oversight (12%), which combined represented 73% of all the issued change orders. Very similar figures were reported in the literature (Hanna 2001). For that reason, more emphasis was given to those 3 specific reasons during the analysis.

Also, while all change orders generally have negative ramifications on a construction project as it indicates change in plans, disruption to existing operations and rework in some cases, Scope Changes have been regarded as a more valuable reason for a change, when compared to Unforeseen Conditions and Design Oversight as explained earlier.

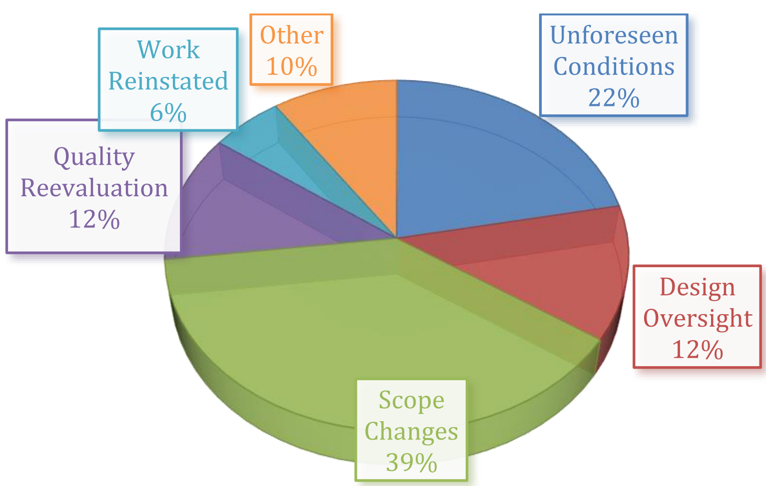


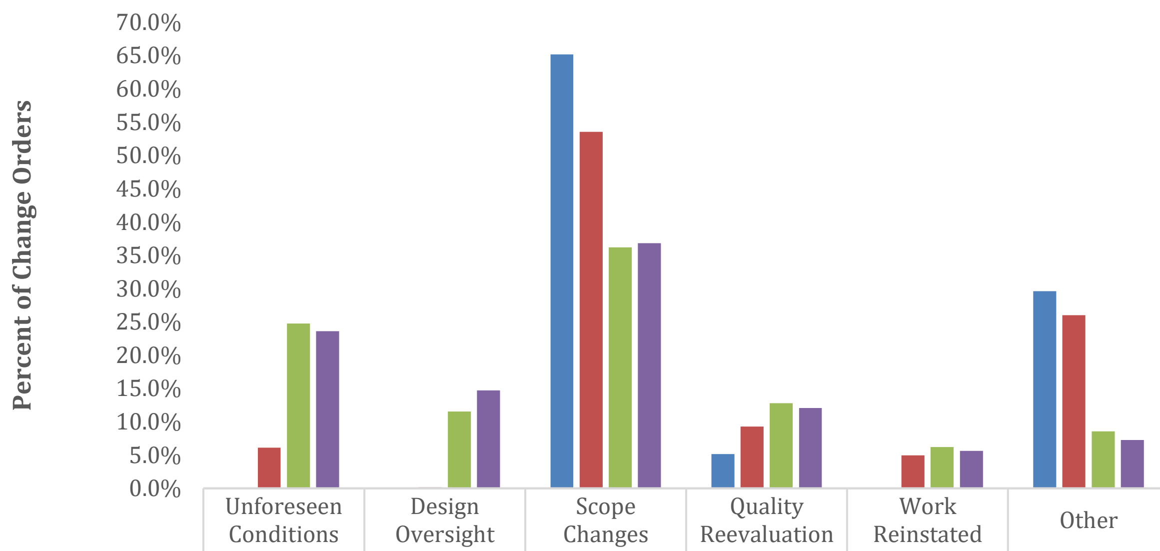
Figure 105: Distribution of Overall Change Orders by Reasons

Since delivery method is the key focus of this research, it was essential to assess the impact of each delivery method on the overall change orders. Referring to the following figure, it can be noticed that there is a clear difference in the average abundance of Unforeseen Conditions change orders where DB and CMAR projects experience minimal average percentages at 0.1% and 6.1% respectively, as opposed to SP and MP projects at on average 24.7% and 23.6% respectively.

Similarly, a clear difference in the average abundance of Design Oversight change orders where DB and CMAR projects experience minimal average percentages at 0% and 0.1% respectively, as opposed to SP and MP projects at, on average, 11.5% and 14.7% respectively.

It can be noticed also that some difference exists when it comes to Scope Changes, with DB having the highest percentage followed by CMAR, SP and MP, which will be studied more closely in the coming sections.

On the other hand, both Quality Reevaluation and Work Reinstated are very similar across the different delivery methods, with hardly any noticed differences. This indicates that those changes are not dependent on the project delivery method. Furthermore, changes due to Other reasons tend to be much higher in DB and CMAR delivery methods at 29.6% and 26% respectively as compared to SP and MP at 8.5% and 7.2% respectively; however, more analysis will be conducted to assess this difference.



	Unforeseen Conditions	Design Oversight	Scope Changes	Quality Reevaluation	Work Reinstated	Other
Design/Build	0.1%	0.0%	65.2%	5.2%	0.0%	29.6%
CMAR	6.1%	0.1%	53.5%	9.3%	5.0%	26.0%
Single prime	24.7%	11.5%	36.2%	12.8%	6.2%	8.5%
Multiple Prime	23.6%	14.7%	36.8%	12.0%	5.6%	7.2%

Design/Build CMAR Single prime Multiple Prime

Figure 106: Average Likelihood of Overall Change Orders by Delivery Method

The following sections shed more light on each change order reasons by providing statistical tests to quantitatively assess the significant differences among the different delivery methods, as well as comment on the variance within each group.

5.2.3.1 Unforeseen Conditions

5.2.3.1.1 <u>Comparative Analysis of Delivery Methods</u>

The first step in conducting this analysis was validating the assumption of normality to determine which significance tests were appropriate to use. The Q-Q plot below shows a clear departure from normality at the overall dataset, and also at the level of each delivery method.

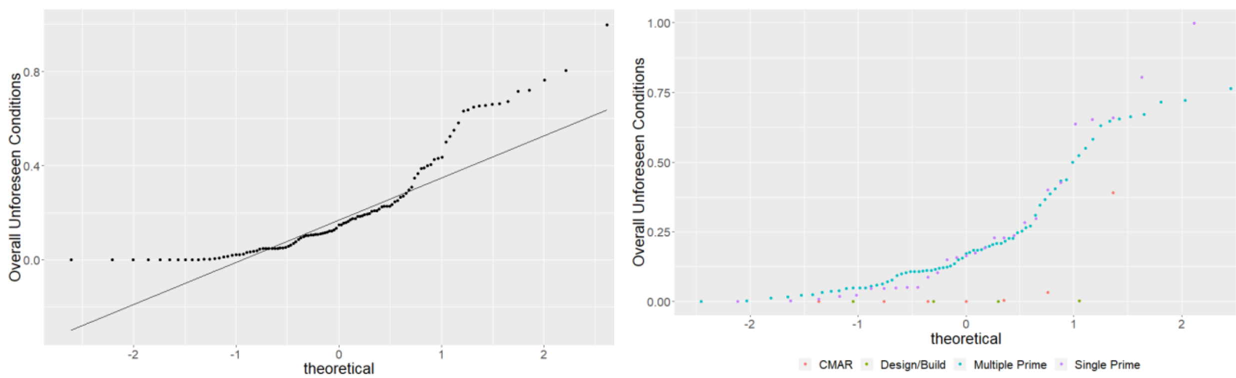


Figure 107: Q-Q Plot of Overall Changes Due to Unforeseen Conditions

Confirming the visual normality of the above Q-Q plots, the table below presents the results of a Shapiro-Wilk test, with a resultant *p*-value of 0.000 for each delivery method, sufficient to reject the null hypothesis of normality at the 95% confidence level.

Table 101: Significance of Normality for Overall Changes Due to Unforeseen Conditions

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.000	Significant	--
	CM at Risk	0.000	Significant	--
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

By using box-and-whisker plots, a clear difference is identified between the Collaborative delivery methods and Traditional Methods. However, within the subgroups there are no visually evident differences between DB and CMAR, while a minor difference was evident between SP and MP. Further, it is visually evident that Traditional delivery methods have higher variability than Collaborative ones from the level of compaction of the boxplot.

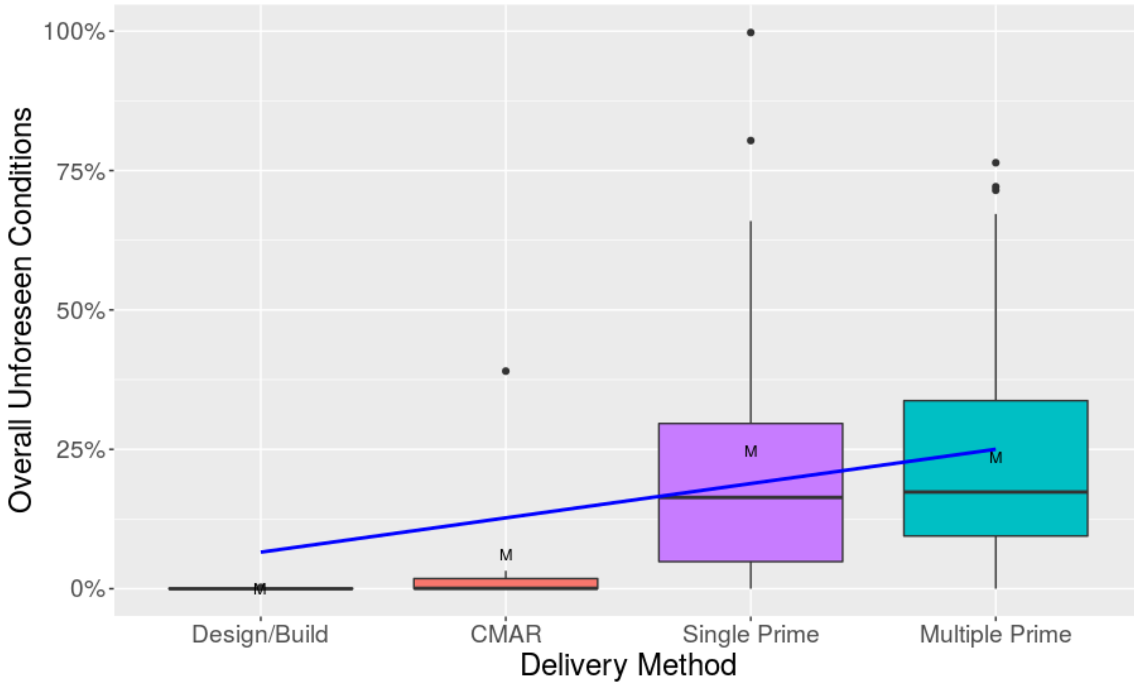


Figure 108: Boxplot of Overall Changes Due to Unforeseen Conditions

This boxplot is broken out into a descriptive statistical table, below. This table suggests a difference in median performance between DB and CMAR (0 and 0.1%) and SP and MP (16.4 and 17.3% on the other hand). Slight difference is highlighted between SP and MP. Therefore, further statistical tests were performed. Variability shows that Collaborative delivery experienced low standard deviations (0.1 and 14.6%) while Traditional was much higher (26.6 and 21.1%). Further tests were performed to assess if these deviations were influenced by outliers.

Table 102: Descriptive Statistics for Overall Changes Due to Unforeseen Conditions Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	0.1%	6.1%	24.7%	23.6%
Maximum	0.2%	39.0%	99.8%	76.4%
75% Percentile	0.2%	3.2%	34.9%	35.1%
Median	0.0%	0.1%	16.4%	17.3%
25% Percentile	0.0%	0.0%	4.8%	8.9%
Minimum	0.0%	0.0%	0.0%	0.0%
Standard Deviation	0.1%	14.6%	26.6%	21.1%

Given the previously stated normality testing, appropriate statistical tests were performed. Fligner-Killeen test resulted in a *p*-value of 0.000, sufficient at the 95% confidence level to reject the null hypothesis of equal variance among delivery methods, confirming the previous suggestion of higher variability among Traditional methods.

Additionally, Kruskal-Wallis H-Test resulted in *p* = 0.000, sufficient to reject the null hypothesis of equal population medians among delivery methods at the 95% confidence level. Post-hoc, Conover-Iman test was performed to evaluate pairwise statistical differences between the individual delivery methods. Statistically, differences were found between DB and SP/MP and CMAR and SP/MP, but not between DB and CMAR nor SP and MP.

Table 103: Significance of Variance & Distribution for Overall Changes Due to Unforeseen Conditions

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.005	Significant	--
Kruskal-Wallis	Overall	0.000	Significant	--
Conover-Iman	DB vs Single	0.003	Significant	--
	DB vs Multiple	0.001	Significant	--
	CMAR vs Single	0.009	Significant	--
	CMAR vs Multiple	0.003	Significant	--
	DB vs CMAR	0.752	Not Significant	Not Significant
	Single vs Multiple	0.930	Not Significant	Not Significant

These findings suggest that DB have a higher performance compared to SP and MP, as well as CMAR relative to SP and MP in terms of overall change orders due to unforeseen conditions.

5.2.3.1.2 Comparative Analysis of Delivery Type

A second quantitative assessment was performed to assess the impact of each delivery type on change orders initiated due to unforeseen conditions. Therefore, a similar strategy as in the preceding section was performed to validate normality and select significance tests. The Q-Q plot shows a clear departure from normality in each delivery type.

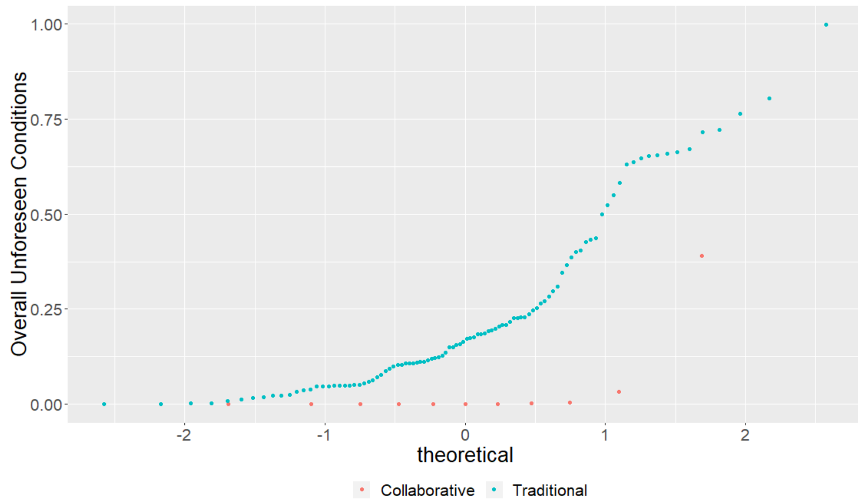


Figure 109: Q-Q Plot of Overall Changes Due to Unforeseen Conditions

As previously, *p*-values of 0.000 from Shapiro-Wilk are sufficient to reject the null hypothesis of normality at the 95% confidence level.

Table 104: Significance of Normality for Overall Changes Due to Unforeseen Conditions

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.000	Significant	--
	Traditional	0.000	Significant	--

A comparative boxplot identified visual differences between Collaborative and Traditional delivery types. Lower variability among Collaborative delivery was also evident.

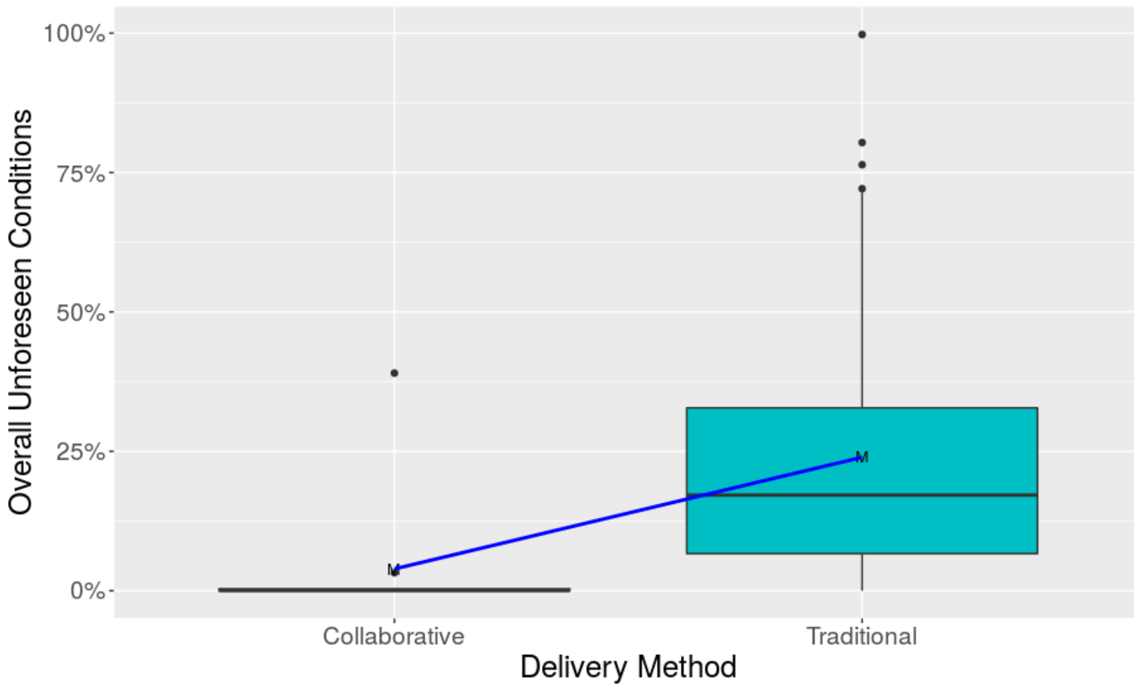


Figure 110: Boxplot of Overall Changes Due to Unforeseen Conditions

The table of descriptive statistics below presents a clear difference in median and mean for Collaborative (0/3.9%) and Traditional (17.1/23.9%) respectively. Collaborative methods also exhibited lower standard deviation than Traditional (11.7% v. 22.7%).

Table 105: Descriptive Statistics for Overall Changes Due to Unforeseen Conditions Per PDT

Statistical Test	Collaborative	Traditional
Mean	3.9%	23.9%
Maximum	39.0%	99.8%
75% Percentile	0.4%	34.7%
Median	0.0%	17.1%
25% Percentile	0.0%	6.3%
Minimum	0.0%	0.0%
Standard Deviation	11.7%	22.7%

Based once again on the normality testing, statistical tests were performed. Fligner-Killeen and Mann-Whitney U tests both returned *p*-values of 0.000, sufficient to reject the appropriate null hypotheses at the 95% confidence level. Thus, it was concluded the inequal variance among delivery types, and that there were not equal medians among the two delivery types.

Table 106: Significance of Variance & Distribution for Overall Changes Due to Unforeseen Conditions

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Mann–Whitney	Overall	0.000	Significant	--

This supports previous evidence to the higher performance of Collaborative delivery over Traditional delivery in terms of overall change orders due to Unforeseen Conditions.

While logic might dictate that unforeseen conditions, given their unforeseen nature, should be independent of delivery method, results suggest that certain delivery methods (the Collaborative delivery methods) were better able to handle unforeseen change and minimize project impact.

5.2.3.2 Design Oversight

5.2.3.2.1 <u>Comparative Analysis of Delivery Methods</u>

As with the other analysis, the initial step was to determine normality. As evident from the Q-Q plots below, normality was not present. This was confirmed at the 95% confidence level by the Shapiro-Wilk test, as shown in the table below.

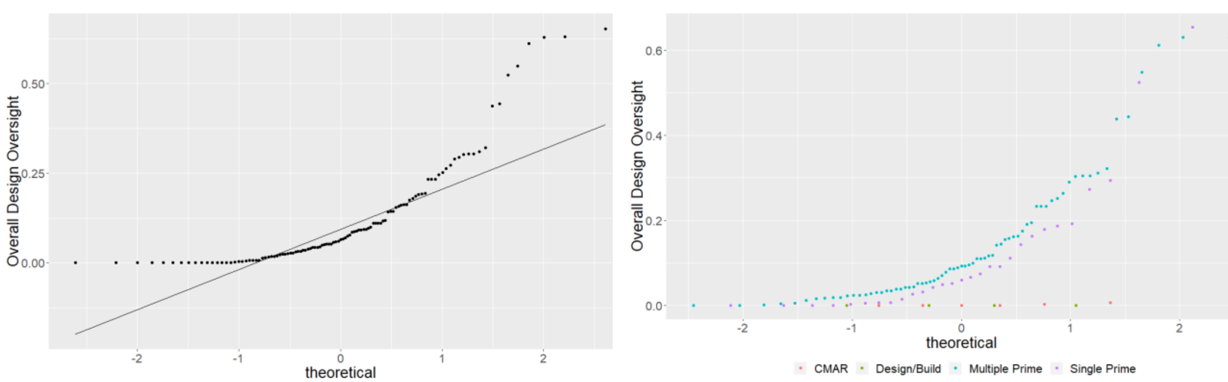


Figure 111: Q-Q Plot of Overall Changes Due to Design Oversight

Table 107: Significance of Normality for Overall Changes Due to Design Oversight

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build*	NA	NA	NA
	CM at Risk	0.000	Significant	--
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

*all the values are equivalent. Normality cannot be tested

By considering the comparative boxplot below, there is a clear visual difference in performance between Collaborative and Traditional delivery methods, but there is no clear difference between DB and CMAR and only a slight one between SP and MP. Traditional delivery also exhibited higher variability.

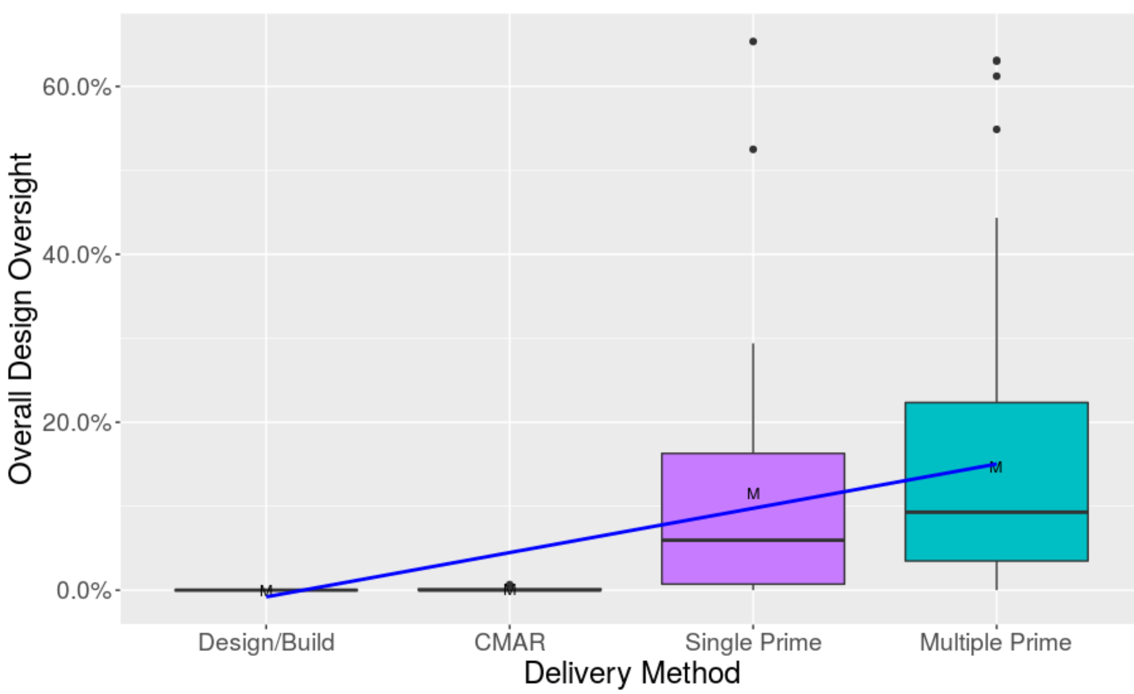


Figure 112: Boxplot of Overall Changes Due to Design Oversight

The table of descriptive statistics, below, was compiled to validate the observations. A clear difference in the median of Collaborative (0/0%) and Traditional (5.9/9.3%). The medians suggest no difference between DB and CMAR, and a slight difference between SP and MP, as was previously speculated. Further testing was performed to validate this assumption. Variability was higher in SP and MP (15.6/15.6%) than DB and CMAR (0.0/0.2%). Once again, statistical tests were performed to validate these findings.

Table 108: Descriptive Statistics for Overall Changes Due to Design Oversight Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	0.0%	0.1%	11.5%	14.7%
Maximum	0.0%	0.7%	65.4%	63.1%
75% Percentile	0.0%	0.2%	17.1%	23.3%
Median	0.0%	0.0%	5.9%	9.3%
25% Percentile	0.0%	0.0%	0.7%	3.4%
Minimum	0.0%	0.0%	0.0%	0.0%
Standard Deviation	0.0%	0.2%	15.6%	15.6%

Fligner-Killeen returned a *p*-value of 0.000, confirming the inequal variance among delivery methods at the 95% confidence level and suggesting a higher variability in SP and MP than in DB and CMAR.

Kruskal-Wallis further returned a *p*-value of 0.000, rejecting the assumption of equal medians at the 95% confidence level. Conover-Iman post-hoc test showed statistically significant differences between DB and SP/MP and CMAR and SP/MP, with no difference between DB and CMAR nor between SP and MP.

Table 109: Significance of Variance & Distribution for Overall Changes Due to Design Oversight

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Kruskal-Wallis	Overall	0.000	Significant	--
Conover-Iman	DB vs Single	0.007	Significant	--
	DB vs Multiple	0.000	Significant	--
	CMAR vs Single	0.001	Significant	--
	CMAR vs Multiple	0.000	Significant	--
	DB vs CMAR	0.951	Not Significant	Not Significant
	Single vs Multiple	0.301	Not Significant	Not Significant

Yet further evidence was found in this section of DB and CMAR's ability to outperform SP and MP, this time in terms of overall change orders due to design oversight. Ergo, DB and CMAR deliver higher design quality than DBB projects, with reduced variability.

5.2.3.2.2 Comparative Analysis of Delivery Type

A second quantitative assessment was performed to study the impact of each delivery type on initiated change orders due to design oversight. Once again, normality was assessed, and a lack of normality was found at the 95% confidence level, as shown in the figure and table below.

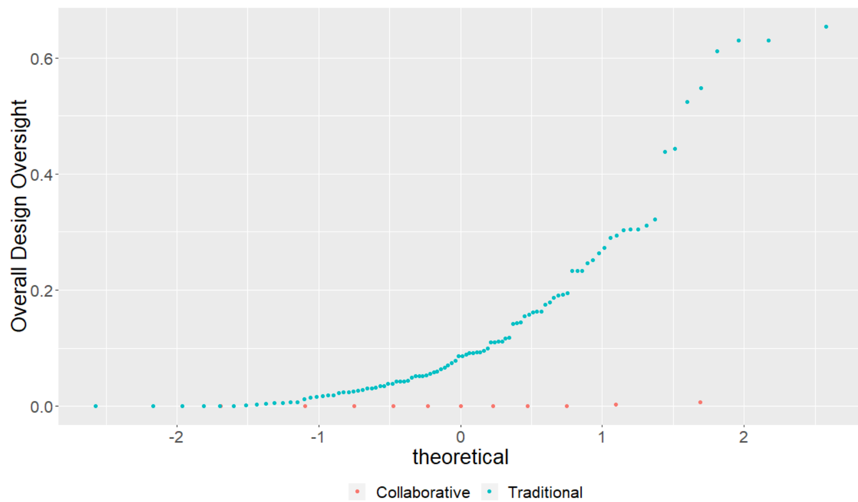


Figure 113: Q-Q Plot of Overall Changes Due to Design Oversight

Table 110: Significance of Normality for Overall Changes Due to Design Oversight

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.000	Significant	--
	Traditional	0.000	Significant	--

The following boxplot highlights a clear difference between Collaborative and Traditional delivery types. Higher variability was also noted in Traditional methods.

Figure 114: Boxplot of Overall Changes Due to Design Oversight

Thus, the following table of descriptive statistics was created, which highlights differences between the medians of Collaborative and Traditional methods (0 and 8.6%) and their means (0.1 and 13.8%). Furthermore, Collaborative delivery methods exhibited lower variability (std. deviation) than Traditional ones.

Table 111: Descriptive Statistics for Overall Changes Due to Design Oversight Per PDT

Statistical Test	Collaborative	Traditional
Mean	0.1%	13.8%
Maximum	0.7%	65.4%
75% Percentile	0.1%	19.1%
Median	0.0%	8.6%
25% Percentile	0.0%	2.8%
Minimum	0.0%	0.0%
Standard Deviation	0.2%	15.6%

To validate the preceding suggestions, statistical tests were performed. Fligner-Killeen returned a *p*-value of 0.000, sufficient to reject the null hypothesis at 95% confidence level. This suggests a higher variability for Traditional delivery types than Collaborative ones.

Mann-Whitney U tests also returned a *p*-value of 0.000, sufficient to reject the null hypothesis of equal medians, once again at the 95% confidence levels.

Table 112: Significance of Variance & Distribution for Overall Changes Due to Design Oversight

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Mann–Whitney	Overall	0.000	Significant	--

These findings validate the better performance of Collaborative delivery over Traditional, in this case in terms of overall change orders due to design oversight.

Largely, this occurs due to improved design quality generated from the early involvement of the contractor during the design process, which can reduce cost and schedule impacts. Additionally, design issues can be resolved prior to a change order needing to be issued as a contractual responsibility of the project parties (especially in the case of DB), which passes less risk along to the owner. Further research is needed to confirm these allegations.

5.2.3.3 Scope Changes

5.2.3.3.1 <u>Comparative Analysis of Delivery Methods</u>

The presumption of normality was once again rejected and confirmed by the Shapiro-Wilk test. All delivery methods were found significant at the 95% confidence levels.

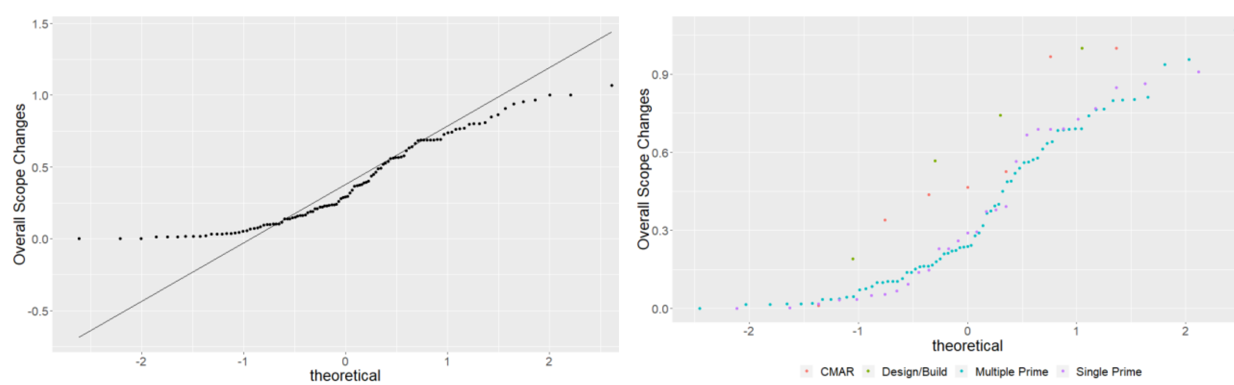


Figure 115: Q-Q Plot of Overall Scope Changes

Table 113: Significance of Normality for Overall Scope Changes

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build*	NA	NA	NA
	CM at Risk	0.000	Significant	--
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

*all the values are equivalent. Normality cannot be tested

The boxplot below shows descending values for DB, CMAR, SP and MP. SP was slightly more variable than the rest of delivery methods.

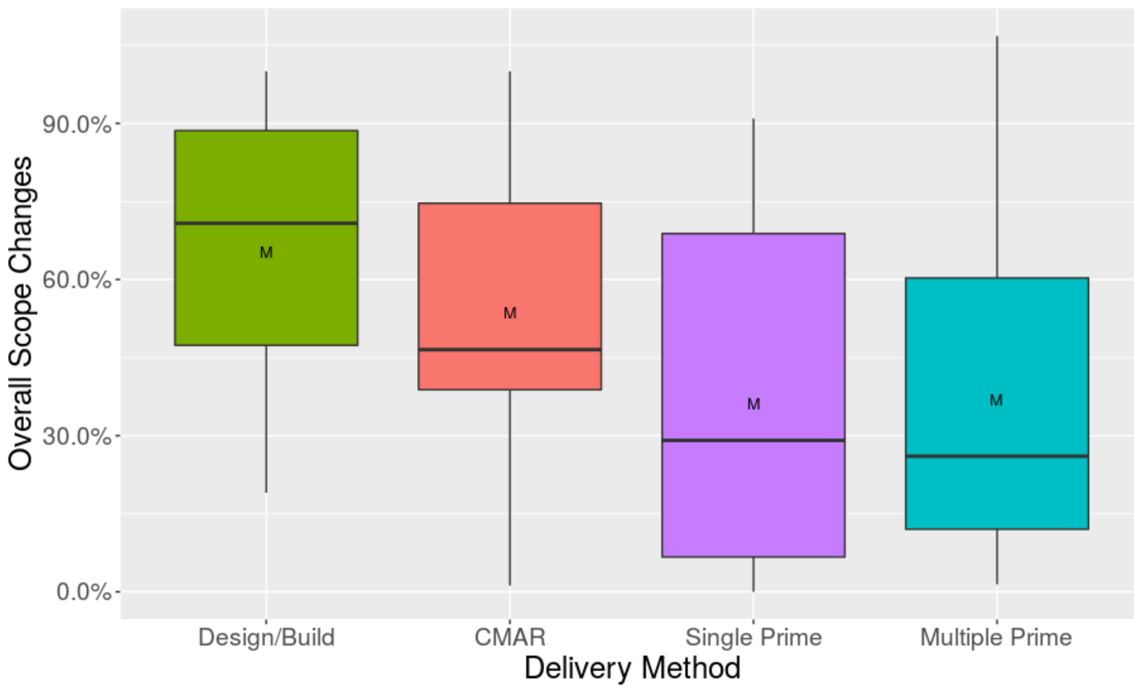


Figure 116: Boxplot of Overall Scope Changes

To provide more accurate findings, the table of descriptive statistics below was drawn. These statistics suggest a significant difference between DB and SP/ MP, with median values (70.8, 29.1, and 26.1%). However, CMAR is midway, with a median of 46.5%, which will be validated for significance with robust statistical methodologies.

All delivery methods experienced similar standard deviations (35.6, 34.9, 30.7 and 28.5%, respectively).

Table 114: Descriptive Statistics for Overall Scope Changes Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	65.2%	53.5%	36.2%	36.8%
Maximum	100.0%	100.0%	90.9%	106.8%
75% Percentile	96.2%	96.7%	68.9%	61.7%
Median	70.8%	46.5%	29.1%	26.1%
25% Percentile	28.5%	33.9%	6.1%	11.2%
Minimum	19.0%	1.2%	0.0%	1.4%
Standard Deviation	35.6%	34.9%	30.7%	28.5%

Statistical validation returned the following results. Fligner-Killeen returned a *p*-value of 0.979, which is insufficient to reject the null hypothesis of equal variance among delivery methods at either the 95 or 90% confidence levels. This suggests no variability between the four delivery methods. Kruskal-Wallis did not return a sufficient value to reject the null hypothesis (*p*-value = 0.248) at either the 95 or 90% confidence levels. No post hoc test was thus performed.

Table 115: Significance of Variance & Distribution for Overall Scope Changes

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.979	Not Significant	Not Significant
Kruskal-Wallis	Overall	0.248	Not Significant	Not Significant
Conover-Iman	DB vs Single	NA	NA	NA
	DB vs Multiple	NA	NA	NA
	CMAR vs Single	NA	NA	NA
	CMAR vs Multiple	NA	NA	NA
	DB vs CMAR	NA	NA	NA
	Single vs Multiple	NA	NA	NA

While DB and CMAR were found to be more prone to scope changes, by their nature they have a less-specified scope that is developed over the course of the project, analysis found that there was insignificant difference between the overall scope changes between Collaborative and Traditional delivery methods.

5.2.3.3.2 Comparative Analysis of Delivery Type

A second analysis was performed to study the impact of delivery type on initiated change orders due to scope. Once again, the assumption of normality was rejected, as shown in the Q-Q plot below and validated by Shapiro-Wilk at the 95% confidence level.

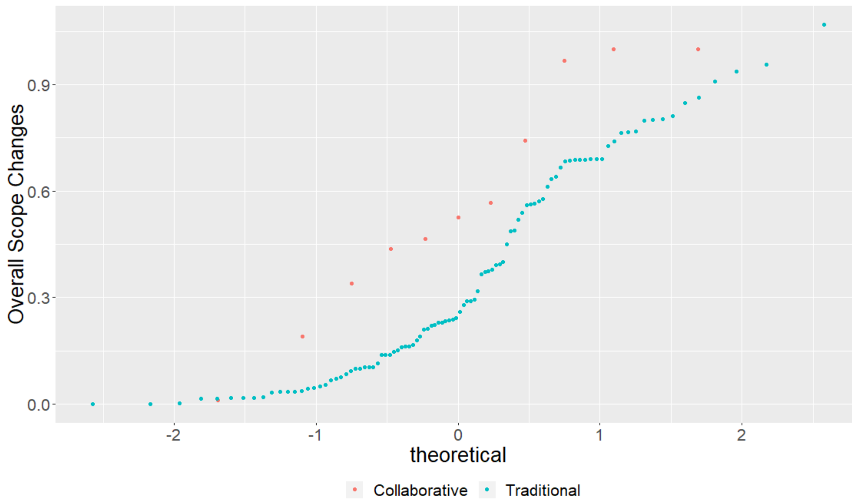


Figure 117: Q-Q Plot of Overall Scope Changes

Table 116: Significance of Normality for Overall Scope Changes

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.508	Not Significant	Not Significant
	Traditional	0.000	Significant	--

As shown in the comparative boxplot, Collaborative has visually higher median values than Traditional, but similar levels of variability.

Figure 118: Boxplot of Overall Scope Changes

To confirm the previous observations, the following table of descriptive statistics was drawn. Collaborative experienced higher median values compared to Traditional delivery (52.7 and 27.9%). However, it exhibited similar standard deviation to Traditional (33.8 vs 29.0%).

Table 117: Descriptive Statistics for Overall Scope Changes Per PDT

Statistical Test	Collaborative	Traditional
Mean	57.8%	36.6%
Maximum	100.0%	106.8%
75% Percentile	96.7%	64.0%
Median	52.7%	27.9%
25% Percentile	33.9%	10.3%
Minimum	1.2%	0.0%
Standard Deviation	33.8%	29.0%

More robust statistical tests were performed to confirm the preceding claims. Fligner-Killeen returned a *p*-value of 0.959 which is insufficient to reject the null hypothesis of equal variance among delivery types at the 95% confidence level. This suggests no variability difference between delivery types as anticipated.

In addition, Mann-Whitney U test resulted in a *p*-value of 0.051, which is barely insufficient to reject the null hypothesis of equal medians at the 95% confidence level.

Table 118: Significance of Variance & Distribution for Overall Scope Changes

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.959	Significant	--
Mann–Whitney	Overall	0.051	Not Significant	Significant

These findings are consistent with previous sections of this research - there is a comparable likelihood of occurrence among all the delivery methods in terms of overall change orders due to scope.

Although one would expect similar results for such methods due to their nature of having a less-specified scope that is typically developed over the course of the project, further analysis found that, there was insignificant difference in overall scope changes between Collaborative and Traditional delivery methods.

5.2.3.4 Quality/Value Reevaluation

5.2.3.4.1 <u>Comparative Analysis of Delivery Methods</u>

First, normality was assessed. The following Q-Q plots show a clear departure from normality. The subsequent table shows validation at the 95% confidence level across all delivery types.

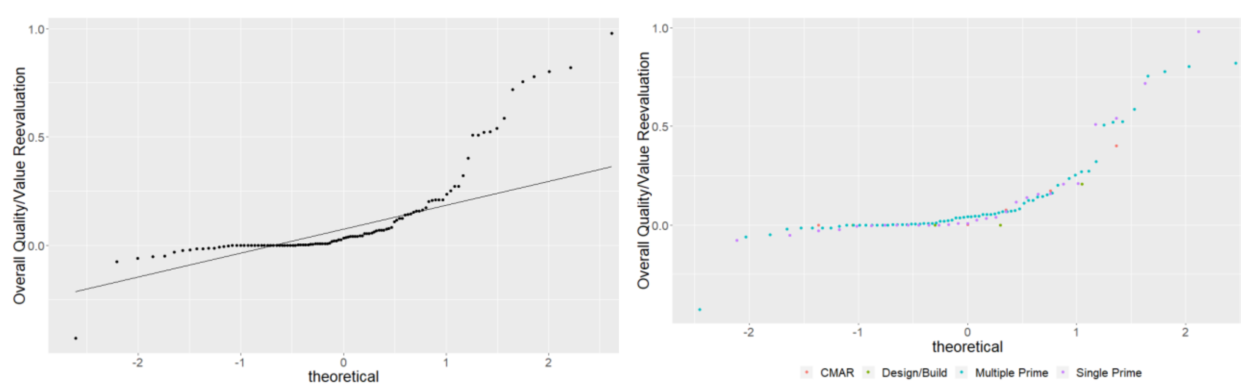


Figure 119: Q-Q Plot of Overall Changes Due to Quality Reevaluation

Table 119: Significance of Normality for Overall Changes Due to Quality Reevaluation

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.001	Significant	-
	CM at Risk	0.006	Significant	--
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

Per the comparative boxplot below, there is no immediately evident delivery method which is better in performance in terms of overall change due to quality/value reevaluation. It is, however, evident that DB has a slightly lower variability than other methods.

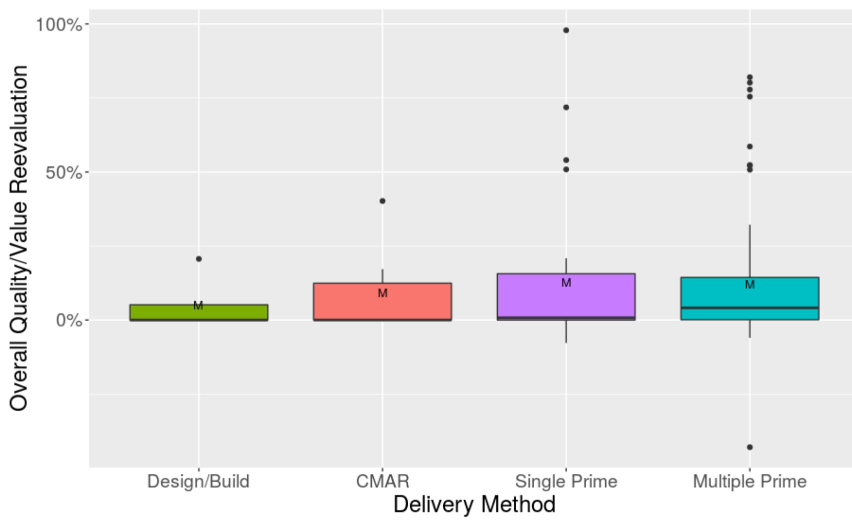


Figure 120: Boxplot of Overall Changes Due to Quality Reevaluation

Per the table of descriptive statistics below, the median values of all four delivery methods are similar: 0, 0, 0.8 and 4.1%. Standard deviations for DB, CMAR, SP, and MP were similar as well (10.3, 15.1, 24.9, and 22.4% respectively). This suggests no difference in performance between the different delivery methods, but more robust statistical testing is needed.

Table 120: Descriptive Statistics for Overall Changes Due to Quality Reevaluation Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	5.2%	9.3%	12.8%	12.0%
Maximum	20.7%	40.2%	97.9%	82.0%
75% Percentile	15.5%	17.2%	15.7%	14.7%
Median	0.0%	0.0%	0.8%	4.1%
25% Percentile	0.0%	0.0%	-0.1%	0.1%
Minimum	0.0%	-0.1%	-7.7%	-42.9%
Standard Deviation	10.3%	15.1%	24.9%	22.4%

Robust statistical tests were performed, with the following results. Fligner-Killeen returned a *p*-value of 0.339, and Kruskal-Wallis returned a *p*-value of 0.653, which failed to reject the null hypotheses of equal variance and equal population of medians at the 95 and 90% confidence level. Therefore, the post-hoc test was not performed.

Table 121: Significance of Variance & Distribution for Overall Changes Due to Quality Reevaluation

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.339	Not Significant	Not Significant
Kruskal-Wallis	Overall	0.653	Not Significant	Not Significant
Conover-Iman	DB vs Single	NA	NA	NA
	DB vs Multiple	NA	NA	NA
	CMAR vs Single	NA	NA	NA
	CMAR vs Multiple	NA	NA	NA
	DB vs CMAR	NA	NA	NA
	Single vs Multiple	NA	NA	NA

5.2.3.4.2 Comparative Analysis of Delivery Type

A secondary analysis of delivery type was performed to assess if there was any correlation between that variable and change orders due to reevaluation. Once again, normality was tested and rejected at the 95% confidence level, per the Q-Q plot and table below.

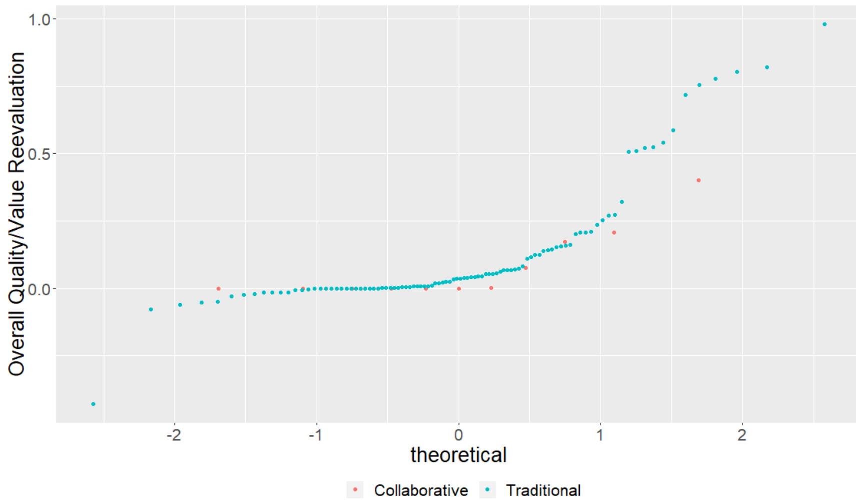


Figure 85: Q-Q Plot of Overall Changes Due to Quality Reevaluation

Table 122: Significance of Normality for Overall Changes Due to Quality Reevaluation

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.000	Significant	--
	Traditional	0.000	Significant	--

The comparative boxplot drawn below shows no immediately evident difference between delivery types. Traditional delivery types were evidently more variable due to outliers.

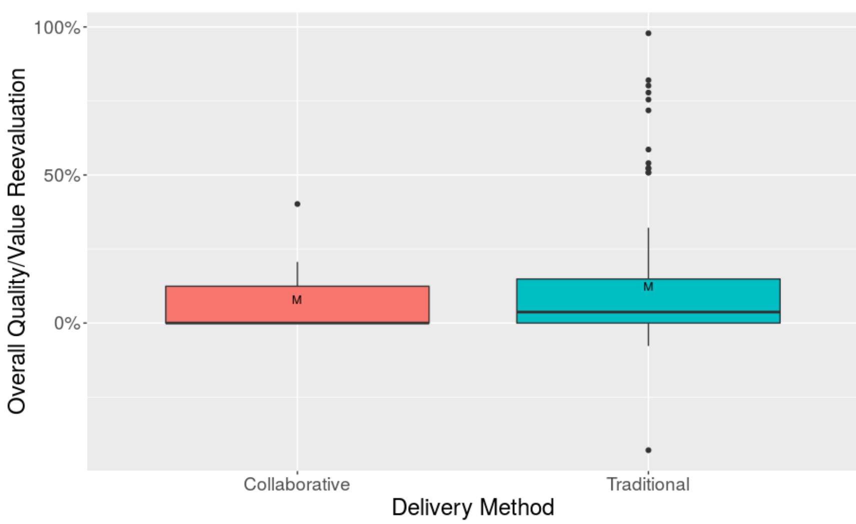


Figure 121: Boxplot of Overall Changes Due to Quality Reevaluation

The following table of descriptive statistics shows that the median values were similar between Collaborative and Traditional delivery (0 and 3.7%). Variability was slightly different, with standard deviations of 13.2 and 23.0%. More robust testing was needed.

Table 123: Descriptive Statistics for Overall Changes Due to Quality Reevaluation Per PDT

Statistical Test	Collaborative	Traditional
Mean	7.8%	12.3%
Maximum	40.2%	97.9%
75% Percentile	17.2%	15.3%
Median	0.0%	3.7%
25% Percentile	0.0%	0.0%
Minimum	-0.1%	-42.9%
Standard Deviation	13.2%	23.0%

Fligner-Killeen test returned a *p*-value of 0.073 which was insufficient to reject the null hypothesis of equal variance among delivery types at the 95% confidence level. This suggests higher variability in Traditional delivery.

Mann-Whitney U test returned a *p*-value of 0.219, which is insufficient to reject the null hypothesis of equal medians at either the 95 or 90% confidence levels.

Table 124: Significance of Variance & Distribution for Overall Changes Due to Quality Reevaluation

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.073	Not Significant	Significant
Mann–Whitney	Overall	0.470	Not Significant	Not Significant

5.2.3.5 Program Work Reinstated

5.2.3.5.1 Comparative Analysis of Delivery Methods

First, normality was assessed and rejected at the 95%, as shown in the Q-Q plots and table below.

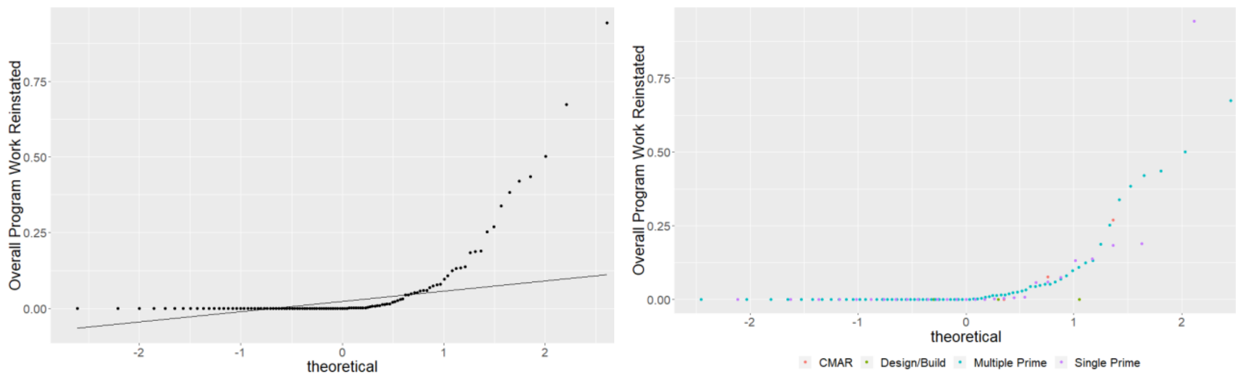


Figure 122: Q-Q Plot of Overall Changes Due to Program Reinstated

Table 125: Significance of Normality for Overall Changes Due to Program Reinstated

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build*	NA	NA	NA
	CM at Risk	0.000	Significant	--
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

*all the values are equivalent. Normality cannot be tested

Per the comparative boxplot below, there is no visually evident difference between the different delivery methods. It is visually evident that there is similar variability between CMAR, SP and MP, which is higher than DB.

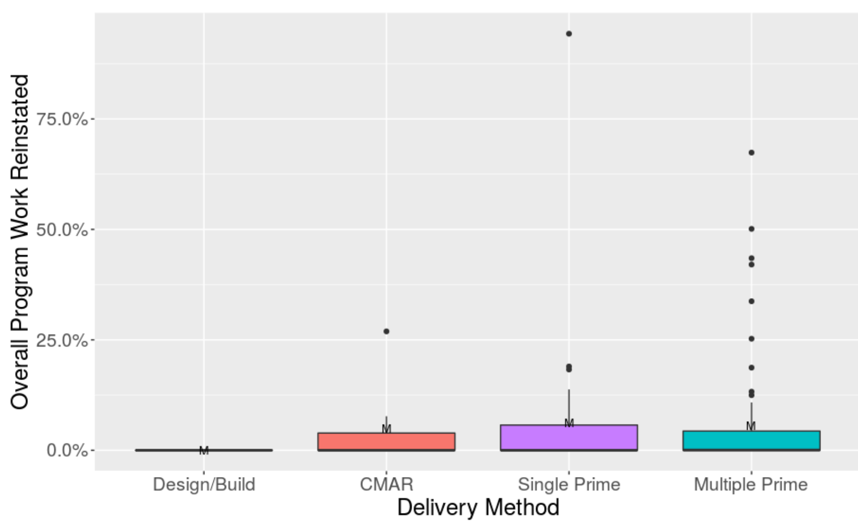


Figure 123: Boxplot of Overall Changes Due to Program Reinstated

The following table of descriptive statistics was drawn to increase the robustness of the conclusions. The median values of each delivery method were: 0%, 0%, 0% and 0% respectively. These statistics suggest no difference between the delivery methods but will still require further statistical tests. CMAR, SP and MP, as visually hypothesized, had similar standard deviations at 10.1, 17.9 and 12.9 percent, while DB at 0.0%.

Table 126: Descriptive Statistics for Overall Changes Due to Program Reinstated Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	0.0%	5.0%	6.2%	5.6%
Maximum	0.0%	26.9%	94.3%	67.4%
75% Percentile	0.0%	7.7%	5.8%	4.5%
Median	0.0%	0.0%	0.0%	0.0%
25% Percentile	0.0%	0.0%	0.0%	0.0%
Minimum	0.0%	0.0%	0.0%	0.0%
Standard Deviation	0.0%	10.1%	17.9%	12.9%

More robust tests were performed to quantitatively confirm the previous findings. Fligner-Killeen returned a *p*-value of 0.034, which was sufficient to reject the null hypothesis of equal variance at the 95% confidence level. Kruskal-Wallis, furthermore, returned a *p*-value of 0.260 which failed to reject the null hypothesis of equal population medians among delivery methods at the 95% confidence level. Accordingly, Conover-Iman post-hoc test was performed.

Table 127: Significance of Variance & Distribution for Overall Changes Due to Program Reinstated

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.034	Significant	--
Kruskal-Wallis	Overall	0.260	Not Significant	Not Significant
Conover-Iman	DB vs Single	NA	NA	NA
	DB vs Multiple	NA	NA	NA
	CMAR vs Single	NA	NA	NA
	CMAR vs Multiple	NA	NA	NA
	DB vs CMAR	NA	NA	NA
	Single vs Multiple	NA	NA	NA

5.2.3.5.2 <u>Comparative Analysis of Delivery Type</u>

A second quantitative assessment was performed to study the impact of each delivery method. The Q-Q plot and table below show a clear departure from normality at the 95% confidence level.

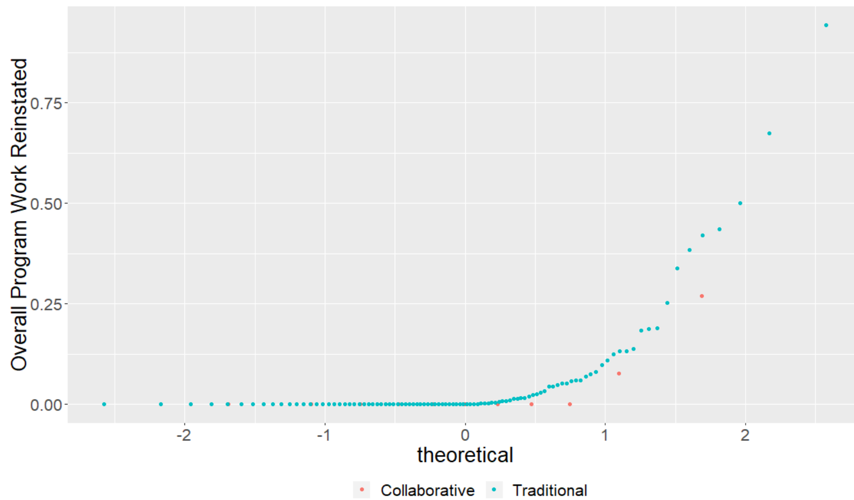


Figure 124: Q-Q Plot of Overall Changes Due to Program Reinstated

Table 128: Significance of Normality for Overall Changes Due to Program Reinstated

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.000	Significant	--
	Traditional	0.000	Significant	--

The comparative boxplot below shows no visually evident difference between delivery types. In terms of variability, Collaborative was much lower than Traditional.

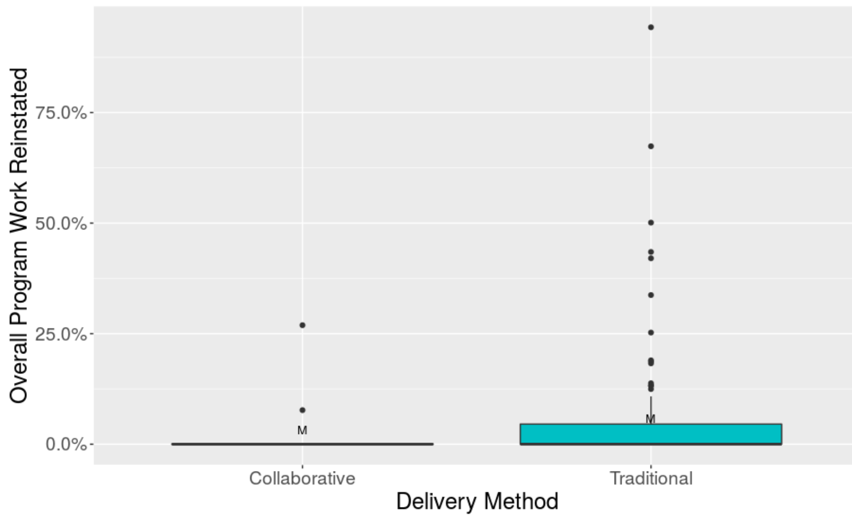


Figure 125: Boxplot of Overall Changes Due to Program Reinstated

Therefore, the descriptive statistical table below was drawn to further characterize these findings. No clear difference in the median of each delivery method was found, with both methods at 0%. The mean values of the delivery methods were also highly similar, 3.2 and 5.8%, respectively. Collaborative methods also experienced similar standard deviation to Traditional (8.2 and 14.4%). Additional analysis was performed.

Table 129: Descriptive Statistics for Overall Changes Due to Program Reinstated Per PDT

Statistical Test	Collaborative	Traditional
Mean	3.2%	5.8%
Maximum	26.9%	94.3%
75% Percentile	0.1%	4.7%
Median	0.0%	0.0%
25% Percentile	0.0%	0.0%
Minimum	0.0%	0.0%
Standard Deviation	8.2%	14.4%

In this analysis, Fligner-Killeen returned a *p*-value of 0.016, which is sufficient to reject the null hypothesis of equal variance among delivery types at the 95% confidence level. Mann-Whitney U test returned a *p*-value of 0.160, which failed to reject the null hypothesis of equal medians among delivery types at either the 95 or 90% confidence levels.

Table 130: Significance of Variance & Distribution for Overall Changes Due to Program Reinstated

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.016	Significant	--
Mann–Whitney	Overall	0.160	Not Significant	Not Significant

5.2.3.6 Other Reasons

5.2.3.6.1 Comparative Analysis of Delivery Methods

Normality was initially assessed and rejected at the 95% confidence level, as shown in the following Q-Q plots and table.

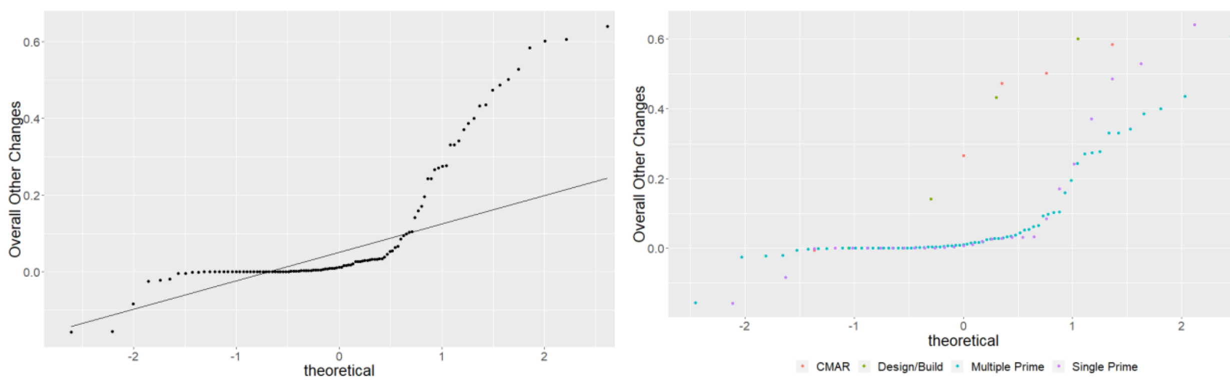


Figure 126: Q-Q Plot of Overall Changes Due to Other Reasons

Table 131: Significance of Normality for Overall Changes Due to Other Reasons

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.741	Not Significant	Not Significant
	CM at Risk	0.078	Not Significant	Significant
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

No clear difference was identified visually among delivery methods. DB and CMAR are visually more variable than SP and MP, as shown below.

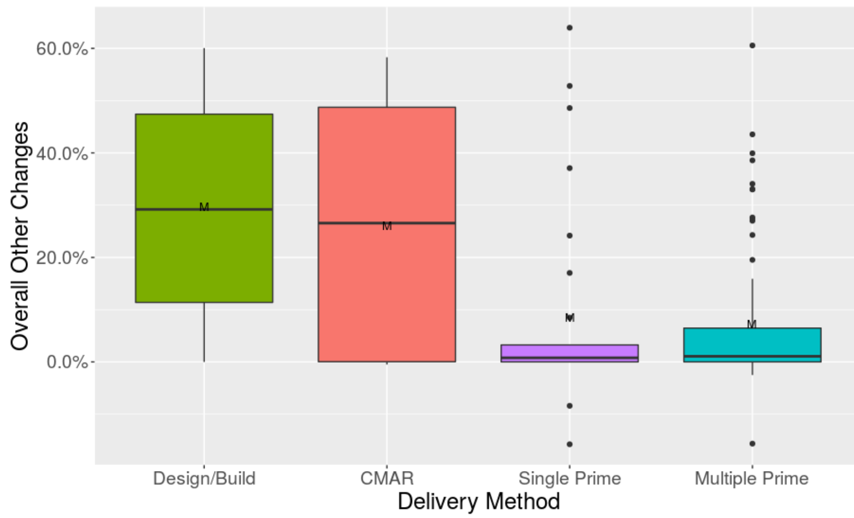


Figure 127: Boxplot of Overall Changes Due to Other Reasons

The table of descriptive statistics, below, highlights higher medians among DB and CMAR (29.2/26.6%), against SP and MP (0.8/1.1%), as well as the slightly higher standard deviations of DB and CMAR (27.1/26.2%) as compared to SP and MP (18.7/13.6%). Further tests were performed to confirm both these findings.

Table 132: Descriptive Statistics for Overall Changes Due to Other Reasons Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	29.6%	26.0%	8.5%	7.2%
Maximum	60.1%	58.3%	64.0%	60.6%
75% Percentile	55.8%	50.2%	5.9%	7.3%
Median	29.2%	26.6%	0.8%	1.1%
25% Percentile	3.8%	0.0%	0.0%	0.0%
Minimum	0.0%	-0.5%	-15.8%	-15.6%
Standard Deviation	27.1%	26.2%	18.7%	13.6%

In this secondary statistical testing, Fligner-Killen returned a *p*-value of 0.046 which is sufficient to reject the null hypothesis at the 95% confidence level and confirm higher variability in DB and CMAR.

Kruskal-Wallis, however, returned a *p*-value of 0.285, which is insufficient to reject the null hypothesis of equal population medians at either the 95 or 90% confidence levels. Ergo no post hoc testing was performed.

Table 133: Significance of Variance & Distribution for Overall Changes Due to Other Reasons

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.046	Significant	--
Kruskal-Wallis	Overall	0.285	Not Significant	Not Significant
Conover-Iman	DB vs Single	NA	NA	NA
	DB vs Multiple	NA	NA	NA
	CMAR vs Single	NA	NA	NA
	CMAR vs Multiple	NA	NA	NA
	DB vs CMAR	NA	NA	NA
	Single vs Multiple	NA	NA	NA

5.2.3.6.2 <u>Comparative Analysis of Delivery Type</u>

A secondary assessment was performed that contrasted performance by delivery type. First, normality was assessed and rejected at the 95% confidence level, as shown in the Q-Q plot and table below.

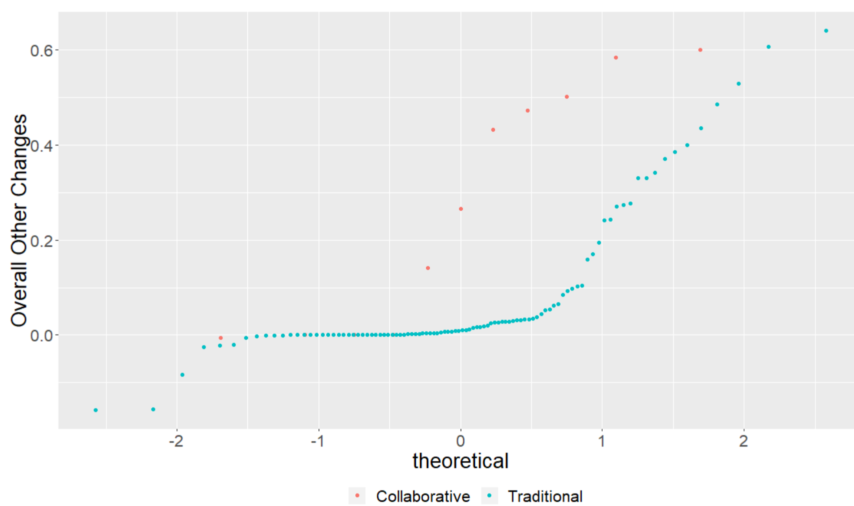


Figure 128: Q-Q Plot of Overall Changes Due to Other Reasons

Table 134: Significance of Normality for Overall Changes Due to Other Reasons

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.034	Significant	--
	Traditional	0.000	Significant	--

The boxplots were then subsequently drawn to highlight visual differences between delivery types. Collaborative delivery methods were notably more variable with higher median value than Traditional ones.

Figure 129: Boxplot of Overall Changes Due to Other Reasons

To create more robust statistical observations, the following table of descriptive statistics was drawn. The median data, as evident, was much higher in Collaborative methods than Traditional at 26.6 and 1.0%, respectively. Collaborative methods also experienced a higher level of variability, with standard deviation of 25.2%, which is notably higher than Traditional delivery methods (15.2%).

Table 135: Descriptive Statistics for Overall Changes Due to Other Reasons Per PDT

Statistical Test	Collaborative	Traditional
Mean	27.3%	7.6%
Maximum	60.1%	64.0%
75% Percentile	50.2%	6.5%
Median	26.6%	1.0%
25% Percentile	0.0%	0.0%
Minimum	-0.5%	-15.8%
Standard Deviation	25.2%	15.2%

Fligner-Killeen and Mann-Whitney U tests were performed to confirm the previous findings and claims. Fligner-Killeen test returned a *p*-value of 0.004, sufficient to reject the null hypothesis of equal variance among delivery types and suggest higher variability for Collaborative delivery types. Mann-Whitney U test, however, returned a *p*-value of 0.067, which is insufficient to reject the null hypothesis of equal medians among the two delivery types at the 95% confidence level.

Table 136: Significance of Variance & Distribution for Overall Changes Due to Other Reasons

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.004	Significant	--
Mann–Whitney	Overall	0.067	Not Significant	Significant

5.2.4 Section Summary

In this section, the likelihood of occurrence of each of the six reasons allowed by the State for issuing change orders were analyzed. The six reasons where studied for each of the cost-related, schedule-related and overall change orders to drive performance conclusions and identify which delivery methods are most impacted by each change order reason. Nevertheless, due to latent ambiguity in the definitions of some change orders' reasons including Quality/Value Reevaluation, Program Work Reinstated and Other, in addition to the minor portion of those reasons, no solid conclusions were drawn from the conducted analysis of these three specific change reasons.

In regard to cost-related change orders, the analysis showed that Unforeseen Conditions, Design Oversight and Scope Changes accounted for more than 76%. When statistically examined, Collaborative delivery methods experienced significantly lower percentage of Unforeseen Conditions at the 95% confidence level with an average of 0.8% as compared to Traditional ones with an average of 23.4%. Similarly, Collaborative delivery methods experienced significantly lower percentage of Design Oversight with an average of 0.2% as compared to Traditional ones with an average of 19.2%. As for Scope changes, no statistical evidence supported any significant difference between Collaborative and Traditional delivery types.

As for schedule-related change orders, the analysis has shown that Unforeseen Conditions, Design Oversight and Scope Changes accounted for around 69%. When statistically examined, Collaborative delivery methods experienced significantly lower percentage of Unforeseen Conditions at the 95% confidence level with an average of 7.1% as compared to Traditional ones with an average of 25%. Similarly, Collaborative delivery methods experienced significantly lower percentage of Design Oversight with an average of 0.0% as compared to Traditional (10.7%). As for Scope changes, no statistical evidence supported any significant difference between Collaborative and Traditional delivery types.

Considering the overall change orders, the analysis has shown that Unforeseen Conditions, Design Oversight and Scope Changes accounted for around 73%. When statistically examined, Collaborative delivery methods experienced significantly lower percentage of Unforeseen Conditions at the 95% confidence level with an average of 3.9% as compared to Traditional ones with an average of 23.9%. Similarly, Collaborative delivery methods experienced significantly lower percentage of Design Oversight with an average of 0.1% as compared to Traditional (13.8%). As for Scope changes, again no statistical evidence supported any significant difference between Collaborative and Traditional delivery types. Table 137 summarizes the findings of overall changes.

Table 137: Statistical Summary of Overall Changes by Delivery Methods & Types

Statistical Test	Parameter	*Unforeseen Conditions*	Design Oversight	Scope Changes	Quality Reevaluation	Work Reinstated	Other
Shapiro-Wilk	Overall	**	**	**	**	**	**
Mann-Whitney	Collaborative vs Traditional	**	**	*			*
Kruskal-Wallis	Overall	**	**				
Conover-Iman	DB vs Single	**	**				
	DB vs Multiple	**	**				
	CMAR vs Single	**	**				
	CMAR vs Multiple	**	**				
	DB vs CMAR						
	Single vs Multiple						

*statistically significant difference at the 90% confidence level (*p*-value <0.10)
**statistically significant difference at the 95% confidence level (*p*-value <0.05)

The above conclusions are substantial for various reasons. Although one would expect unforeseen conditions to be independent of the delivery method, it is believed that the better performance of Collaborative projects is driven by the project's participants ability to coordinate in order to identify and resolve those unanticipated issues much more effective compared to those operating under the Traditional structure.

Conversely, although one would anticipate Collaborative projects to experience significantly higher scope changes due to the fact that the general contractor are typically engaged early on in the process when the design is 0-20% complete (under DB approach) or 20-40% complete (under CMAR approach), the results did not present any significant difference at either the 95% of the 90% confidence levels for either cost-related, schedule-related or overall change orders. The findings provide an empirical evidence of the Collaborative project's ability to reduce scope-related change orders, even with the scope being fairly defined. Arguably, this might be due to the relatively high accuracy in defining the project scope generated by the mutual collaboration of project parties, or due to the higher emphasis on specialty and past experience during the selection process, which is based on merit rather than solely lowest bid.

Further, the analysis showed that Collaborative projects experienced much lower changes due to Design Oversight compared to Traditional ones. It is believed that the higher design performance associated with Collaborative delivery methods stems from the higher collaboration level between project parties, which significantly reduces design-related changes during the construction phase. It is also believed that the early involvement of the contractor provides a more thorough constructability analysis and professional feedback to the design team.

Table 138 summarizes the comparative analysis between Collaborative and Traditional delivery types discussed in this section.

Table 138: Significance Summary of Collaborative vs Traditional Delivery Types

Change Reasons	Cost-Related Changes	Schedule-Related Changes	Overall Changes
Unforeseen Conditions	**	**	**
Design Oversight	**	**	**
Scope Changes			*
Quality/Value Reevaluation			
Program Work Reinstated	*		
Other Reasons		**	*

*statistically significant difference at the 90% confidence level (*p*-value <0.10)
**statistically significant difference at the 95% confidence level (*p*-value <0.05)

5.3 Comparable Delivery Method Performance Assessment

The performance of a given project can be subjective depending on the metrics used in the evaluation process. While ordinal data can be very useful in representing the individual's feedback and perception about the success of a given project, those are by nature less reliable and more prone to the subjectivity of the individual. Therefore, this research identified key performance areas and numerical performance metrics within those areas that allow for data-driven analysis to be conducted and also guarantee much higher confidence in the outcomes and conclusions.

With that, 4 performance areas and 8 associated performance metrics were identified a listed earlier in Chapter 3. Given that many other metrics exist in the literature for measuring quality, safety and customer satisfaction, those identified metrics were the only ones tracked by the State and hence, the only numerical metrics that could be used.

Through the coming sections, the performance of the four delivery methods will be analyzed and compared in light of the eight performance metrics in order to draw thoughtful conclusions and recommendations regarding each performance area.

5.3.1 Cost Performance Area

In this section, we study the overall performance of the Wisconsin State construction projects from a cost standpoint. To do that, the metric cost (in dollars) per gross square foot is used and compared to a standard benchmark gathered from RS Means 2018.

5.3.1.1 Unit Cost

Before evaluating the collected data set, RS Means 2018 was utilized as a reference to benchmark the projects average cost per gross square foot. With 11 facility type considered and 140 projects, the below chart shows that the cost per GSF to construct some facilities like Residence Halls, Academic Buildings and Detention Centers is less than the National Average values for those facilities. Whereas, the cost to construct the rest of the facilities including Treatment Facility, Office Buildings, Research Buildings, etc. is higher than the respective National Average values.

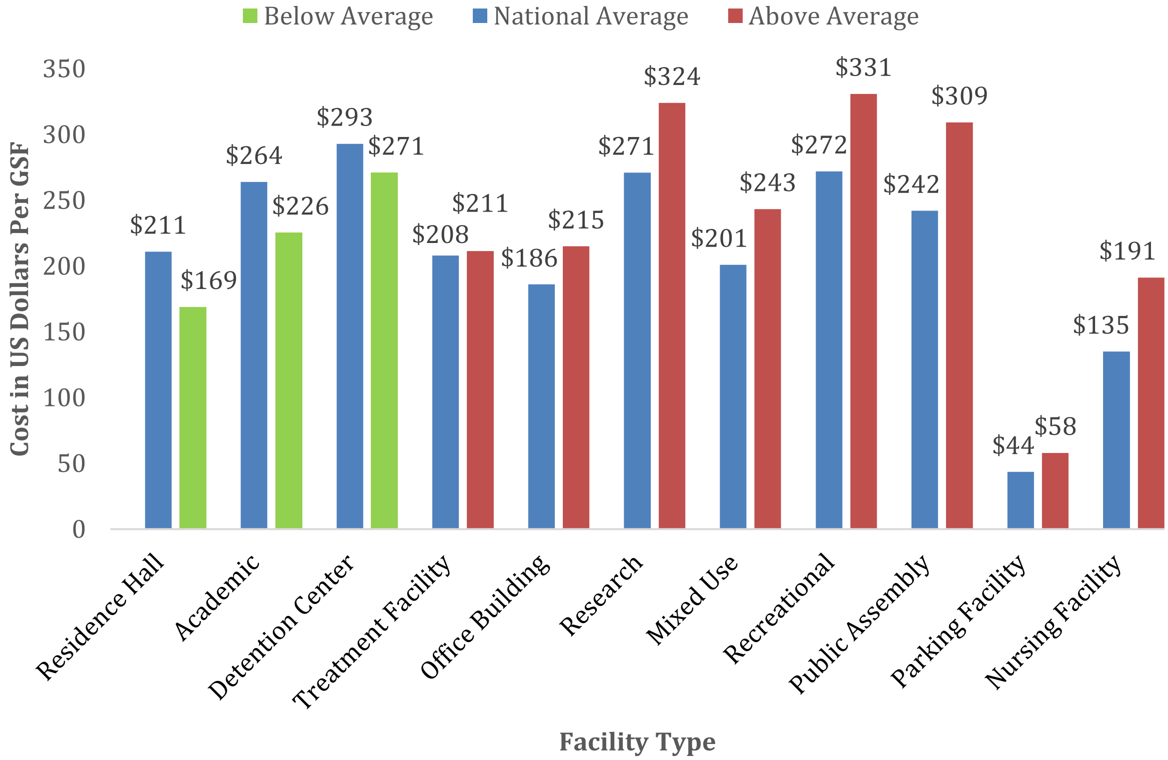


Figure 130: Unit Cost of State Projects Against the National Average

To put these differences into perspective, the chart below has been developed to represent the Percent Difference between the facility's cost per GSF against its contemporary benchmark. The results show that the first 3 facility types cost approximately 14% less than the National Average with the lowest difference being Detention Centers at 7% and the highest difference occurring for Residence Halls at 20%.

The other eight facility types cost approximately 23% above the National Average with the lowest difference being Treatment Facilities at 2% and highest occurring for Nursing Facilities at 42%. All in all, the overall average construction cost per GSF for all facility types combined is 13% above the National Average. It is worth mentioning that RS Means computes the National Average values using data from 30 major U.S. cities, including San Francisco, Los Angeles, Washington D.C. and others (review Chapter 1 for a list of the 30 cities).

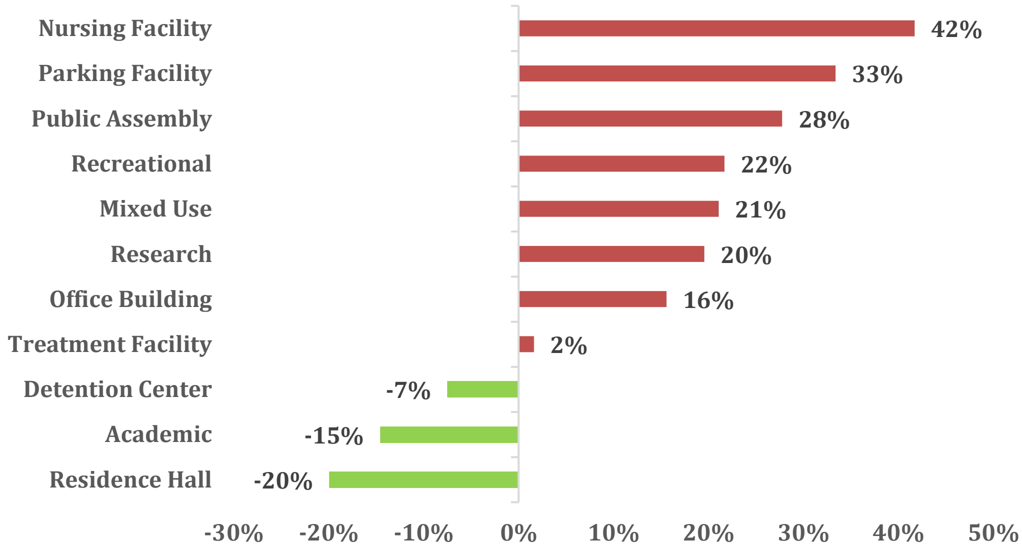


Figure 131: Unit Cost of Wisconsin State Projects Against the National Average

5.3.2 Schedule Performance Area

In this section, we study and compare the performance of each delivery method in terms of its ability to abide by the contractual construction duration. In that sense, two schedule performance metrics are used: Schedule Growth which measures project's delay as compared to the initial contractual duration, and Schedule Factor which measures project's delay as compared to the contractual duration plus the approved time extensions.

5.3.2.1 Schedule Growth

Schedule Growth is a metric that is used to measure the percentage of delay of a project as compared to the Contractual Construction Duration. While this metric does not incorporate the approved time extensions, it is still very informative from an end user's perspective. Some delays are excusable under the contract but the impact on a researcher due to a delayed testing lab for instance is essential to consider.

$$Schedule\ Growth = \frac{Actual\ Const.Duration - Contractual\ Const.Duration}{Contractual\ Construction\ Duration}$$

5.3.2.1.1 <u>Comparative Analysis of Delivery Methods</u>

As evident from the Q-Q plots below, normality was not present. This was confirmed at the 95% confidence level by the Shapiro-Wilk test, as shown in the table below.

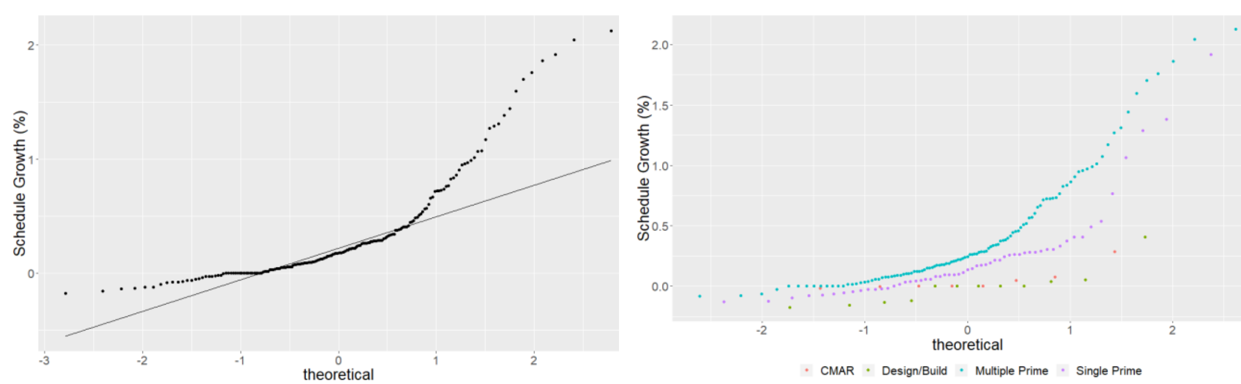


Figure 132: Q-Q Plot of Schedule Growth

Table 139: Significance of Normality for Schedule Growth

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.000	Significant	--
	CM at Risk	0.000	Significant	--
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

By considering the comparative boxplot below, there is a clear visual difference in performance between Collaborative and Traditional delivery methods, but there is no clear difference between DB and CMAR and only a slight one between SP and MP. Traditional delivery also exhibited higher variability, represented by the compaction level of each boxplot.

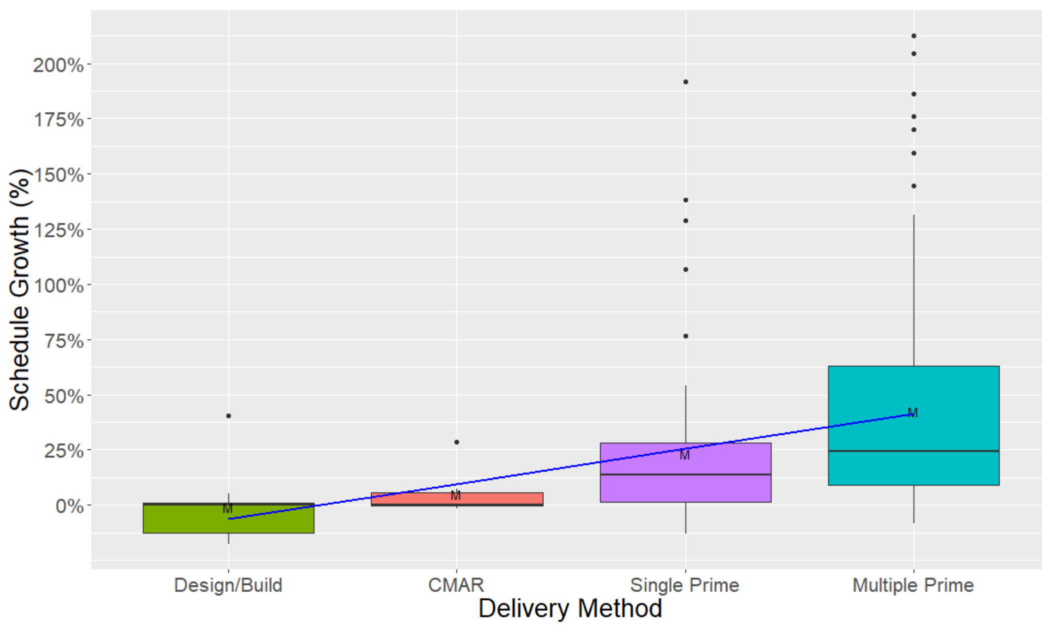


Figure 133: Boxplot of Schedule Growth Against PDMs

The table of descriptive statistics, below, was compiled to validate the observations. A clear difference in the median of Collaborative (0/0%) and Traditional (13.8/25.4%). The medians suggest no difference between DB and CMAR, and a slight difference between SP and MP, as was previously speculated. Further testing was performed to validate this assumption. Variability was higher in SP and MP (38.2/62.4%) than DB and CMAR (15.3/10.1%). Once again, statistical tests were performed to validate these findings.

Table 140: Descriptive Statistics for Schedule Growth Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	-0.7%	4.8%	23.3%	46.0%
Maximum	40.7%	28.7%	191.8%	461.7%
75% Percentile	1.0%	5.4%	28.0%	65.9%
Median	0.0%	0.0%	13.8%	25.4%
25% Percentile	-12.4%	-0.1%	1.3%	9.0%
Minimum	-17.7%	-1.6%	-13.2%	-8.3%
Standard Deviation	15.3%	10.1%	38.2%	62.4%

Fligner-Killeen returned a *p*-value of 0.000, confirming the inequal variance among delivery methods at the 95% confidence level and suggesting a higher variability in SP and MP than in DB and CMAR.

Kruskal-Wallis further returned a *p*-value of 0.000, rejecting the assumption of equal medians at the 95% confidence level. Conover-Iman post-hoc test showed statistically significant differences between DB and SP/MP, CMAR and MP and between SP and MP, with no difference shown between DB and CMAR or CMAR and SP.

Table 141: Significance of Distribution for Schedule Growth Per PDM

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Kruskal-Wallis	Overall	0.000	Significant	--
Conover-Iman	DB vs Single	0.007	Significant	--
	DB vs Multiple	0.000	Significant	--
	CMAR vs Single	0.162	Not Significant	Not Significant
	CMAR vs Multiple	0.002	Significant	--
	DB vs CMAR	0.871	Not Significant	Not Significant
	Single vs Multiple	0.004	Significant	--

Yet further evidence was found in this section of the better performance of DB and CMAR over SP and MP, this time in terms of schedule growth. That is, DB and CMAR provide better performance over schedule than DBB projects, with reduced variability.

5.3.2.1.2 Comparative Analysis of Delivery Types

Once again, normality was assessed, and a lack of normality was found at the 95% confidence level, as shown in the figure and table below.

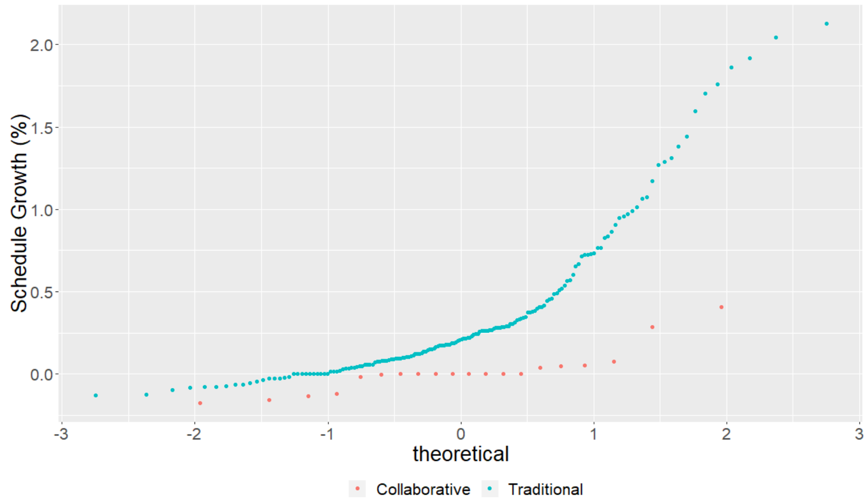


Figure 134: Q-Q Plot of Schedule Growth

Table 142: Significance of Normality for Schedule Growth by Delivery Type

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.000	Significant	--
	Traditional	0.000	Significant	--

The following boxplot highlights a clear difference between Collaborative and Traditional delivery types. Higher variability was also noted in Traditional methods.

Figure 135: Boxplot of Schedule Growth Against PDTs

Thus, the following table of descriptive statistics was created, which highlights differences between the medians of Collaborative and Traditional methods (0 and 21.1%) and their means (1.5 and 38.4%). Furthermore, Collaborative delivery methods exhibited lower variability (13.5%) than Traditional ones (56.3%).

Table 143: Descriptive Statistics for Schedule Growth Per PDT

Statistical Test	Collaborative	Traditional
Mean	1.5%	38.4%
Maximum	40.7%	461.7%
75% Percentile	4.1%	45.9%
Median	0.0%	21.2%
25% Percentile	-0.8%	5.7%
Minimum	-17.7%	-13.2%
Standard Deviation	13.5%	56.3%

To validate the preceding suggestions, statistical tests were performed. Fligner-Killeen returned a p-value of 0.000, sufficient to reject the null hypothesis at 95% confidence level. This suggests a higher variability for Traditional delivery types than Collaborative ones.

Mann-Whitney U tests also returned a p-value of 0.000, sufficient to reject the null hypothesis of equal medians, once again at the 95% confidence levels.

Table 144: Significance of Variance & Distribution for Schedule Growth

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Mann–Whitney	Overall	0.000	Significant	--

These findings confirm the previous conclusion of the better performance of Collaborative delivery types over Traditional ones, as well as elaborate the higher risk associated with Traditional projects compared to Collaborative ones explained by the higher variability.

5.3.2.2 Schedule Factor

Schedule factor is a metric that is used to measure the percentage of delay of a project as compared to the Extended Contractual Construction Duration. This metric is very effective in delivering an accurate understanding of the performance of each delivery method as it incorporates excusable approved extensions. Along with schedule growth, schedule factor examines the schedule performance from a different perspective, which comprehends that of the schedule growth.

$$Schedule\ Factor = \frac{Actual\ Construction\ Duration}{Extended\ Contractual\ Const.Duration + Total\ Approved\ Extensions}$$

5.3.2.2.1 <u>Comparative Analysis of Delivery Methods</u>

As evident from the Q-Q plots below, normality was not present. This was confirmed at the 95% confidence level by the Shapiro-Wilk test, as shown in the table below.

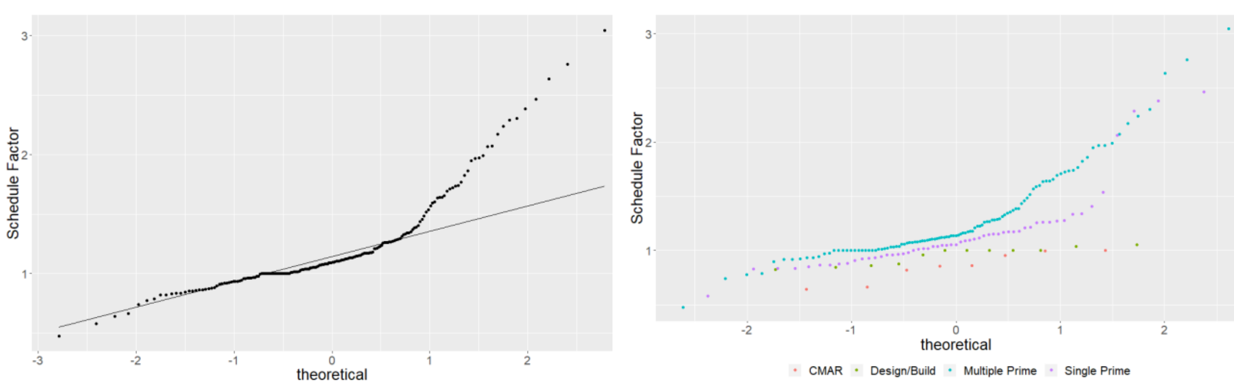


Figure 136: Q-Q Plot of Schedule Factor

Table 145: Significance of Normality for Schedule Factor

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.043	Significant	--
	CM at Risk	0.241	Not Significant	Not Significant
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

By considering the comparative boxplot below, there is a clear visual difference in performance between Collaborative and Traditional delivery methods, but there is no clear difference between DB and CMAR and only a slight one between SP and MP. MP delivery also exhibited higher variability.

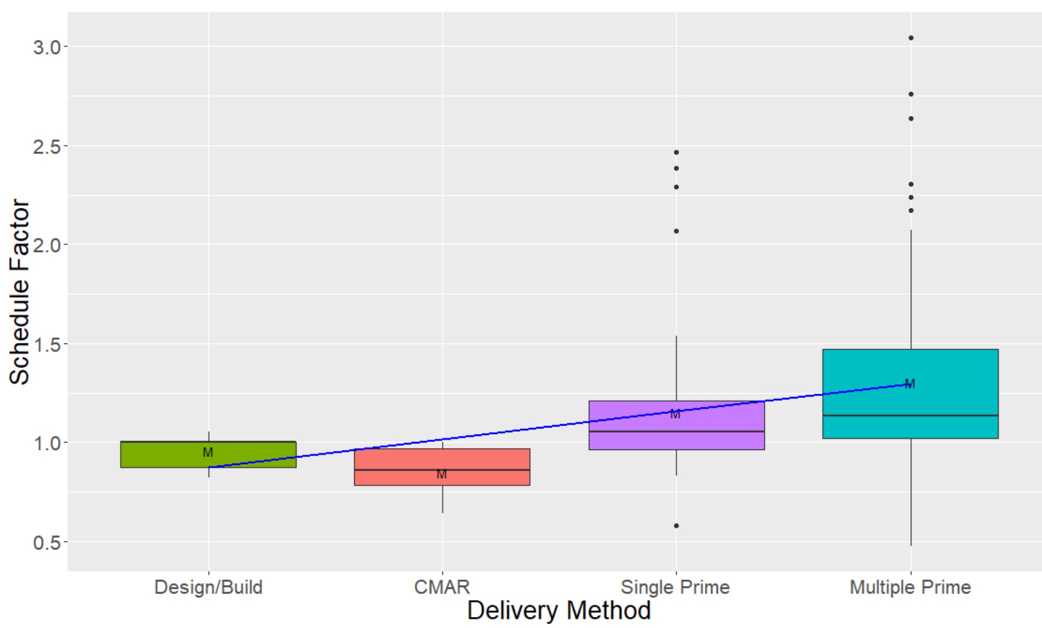


Figure 137: Boxplot of Schedule Factor Against PDMs

The table of descriptive statistics, below, was compiled to validate the observations. A clear difference in the median of Collaborative (1.0/0.86) and Traditional (1.06/1.14). The medians suggest no difference between DB and CMAR, and a slight difference between SP and MP, as was previously speculated. Further testing was performed to validate this assumption. Variability was higher in SP and MP (0.36/0.56) than DB and CMAR (0.08/0.14). Once again, statistical tests were performed to validate these findings.

Table 146: Descriptive Statistics for Schedule Factor Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	0.96	0.85	1.15	1.34
Maximum	1.05	1.00	2.47	5.21
75% Percentile	1.00	0.97	1.21	1.49
Median	1.00	0.86	1.06	1.14
25% Percentile	0.88	0.78	0.96	1.02
Minimum	0.82	0.64	0.58	0.48
Standard Deviation	0.08	0.14	0.36	0.56

Fligner-Killeen returned a *p*-value of 0.006, confirming the inequal variance among delivery methods at the 95% confidence level and suggesting a higher variability in SP and MP than in DB and CMAR.

Kruskal-Wallis further returned a *p*-value of 0.000, rejecting the assumption of equal medians at the 95% confidence level. Conover-Iman post-hoc test showed statistically significant differences between DB and SP/MP, CMAR and SP/MP and between SP and MP, with no difference shown between DB and CMAR.

Table 147: Significance of Distribution for Schedule Factor

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.006	Significant	--
Kruskal-Wallis	Overall	0.000	Significant	--
Conover-Iman	DB vs Single	0.020	Significant	--
	DB vs Multiple	0.000	Significant	--
	CMAR vs Single	0.003	Significant	--
	CMAR vs Multiple	0.000	Significant	--
	DB vs CMAR	0.701	Not Significant	Not Significant
	Single vs Multiple	0.008	Significant	--

Further evidence was found in this section of the better performance of DB and CMAR over SP and MP, this time in terms of schedule factor. That is DB and CMAR provide better performance over schedule than DBB projects, with reduced variability.

5.3.2.2.2 Comparative Analysis of Delivery Types

Normality was assessed, and a lack of normality was found at the 95% confidence level, as shown in the figure and table below.

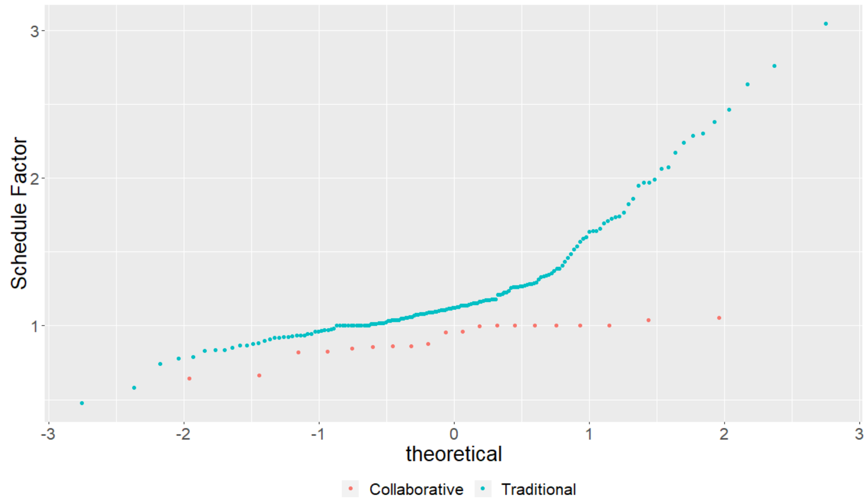


Figure 138: Q-Q Plot of Schedule Factor

Table 148: Significance of Normality for Schedule Factor

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.010	Significant	--
	Traditional	0.000	Significant	--

The following boxplot highlights a clear difference between Collaborative and Traditional delivery types. Higher variability was also noted in Traditional methods.

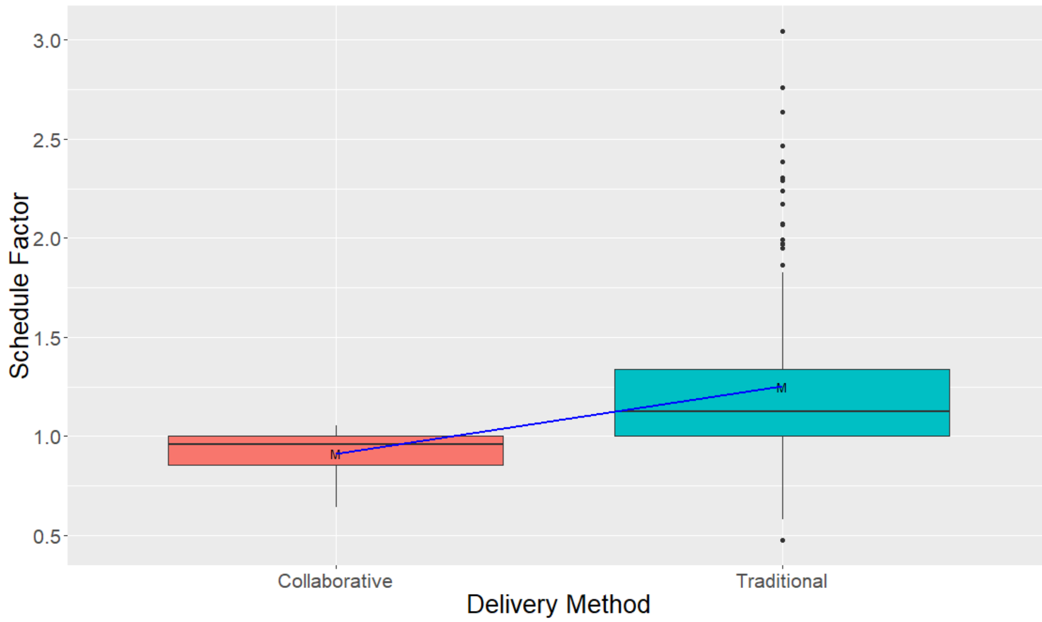


Figure 139: Boxplot of Schedule Factor Against PDTs

Thus, the following table of descriptive statistics was created, which highlights differences between the medians of Collaborative and Traditional methods (0.96 and 1.12) and their means (0.91 and 1.27). Furthermore, Collaborative delivery methods exhibited lower variability (0.12) than Traditional ones (0.51).

Table 149: Descriptive Statistics for Schedule Factor Per PDT

Statistical Test	Collaborative	Traditional
Mean	0.91	1.27
Maximum	1.05	5.21
75% Percentile	1.00	1.34
Median	0.96	1.12
25% Percentile	0.85	1.00
Minimum	0.64	0.48
Standard Deviation	0.12	0.51

To validate the preceding suggestions, statistical tests were performed. Fligner-Killeen returned a p-value of 0.005, sufficient to reject the null hypothesis at 95% confidence level. This suggests a higher variability for Traditional delivery types than Collaborative ones.

Mann-Whitney U tests also returned a p-value of 0.000, sufficient to reject the null hypothesis of equal medians, once again at the 95% confidence levels.

Table 150: Significance of Variance & Distribution for Schedule Factor

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.005	Significant	--
Mann–Whitney	Overall	0.000	Significant	--

These findings confirm the previous conclusion of the better performance of Collaborative delivery types over Traditional ones, as well as elaborate the higher risk associated with Traditional projects compared to Collaborative ones when it comes to schedule factor.

5.3.3 Communication Performance Area

In this section, we study and compare the performance of each delivery method in light of its communication and collaboration capacity among the different project parties. In that sense, two communication performance metrics are used: RFI Processing Time which measures the average number of days consumed before a contractor receives a final response to a request for information, and Number of RFIs Per Million Dollars which measures the intensity of RFIs for every million dollars of construction cost.

Although the response time and frequency of RFIs depend on various factors such as the level of complexity and the project phase during which an RFI was issued, it is assumed that the high volume of projects and respective RFIs (+47,000 RFIs) present in the dataset are sufficient to account for those intangible factors. In addition, multiple studies in the literature have also reported reliable results using these metrics (Navigant 2013; El Asmar et al. 2015; Ibrahim 2018). It is worth mentioning that while DB does not capture RFI-related activities as RFIs are typically handled by the design-builder, it was included in the analysis due to the generated savings to the Owner by delegating such RFI-related activities (Navigant 2013).

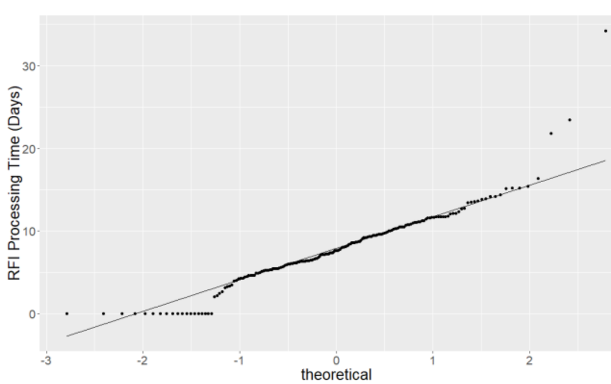
5.3.3.1 RFI Processing Time

RFI Processing Time is a metric that is used to measure the average number of days consumed before a contractor receives a final response to a request for information. This metric is appropriate in depicting the quality of communication between the Owner, GC and Architect.

$$RFI\ Processing\ Time = RFI\ Submission\ Date - RFI\ Final\ Response\ Date$$

5.3.3.1.1 <u>Comparative Analysis of Delivery Methods</u>

As evident from the Q-Q plots below, normality was not present. This was confirmed at the 95% confidence level by the Shapiro-Wilk test, as shown in the table below.



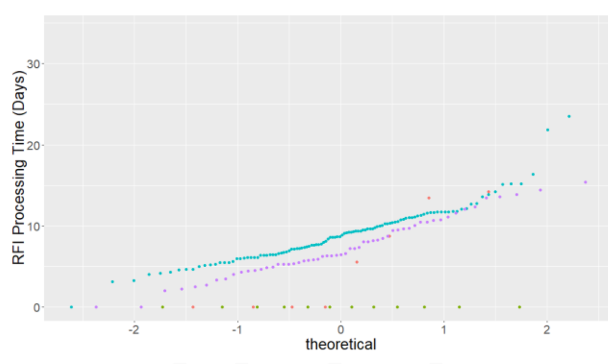


Figure 140: Q-Q Plot of RFI Processing Time

Table 151: Significance of Normality for RFI Processing Time

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build*	NA	NA	NA
	CM at Risk	0.028	Significant	--
	Single-Prime	0.496	Not Significant	Not Significant
	Multiple-Prime	0.000	Significant	--

*all the values are equivalent. Normality cannot be tested

By considering the comparative boxplot below, there is a clear visual difference in performance between Collaborative and Traditional delivery methods. CMAR delivery also exhibited higher variability.

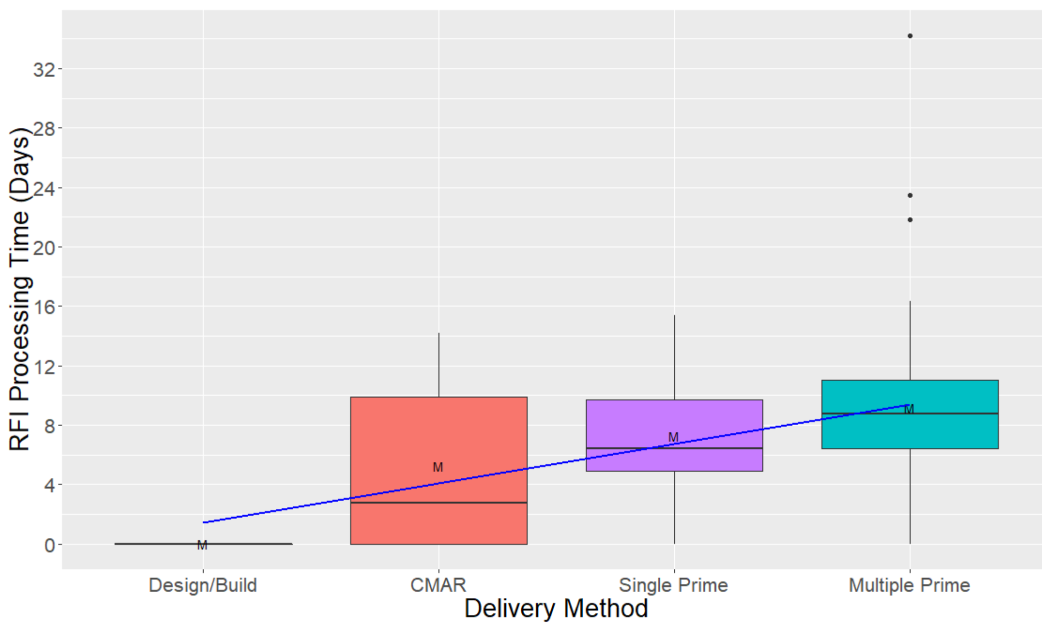


Figure 141: Boxplot of RFI Processing Time Against PDMs

The table of descriptive statistics, below, was compiled to validate the observations. A clear difference in the median of Collaborative (0/2.8) and Traditional (6.5/8.8). The medians suggest no difference between DB and CMAR, and a slight difference between SP and MP, as was previously speculated. Further testing was performed to validate this assumption. Variability was higher in CMAR (6.2) than DB, SP and MP (0.0/3.6/4.2). Once again, statistical tests were performed to validate these findings.

Table 152: Descriptive Statistics for RFI Processing Time Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	0.0	5.2	7.3	9.1
Maximum	0.0	14.2	15.4	34.2
75% Percentile	0.0	9.9	9.7	11.0
Median	0.0	2.8	6.5	8.8
25% Percentile	0.0	0.0	4.9	6.4
Minimum	0.0	0.0	0.0	0.0
Standard Deviation	0.0	6.2	3.6	4.2

Fligner-Killeen returned a *p*-value of 0.000, confirming the inequal variance among delivery methods at the 95% confidence level and suggesting a higher variability in CMAR than in DB, SP and MP.

Kruskal-Wallis further returned a *p*-value of 0.000, rejecting the assumption of equal medians at the 95% confidence level. Conover-Iman post-hoc test showed statistically significant differences between DB and SP/MP, DB and CMAR, and between SP and MP, with no difference shown between CMAR and SP/MP.

Table 153: Significance of Distribution for RFI Processing Time

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Kruskal-Wallis	Overall	0.000	Significant	--
Conover-Iman	DB vs Single	0.000	Significant	--
	DB vs Multiple	0.000	Significant	--
	CMAR vs Single	0.702	Not Significant	Not Significant
	CMAR vs Multiple	0.073	Not Significant	Significant
	DB vs CMAR	0.023	Significant	--
	Single vs Multiple	0.010	Significant	--

The previous informative findings provide in-depth inference about the better performance of DB over SP and MP as well as that of CMAR over MP, as well as the performance of SP over MP when it comes to RFI processing time.

5.3.3.1.2 Comparative Analysis of Delivery Types

again, normality was assessed, and a lack of normality was found at the 95% confidence level, as shown in the figure and table below.

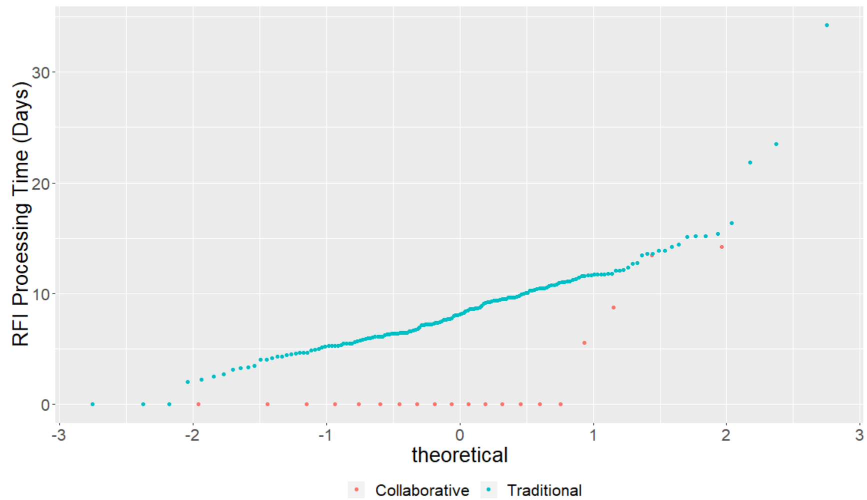


Figure 142: Q-Q Plot of RFI Processing Time

Table 154: Significance of Normality for RFI Processing Time

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.000	Significant	--
	Traditional	0.000	Significant	--

The following boxplot highlights a clear difference between Collaborative and Traditional delivery types. Higher variability was also noted in Traditional methods.

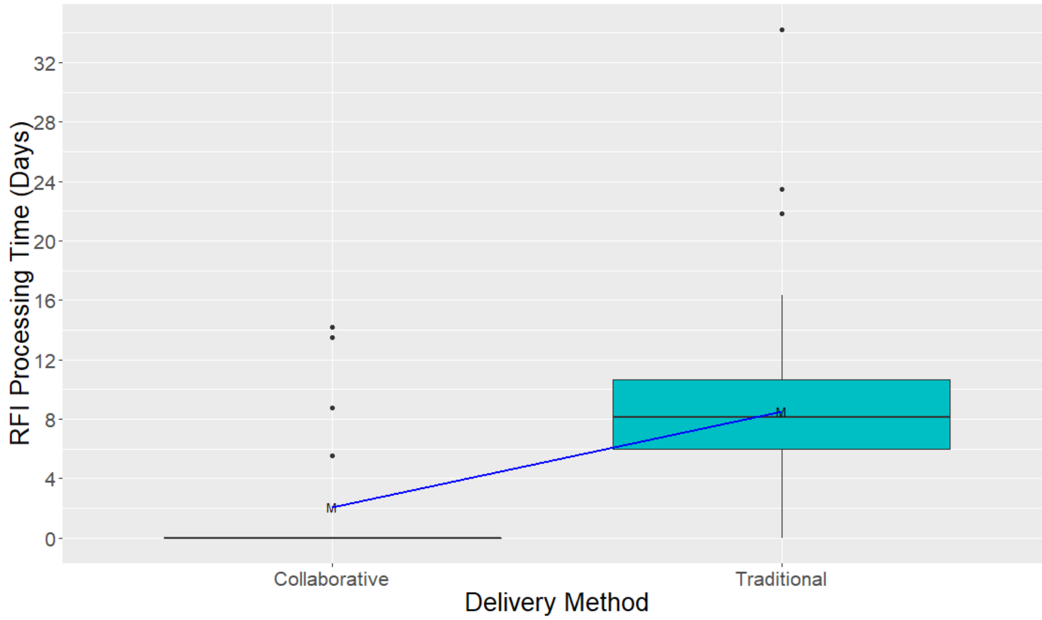


Figure 143: Boxplot of RFI Processing Time Against PDTs

Thus, the following table of descriptive statistics was created, which highlights differences between the medians of Collaborative and Traditional methods (0.0 and 8.1) and their means (2.1 and 8.5). Furthermore, Collaborative delivery methods exhibited similar variability compared to Traditional ones.

Table 155: Descriptive Statistics for RFI Processing Time Per PDT

Statistical Test	Collaborative	Traditional
Mean	2.1	8.5
Maximum	14.2	34.2
75% Percentile	0.0	10.7
Median	0.0	8.1
25% Percentile	0.0	6.0
Minimum	0.0	0.0
Standard Deviation	4.6	4.1

To validate the preceding suggestions, statistical tests were performed. Fligner-Killeen returned a *p*-value of 0.011, sufficient to reject the null hypothesis at 95% confidence level. This suggests a higher variability for Traditional delivery types than Collaborative ones.

Mann-Whitney U tests also returned a *p*-value of 0.000, sufficient to reject the null hypothesis of equal medians, once again at the 95% confidence levels.

Table 156: Significance of Variance & Distribution for RFI Processing Time

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.011	Significant	--
Mann–Whitney	Overall	0.000	Significant	--

The findings confirm the previous conclusion about the more robust performance of Collaborative delivery types over Traditional ones, as well as elaborate the higher risk associated with Traditional projects compared to Collaborative ones when it comes to RFI processing time from the owner perspective.

5.3.3.2 Number of RFIs Per Million Dollars

Number of RFIs per million dollars is a metric that is widely used to measure the intensity of RFIs for every million dollars of construction. Because each RFI consumes additional time for administrative and technical processing, RFIs represent extra cost to the project (Navigant 2013).

$$Number\ of\ RFIs\ Per\ Million\ Dollars = \frac{Number\ of\ RFIs}{Final\ Adjusted\ Construction\ Cost}$$

5.3.3.2.1 <u>Comparative Analysis of Delivery Methods</u>

As with the other analysis, the initial step was to determine normality. As evident from the Q-Q plots below, normality was not present. This was confirmed at the 95% confidence level by the Shapiro-Wilk test, as shown in the table below.

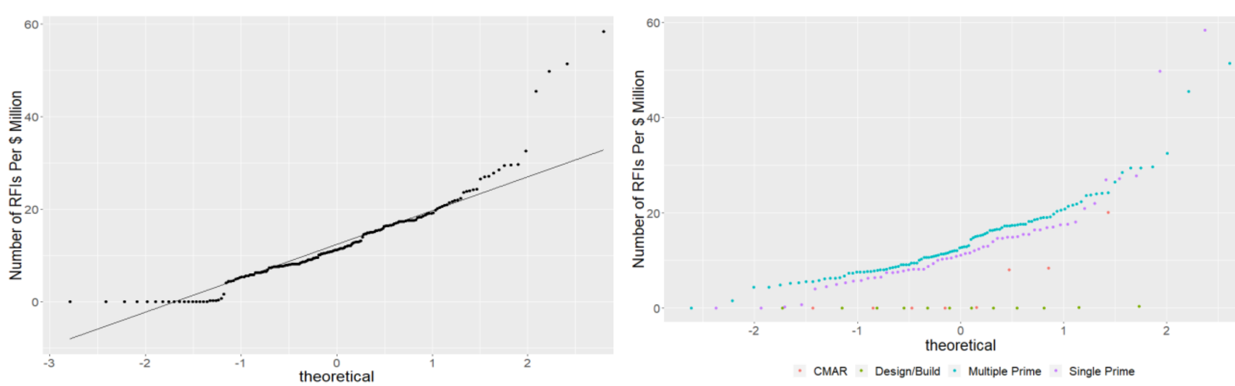


Figure 144: Q-Q Plot of Number of RFIs Per Million Dollar

Table 157: Significance of Normality for Number of RFIs Per Million Dollars

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build*	NA	NA	NA
	CM at Risk	0.000	Significant	--
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

*all the values are equivalent. Normality cannot be tested

By considering the comparative boxplot below, there is a clear visual difference in performance between Collaborative and Traditional delivery methods, but there is no clear difference between DB and CMAR and only a slight one between SP and MP. Traditional delivery also exhibited higher variability.

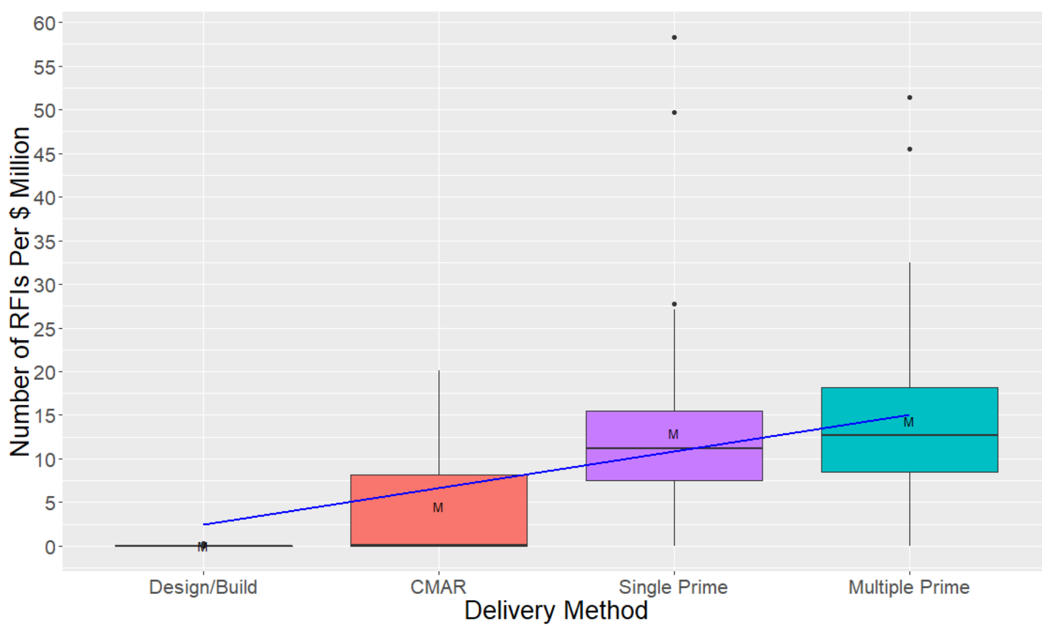


Figure 145: Boxplot of Number of RFIs Per Million Dollar Against PDMs

The table of descriptive statistics, below, was compiled to validate the observations. A clear difference in the median of Collaborative (0/0) and Traditional (11.1/12.7). The medians suggest no difference between DB and CMAR, and a slight difference between SP and MP, as was previously speculated. Further testing was performed to validate this assumption. Variability was higher in SP and MP (10.1/8.1) than DB and CMAR (0.0/7.3). Once again, statistical tests were performed to validate these findings.

Table 158: Descriptive Statistics for Number of RFIs Per Million Dollars Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	0.0	4.6	13.0	14.4
Maximum	0.0	20.1	58.3	51.4
75% Percentile	0.0	8.1	15.5	18.2
Median	0.0	0.0	11.1	12.7
25% Percentile	0.0	0.0	7.5	8.5
Minimum	0.0	0.0	0.0	0.0
Standard Deviation	0.0	7.3	10.1	8.1

Fligner-Killeen returned a *p*-value of 0.000, confirming the inequal variance among delivery methods at the 95% confidence level and suggesting a higher variability in SP and MP than in DB and CMAR.

Kruskal-Wallis further returned a *p*-value of 0.000, rejecting the assumption of equal medians at the 95% confidence level. Conover-Iman post-hoc test showed statistically significant differences between DB and SP/MP, and CMAR and SP/MP, with no difference shown between DB and CMAR, nor between SP and MP.

Table 159: Significance of Distribution for Number of RFIs Per Million Dollars

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Kruskal-Wallis	Overall	0.000	Significant	--
Conover-Iman	DB vs Single	0.000	Significant	--
	DB vs Multiple	0.000	Significant	--
	CMAR vs Single	0.021	Significant	--
	CMAR vs Multiple	0.001	Significant	--
	DB vs CMAR	0.326	Not Significant	Not Significant
	Single vs Multiple	0.186	Not Significant	Not Significant

The previous informative findings provide in-depth inference about the better performance of DB over SP and MP in the communication metrics, as well as that of CMAR over SP and MP when it comes to the number of RFIs per million dollars.

5.3.3.2.2 Comparative Analysis of Delivery Types

Once again, normality was assessed, and a lack of normality was found at the 95% confidence level, as shown in the figure and table below.

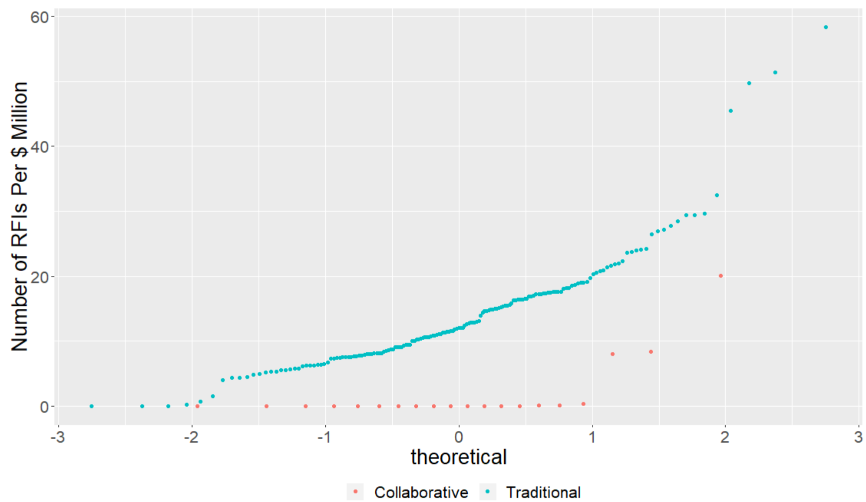


Figure 146: Q-Q Plot of Number of RFIs Per Million Dollar

Table 160: Significance of Normality for Number of RFIs Per Million Dollars

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.000	Significant	--
	Traditional	0.000	Significant	--

The following boxplot highlights a clear difference between Collaborative and Traditional delivery types. Higher variability was also noted in Traditional methods.

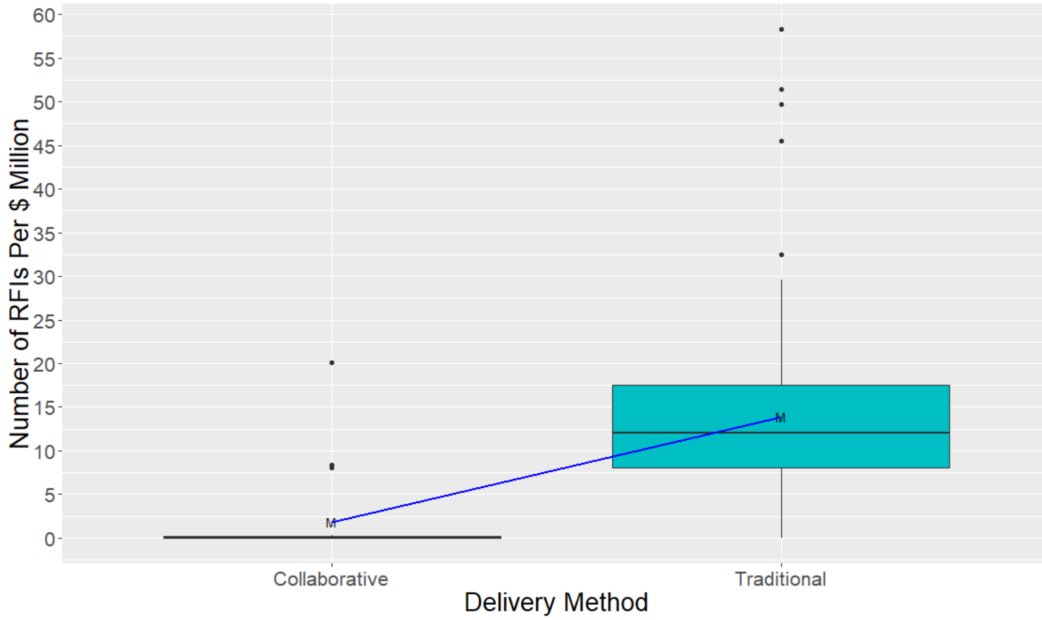


Figure 147: Boxplot of Number of RFIs Per Million Dollar Against PDTs

Thus, the following table of descriptive statistics was created, which highlights differences between the medians of Collaborative and Traditional methods (0 and 12.0) and their means (1.8 and 13.9). Furthermore, Collaborative delivery methods exhibited lower variability (5.0) than Traditional ones (8.8).

Table 161: Descriptive Statistics for Number of RFIs Per Million Dollars Per PDT

Statistical Test	Collaborative	Traditional
Mean	1.8	13.9
Maximum	20.1	58.3
75% Percentile	0.0	17.5
Median	0.0	12.0
25% Percentile	0.0	8.1
Minimum	0.0	0.0
Standard Deviation	5.0	8.8

To validate the preceding suggestions, statistical tests were performed. Fligner-Killeen returned a *p*-value of 0.000, sufficient to reject the null hypothesis at 95% confidence level. This suggests a higher variability for Traditional delivery types than Collaborative ones.

Mann-Whitney U tests also returned a *p*-value of 0.000, sufficient to reject the null hypothesis of equal medians, once again at the 95% confidence levels.

Table 162: Significance of Variance & Distribution for Number of RFIs Per Million Dollars

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Mann–Whitney	Overall	0.000	Significant	--

The findings confirm the previous conclusion about the better performance of Collaborative delivery types over Traditional ones, as well as elaborate the higher risk associated with Traditional projects compared to Collaborative ones.

5.3.3.2.3 RFIs as a Trigger for Change Orders

Projects can still be successful, even with a high number of RFIs, as they enable a higher level of clarity among the team and enhance communication. However, it is essential to notice that this metric is standardized for the project size (in million dollars). Therefore, the high number of RFIs is not absolute, but rather relative to the project size. In addition, RFIs are typically used for documentation purposes, and in many cases as a trigger for change orders.

This understanding can be further explained using the below scatter plot where a strong direct relationship can be noticed between the number of RFIs per million dollars and the number of change orders per million dollars. The relationship is statistically significant at the 95% confidence level with a *p*-value of 0.000 and a Spearman's Rho correlation coefficient of 0.73. It should be noticed that the linear regression line is for presentation purpose only and is not part of the statistical analysis.

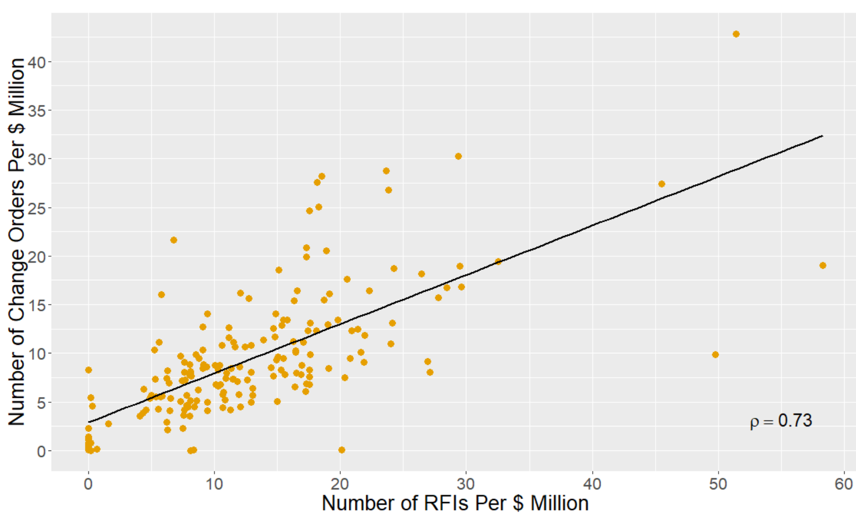


Figure 148: Number of Change Orders Per Million Dollars vs Number of RFIs Per Million Dollars

5.3.4 Change Management Performance Area

In this section, we study and compare the performance of each delivery method in terms of its ability to manage the project execution with the least possible changes. In that sense, two change management metrics are used: Number of Change Orders Per Million Dollars, which measures the intensity of issued change orders for every million dollars of construction, and Project Percentage Change, which measures the change in project's initial cost compared to the final project cost.

5.3.4.1 Number of Chance Orders Per Million Dollars

Number of change orders per million dollars is a metric that is used to measure the construction change orders approved by the State. Only approved change orders are taken into consideration in this research. While those requested but denied consume some of the team's time for discussion and reaching a consensus, which can impact productivity, those change orders are not included in this study because of their minor impact on construction flow of activities and cost.

$$Number\ of\ Change\ Orders\ Per\ Million\ Dollars = \frac{Number\ of\ Approved\ Change\ Orders}{Final\ Adjusted\ Construction\ Cost}$$

5.3.4.1.1 <u>Comparative Analysis of Delivery Methods</u>

First, normality was assessed and dismissed at the 95% confidence level, per the Q-Q plots and table below.

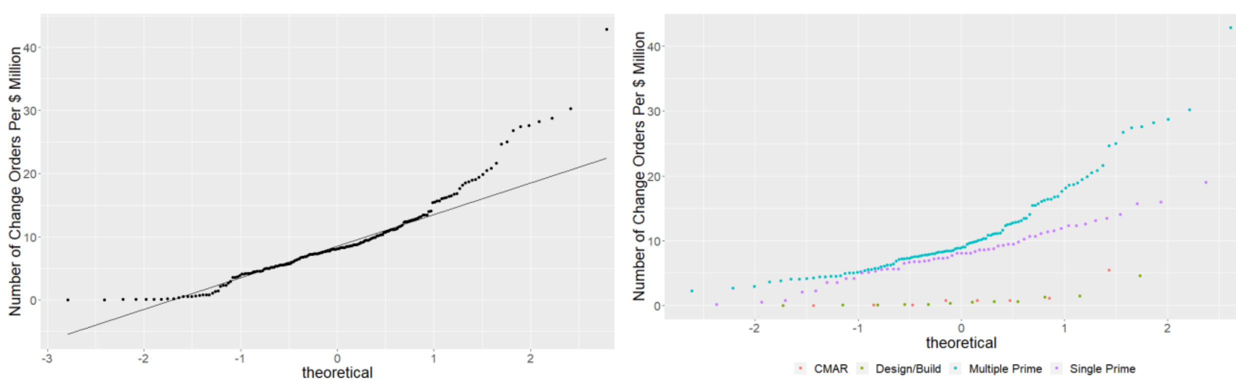


Figure 149: Q-Q Plot of Number of Change Orders Per Million Dollars

Table 163: Significance of Normality for Number of Change Orders Per Million Dollars

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.000	Significant	--
	CM at Risk	0.000	Significant	--
	Single-Prime	0.807	Not Significant	Not Significant
	Multiple-Prime	0.000	Significant	--

The comparative boxplot below shows a clear difference between DB, and CMAR, on one hand, and SP and MP on the other. However, there were no clear differences between DB and CMAR as a group. SP and MP both exhibit higher variability than the Collaborative delivery methods.

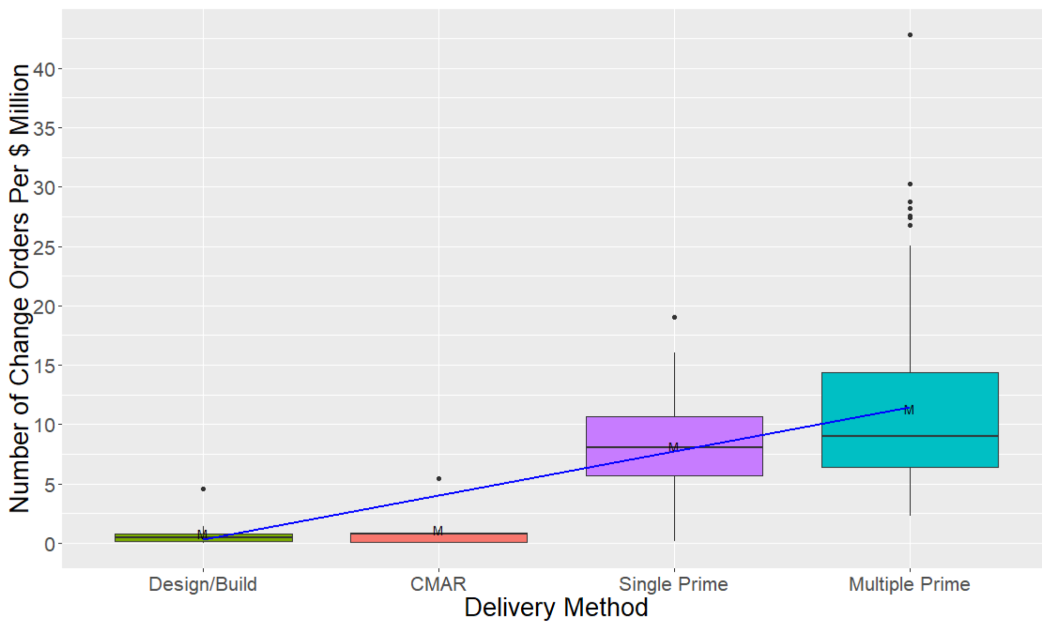


Figure 150: Boxplot of Number of Change Orders Per Million Dollars Against PDMs

Thus, the below table of descriptive statistics was drawn, in which a clear difference in medians was identified (0.4, 0.8, 8.1, 9.0). This suggests there is a considerable difference in performance between DB/CMAR and SP/MP projects which will need to be investigated further. In terms of variability, DB and CMAR had low standard deviations at 1.3 and 1.8, while SP and MP had notably higher values (3.9, and 7.1).

Table 164: Descriptive Statistics for Number of Change Orders Per Million Dollars Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	0.8	1.1	8.2	11.3
Maximum	4.6	5.5	19.1	42.9
75% Percentile	0.8	0.9	10.7	14.4
Median	0.4	0.8	8.1	9.0
25% Percentile	0.1	0.1	5.7	6.4
Minimum	0.0	0.0	0.1	2.3
Standard Deviation	1.3	1.8	3.9	7.1

Fligner-Killeen test returned a *p*-value of 0.000, sufficient to reject the null hypothesis of equal variance at the 95% confidence level. This suggests a higher variability of SP and MP projects. Kruskal-Wallis test returned a *p*-value of 0.000, followed by the post-hoc Conover-Iman test. Significant results were present between DB and SP/MP, CMAR and SP/MP, and between SP and MP, with no significant difference between DB and CMAR.

Table 165: Significance of Distribution for Number of Change Orders Per Million Dollars

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Kruskal-Wallis	Overall	0.000	Significant	--
Conover-Iman	DB vs Single	0.000	Significant	--
	DB vs Multiple	0.000	Significant	--
	CMAR vs Single	0.000	Significant	--
	CMAR vs Multiple	0.000	Significant	--
	DB vs CMAR	0.971	Not Significant	Not Significant
	Single vs Multiple	0.0344	Significant	--

The previous informative findings provide in-depth inference about the more robust performance of DB over SP and MP, CMAR over SP and MP, as well as SP against MP when it comes to maintaining better change management practices. This is reflected by the significantly lower number of change orders.

5.3.4.1.2 <u>Comparison of Delivery Types</u>

As is typical of this research, a second quantitative assessment was performed which compared performance by delivery type (Collaborative v Traditional). In this analysis, normality was dismissed at the 95% confidence level, shown in the Q-Q plot and table below.

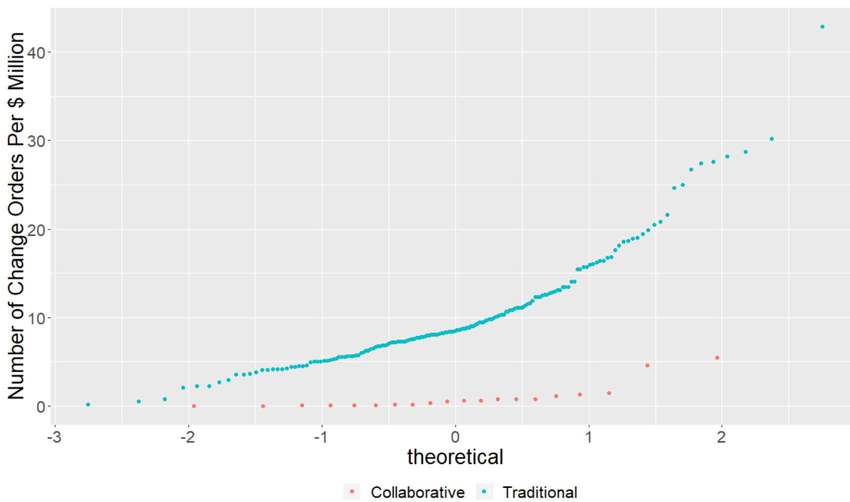


Figure 151: Q-Q Plot of Number of Change Orders Per Million Dollars

Table 166: Significance of Normality for Number of Change Orders Per Million Dollars

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.000	Significant	--
	Traditional	0.000	Significant	--

There is a clear differentiation in the boxplot below between Traditional and Collaborative projects. Further, there was lower variability in Collaborative projects as compared to Traditional ones.

Figure 152: Boxplot of Number of Change Orders Per Million Dollars Against PDMs

The table of descriptive statistics below was drawn to provide perspective. There is an evident difference between the medians of Collaborative and Traditional methods (0.6 and 8.5), as well as their means (1.0 and 10.3). Collaborative methods had lower standard deviation of 1.5 compared to that Traditional ones of 6.4. As with all conclusions, statistical validation was used.

Table 167: Descriptive Statistics for Number of Change Orders Per Million Dollars Per PDT

Statistical Test	Collaborative	Traditional
Mean	1.0	10.3
Maximum	5.5	42.8
75% Percentile	0.9	12.6
Median	0.6	8.5
25% Percentile	0.1	6.2
Minimum	0.0	0.1
Standard Deviation	1.5	6.4

To quantitatively confirm our previous findings, using the normality test previously assessed. Fligner-Killeen test was performed, returning a *p*-value of 0.000 which is sufficient to reject the null hypothesis of equal variance among delivery types at the 95% confidence level. This suggests higher variability in Traditional delivery types as compared to Collaborative ones. Mann-Whitney U test returned a *p*-value of 0.000 –also sufficient to reject the null hypothesis of equal medians among the two delivery types.

Table 168: Significance of Variance & Distribution for Schedule Growth

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Mann–Whitney	Overall	0.000	Significant	--

The findings confirm the previous conclusion about the more robust performance of Collaborative delivery types over Traditional ones, as well as elaborate on the higher risk associated with Traditional projects compared to Collaborative ones when it comes to experiencing higher number of change orders for every million dollars of construction.

5.3.4.2 Project Percentage Change

Project percentage change is a metric that is used to measure the percentage cost change (either decrease or increase) as compared to the Base Bid. Similar high-level analysis was discussed in earlier sections, with emphasis on the highlighting which delivery method is likely to experience cost overrun. Nevertheless, the motive here is different as the research quantify the percentage of cost change and compare it among the four delivery methods.

$$Project\ Percentage\ Change = \frac{Actual\ Project\ Cost - Base\ Bid}{Base\ Bid}$$

5.3.4.2.1 <u>Comparative Analysis of Delivery Methods</u>

First, normality was assessed and dismissed at the 95% confidence level, per the Q-Q plots and table below.

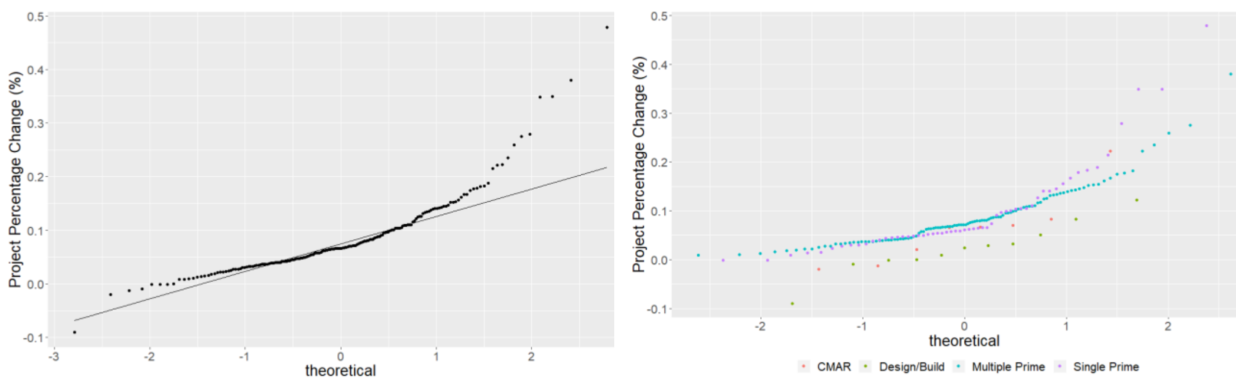


Figure 153: Q-Q Plot of Project Percentage Change

Table 169: Significance of Normality for Project Percentage Change

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.950	Not Significant	Not Significant
	CM at Risk	0.854	Not Significant	Not Significant
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

The comparative boxplot below shows a declining trend, starting with DB with the best performance, and SP with the worst. The four delivery methods exhibit similar variability.

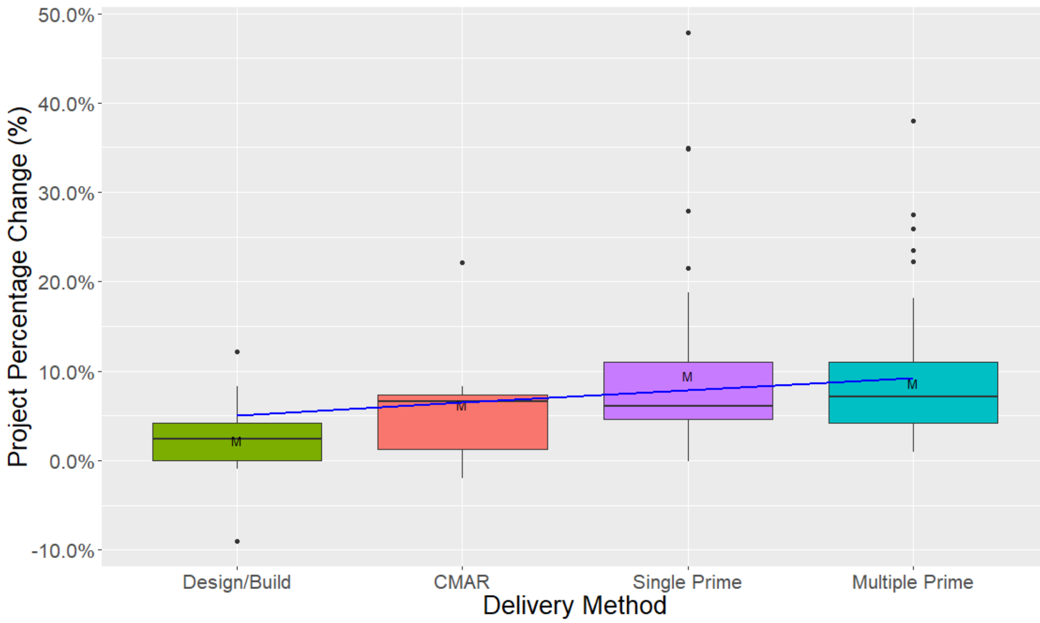


Figure 154: Boxplot of Project Percentage Change Against PDMs

Thus, the below table of descriptive statistics was drawn, in which difference in medians were identified (1.6, 6.6, 6.1, and 7.1%). This suggests there is a considerable difference in performance of DB projects which will need to be investigated further. In terms of variability, all methods had similar standard deviations at 5.2, 7.5, 9.2 and 6.0%, respectively.

Table 170: Descriptive Statistics for Project Percentage Change Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	2.3%	6.2%	9.5%	8.6%
Maximum	12.2%	22.2%	47.9%	38.0%
75% Percentile	3.7%	7.3%	11.1%	11.0%
Median	1.6%	6.6%	6.1%	7.1%
25% Percentile	0.0%	1.3%	4.6%	4.2%
Minimum	-9.0%	-2.0%	-0.1%	0.9%
Standard Deviation	5.2%	7.5%	9.2%	6.0%

Two statistical tests were further performed to provide more robust results. Fligner-Killeen test returned a *p*-value of 0.807, insufficient to reject the null hypothesis of equal variance at the 95% confidence level. Kruskal-Wallis test returned a *p*-value of 0.005, which rejects the null hypothesis at the 95% confidence level. As this test was significant, the post-hoc Conover-Iman test was also performed. DB was found to be significantly better than SP and MP, with no other significant difference.

Table 171: Significance of Distribution for Project Percentage Change

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.807	Not Significant	Not Significant
Kruskal-Wallis	Overall	0.005	Significant	--
Conover-Iman	DB vs Single	0.005	Significant	--
	DB vs Multiple	0.001	Significant	--
	CMAR vs Single	0.644	Not Significant	Not Significant
	CMAR vs Multiple	0.494	Not Significant	Not Significant
	DB vs CMAR	0.377	Not Significant	Not Significant
	Single vs Multiple	0.900	Not Significant	Not Significant

The previous informative findings provide in-depth inference about the higher performance of DB over SP and MP when it comes to reducing cost overrun.

5.3.4.2.2 <u>Comparative Analysis of Delivery Types</u>

As is typical of this research, a second quantitative assessment was performed which compared performance by delivery type (Collaborative v Traditional). In this analysis, normality was dismissed at the 95% confidence level, shown in the Q-Q plot and table below.

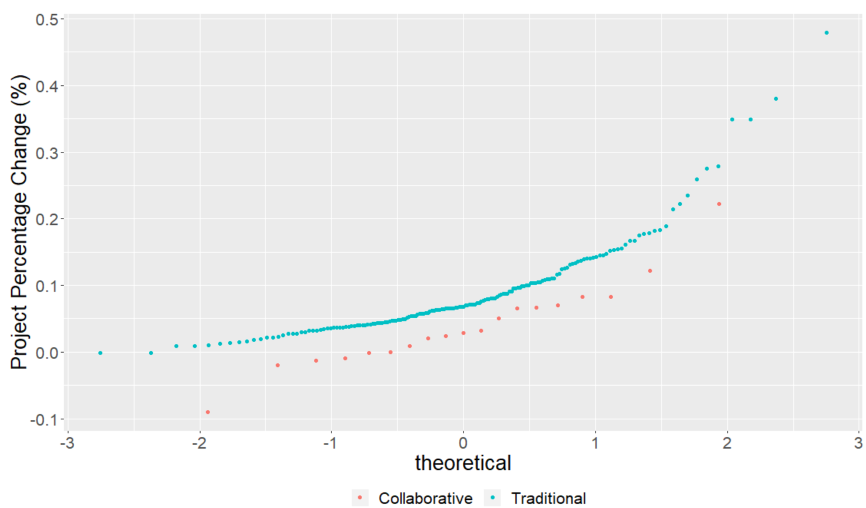


Figure 155: Q-Q Plot of Project Percentage Change

Table 172: Significance of Normality for Project Percentage Change

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.005	Significant	--
	Collaborative	0.164	Not Significant	Not Significant
	Traditional	0.000	Significant	--

There is a clear differentiation in the boxplot below between Traditional and Collaborative projects. Further, there was no difference in variability between Collaborative and Traditional projects.

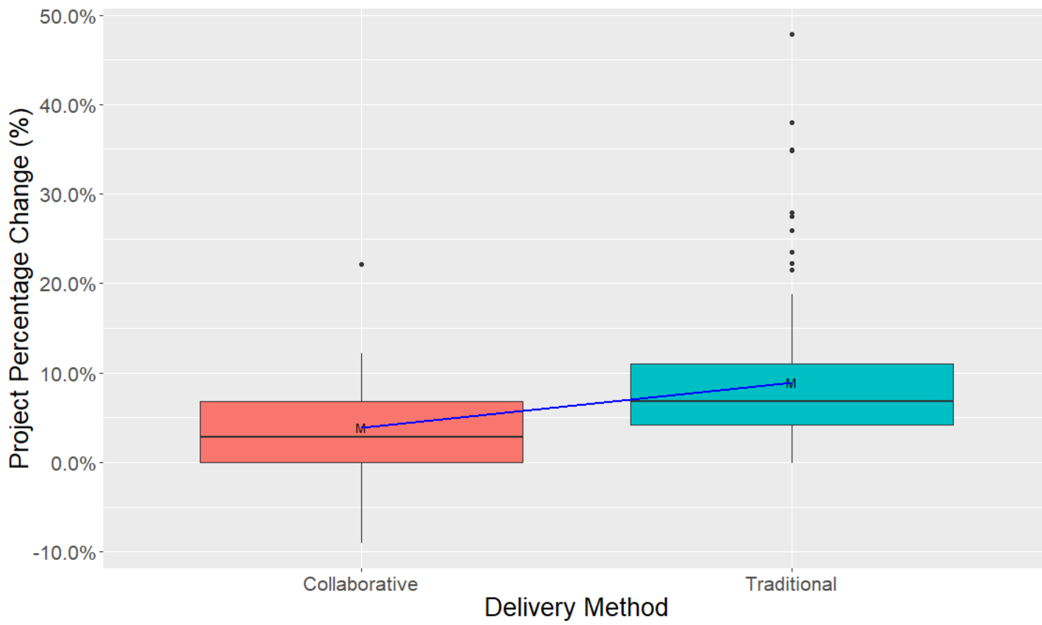


Figure 156: Boxplot of Project Percentage Change Against PDTs

The table of descriptive statistics below was drawn to provide perspective. There is an evident difference between the medians of Collaborative and Traditional methods (2.6 and 6.8%), as well as their means (3.9 and 8.9%). Collaborative methods had slightly lower standard deviation of 6.4% compared to that Traditional ones of 7.2%. As with all conclusions, statistical validation was used.

Table 173: Descriptive Statistics for Project Percentage Change Per PDT

Statistical Test	Collaborative	Traditional
Mean	3.9%	8.9%
Maximum	22.2%	47.9%
75% Percentile	6.7%	11.0%
Median	2.6%	6.8%
25% Percentile	0.0%	4.2%
Minimum	-9.0%	-0.1%
Standard Deviation	6.4%	7.2%

To quantitatively confirm our previous findings, using the normality test previously assessed. Fligner-Killeen test was performed, returning a *p*-value of 0.820 – insufficient to reject the null hypothesis of equal variance among delivery types at the 95% confidence level. Mann-Whitney U test returned a *p*-value of 0.001 –sufficient to reject the null hypothesis of equal medians among the two delivery types.

Table 174: Significance of Variance & Distribution for Project Percentage Change

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.820	Not Significant	Not Significant
Mann–Whitney	Overall	0.001	Significant	--

The findings confirm the previous conclusion about the more robust performance of DB delivery method in specific, and Collaborative delivery types in general when compared to Traditional ones when managing changes effectively and keeping the final cost closer to the initial cost.

5.3.5 Spending Performance Area

In this section, we study and compare the performance of each delivery method in terms of its ability to execute the project in a timely manner. In that sense, two metrics are utilized: Construction Speed, which measures the spending speediness of the construction cost in thousand dollars per day, and Delivery Speed, which measures the spending speediness of the delivery (construction and design) cost in thousand dollars per day.

5.3.5.1 Construction Speed

Construction speed is a metric that is calculated by dividing Actual Construction Adjusted Cost by Actual Construction Duration. While this metric can sometimes be computed as the executed square footage per day, utilizing dollar spending in our case is more reasonable since the square footage data was not as reliable and might have led to latent errors in the results. In addition, the second metric Delivery Speed could not be computed as a function of square footage since it incorporates design speed. Therefore, this computation method is selected to provide a consistent and comparable evaluation.

$$Construction\ Speed = \frac{Actual\ Adjusted\ Construction\ Cost}{Actual\ Construction\ Duration * 1000}$$

5.3.5.1.1 <u>Comparative Analysis of Delivery Methods</u>

As evident from the Q-Q plots below, normality was not present. This was confirmed at the 95% confidence level by the Shapiro-Wilk test, as shown in the table below.

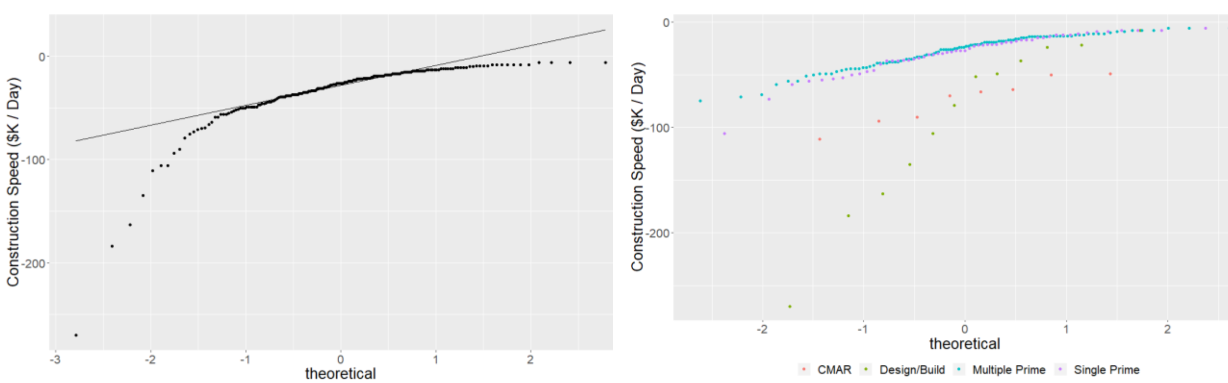


Figure 157: Q-Q Plot of Construction Speed

Table 175: Significance of Normality for Construction Speed

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.153	Not Significant	Not Significant
	CM at Risk	0.496	Not Significant	Not Significant
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

By considering the comparative boxplot below, there is a clear visual difference in performance between Collaborative and Traditional delivery methods, but there is no clear difference between DB and CMAR, and between SP and MP. DB delivery exhibited the highest variability.

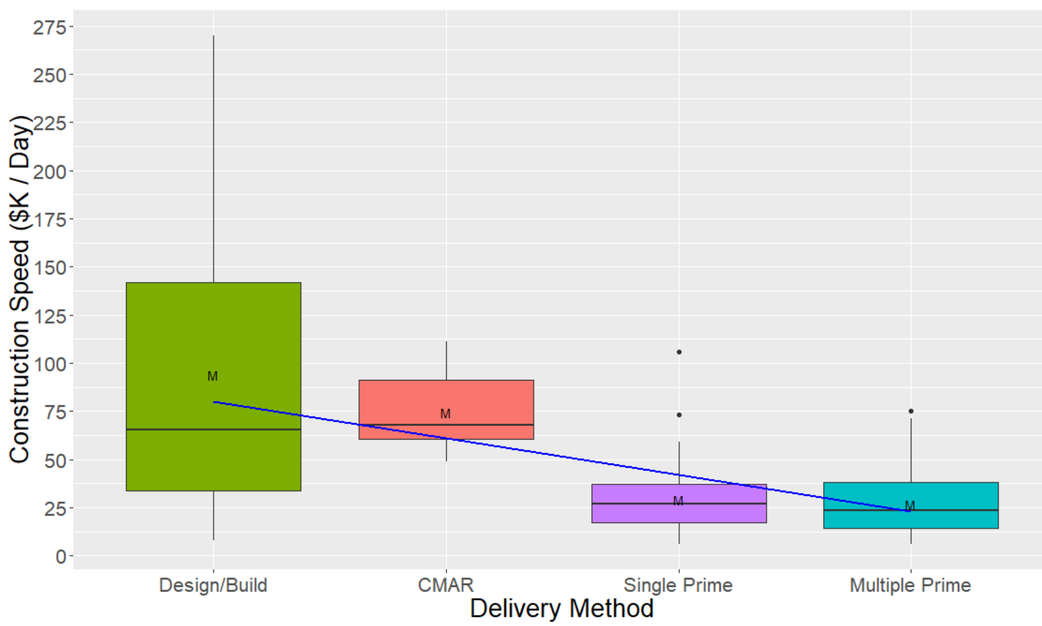


Figure 158: Boxplot of Construction Speed Against PDMs

The table of descriptive statistics, below, was compiled to validate the observations. A clear difference in the median of Collaborative (66/68) and Traditional (27/24). The medians suggest no difference between DB and CMAR, nor between SP and MP, as was previously speculated. Further testing was performed to validate this assumption. Variability was higher in DB (80) than CMAR, SP and MP (22/19/15). Once again, statistical tests were performed to validate these findings.

Table 176: Descriptive Statistics for Construction Speed Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	94	74	29	27
Maximum	270	111	106	75
75% Percentile	142	91	37	38
Median	66	68	27	24
25% Percentile	34	60	17	14
Minimum	8	49	6	6
Standard Deviation	80	22	19	15

Fligner-Killeen returned a *p*-value of 0.000, confirming the inequal variance among delivery methods at the 95% confidence level and suggesting a lower variability in CMAR, SP and MP than in DB.

Kruskal-Wallis further returned a *p*-value of 0.000, rejecting the assumption of equal medians at the 95% confidence level. Conover-Iman post-hoc test showed statistically significant differences between DB and SP/MP, CMAR and SP/ MP, with no difference shown between DB and CMAR nor between SP and MP.

Table 177: Significance of Distribution for Construction Speed

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Kruskal-Wallis	Overall	0.000	Significant	--
Conover-Iman	DB vs Single	0.000	Significant	--
	DB vs Multiple	0.003	Significant	--
	CMAR vs Single	0.000	Significant	--
	CMAR vs Multiple	0.000	Significant	--
	DB vs CMAR	0.498	Not Significant	Not Significant
	Single vs Multiple	0.860	Not Significant	Not Significant

The previous informative findings provide in-depth inference about the more robust performance of DB over SP and MP and CMAR over SP and MP when it comes to quickly executing the construction phase with significantly higher speed.

5.3.5.1.2 Comparative Analysis of Delivery Types

Once again, normality was assessed, and a lack of normality was found at the 95% confidence level, as shown in the figure and table below.

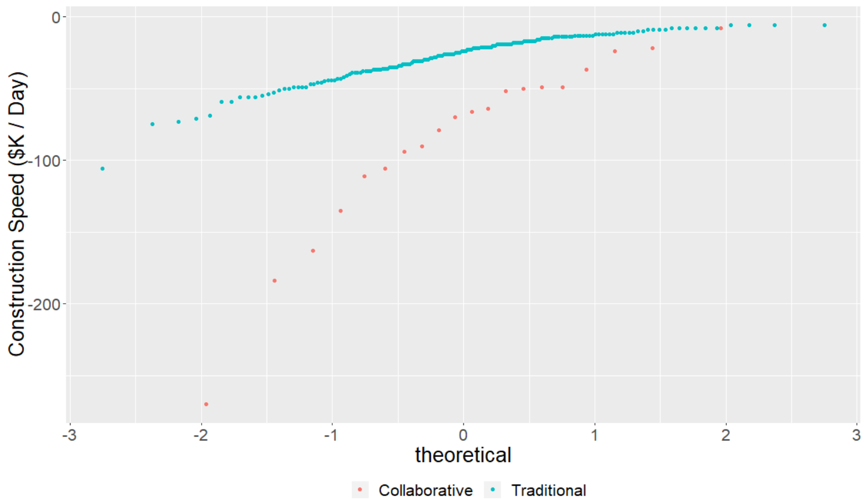


Figure 159: Q-Q Plot of Construction Speed

Table 178: Significance of Normality for Schedule Growth

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.017	Significant	--
	Traditional	0.000	Significant	--

The following boxplot highlights a clear difference between Collaborative and Traditional delivery types. Higher variability was also noted in Collaborative methods.

Figure 160: Boxplot of Construction Speed Against PDTs

Thus, the following table of descriptive statistics was created, which highlights differences between the medians of Collaborative and Traditional methods (68 and 24) and their means (86 and 28). Furthermore, Collaborative delivery methods exhibited higher variability (63) than Traditional ones (16).

Table 179: Descriptive Statistics for Construction Speed Per PDT

Statistical Test	Collaborative	Traditional
Mean	86	28
Maximum	270	106
75% Percentile	107	37
Median	68	24
25% Percentile	49	15
Minimum	8	6
Standard Deviation	63	16

To validate the preceding suggestions, statistical tests were performed. Fligner-Killeen returned a *p*-value of 0.000, sufficient to reject the null hypothesis at 95% confidence level. This suggests a higher variability for Collaborative delivery types than Traditional ones.

Mann-Whitney U tests also returned a *p*-value of 0.000, sufficient to reject the null hypothesis of equal medians in favor of Collaborative projects, once again at the 95% confidence level.

Table 180: Significance of Variance & Distribution for Construction Speed

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Mann–Whitney	Overall	0.000	Significant	--

The findings confirm the previous conclusion about the more robust performance of Collaborative delivery types over Traditional ones when it comes to construction speed and timely execution. However, Collaborative projects are on average more risky than Traditional one such that a higher potential of deviating from the mean exists for Collaborative projects than Traditional.

5.3.5.2 Delivery Speed

Delivery speed is a metric that is calculated by dividing the Actual Adjusted Delivery Cost by the Actual Delivery Duration.

$$Delivery\ Speed = \frac{Actual\ Adjusted\ Delivery\ Cost}{Actual\ Delivery\ Duration * 1000}$$

5.3.5.2.1 <u>Comparative Analysis of Delivery Methods</u>

As evident from the Q-Q plots below, normality was not present. This was confirmed at the 95% confidence level by the Shapiro-Wilk test, as shown in the table below.

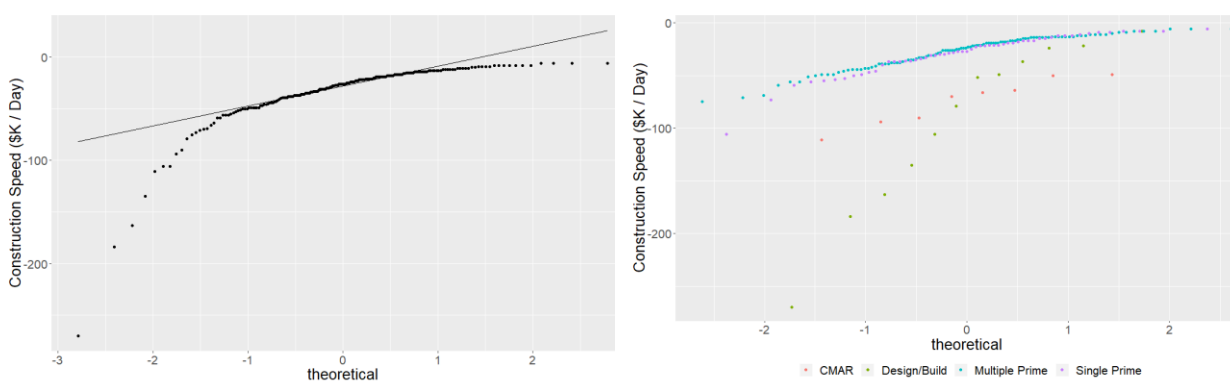


Figure 161: Q-Q Plot of Delivery Speed

Table 181: Significance of Normality for Delivery Speed

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.314	Not Significant	Not Significant
	CM at Risk	0.884	Not Significant	Not Significant
	Single-Prime	0.000	Significant	--
	Multiple-Prime	0.000	Significant	--

By considering the comparative boxplot below, there is a clear visual difference in performance between Collaborative and Traditional delivery methods, but there is no clear difference between DB and CMAR, and between SP and MP. DB delivery exhibited the highest variability.

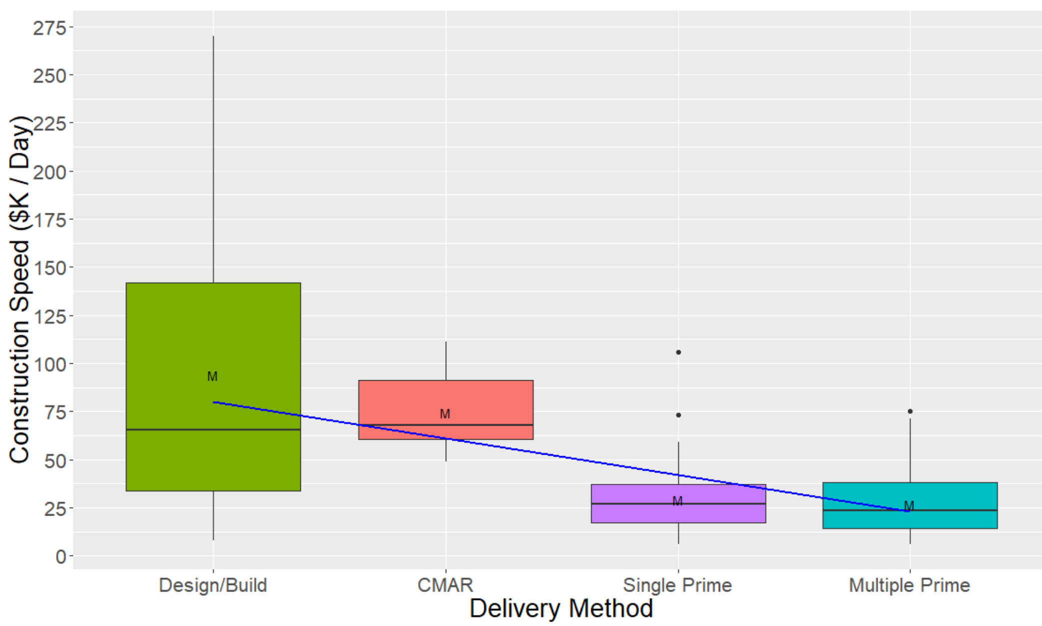


Figure 162: Boxplot of Delivery Speed Against PDMs

The table of descriptive statistics, below, was compiled to validate the observations. A clear difference in the median of Collaborative (37/34) and Traditional (9/10). The medians suggest no difference between DB and CMAR, nor between SP and MP, as was previously speculated. Further testing was performed to validate this assumption. Variability was higher in DB and CMAR (30/15) than SP and MP (9/7). Once again, statistical tests were performed to validate these findings.

Table 182: Descriptive Statistics for Delivery Speed Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	40	38	12	11
Maximum	98	63	61	38
75% Percentile	53	47	15	14
Median	37	34	9	10
25% Percentile	14	30	5	6
Minimum	5	17	3	3
Standard Deviation	30	15	9	7

Fligner-Killeen returned a *p*-value of 0.006, confirming the inequal variance among delivery methods at the 95% confidence level and suggesting a lower variability in SP and MP than in DB and CMAR.

Kruskal-Wallis further returned a *p*-value of 0.000, rejecting the assumption of equal medians at the 95% confidence level. Conover-Iman post-hoc test showed statistically significant differences between DB and SP/MP, CMAR and SP/MP, with no difference shown between DB and CMAR nor between SP and MP.

Table 183: Significance of Variance & Distribution for Delivery Speed

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.006	Significant	--
Kruskal-Wallis	Overall	0.000	Significant	--
Conover-Iman	DB vs Single	0.003	Significant	--
	DB vs Multiple	0.004	Significant	--
	CMAR vs Single	0.000	Significant	--
	CMAR vs Multiple	0.000	Significant	--
	DB vs CMAR	0.439	Not Significant	Not Significant
	Single vs Multiple	0.918	Not Significant	Not Significant

The previous informative findings provide in-depth inference about the higher performance of DB over SP and MP, and CMAR over SP and MP when it comes to executing the design and construction phases with significantly higher speed.

5.3.5.2.2 Comparative Analysis of Delivery Types

Normality was assessed, and a lack of normality was found at the 95% confidence level, as shown in the figure and table below.

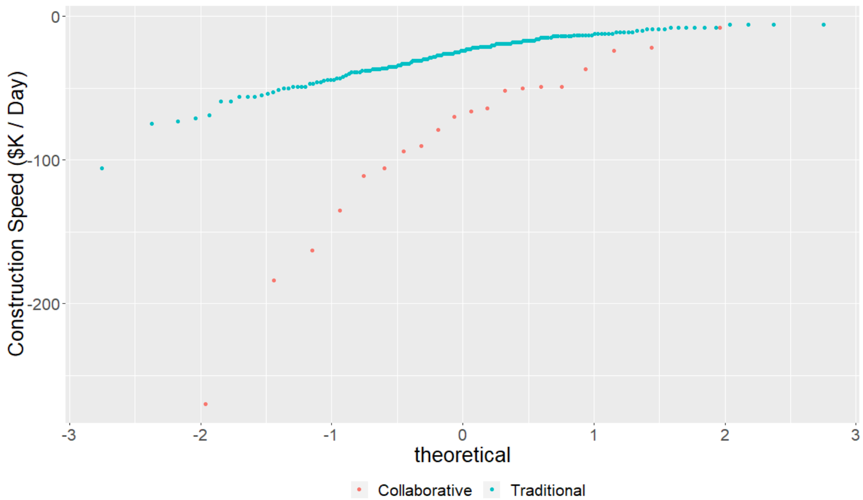


Figure 163: Q-Q Plot of Delivery Speed

Table 184: Significance of Normality for Delivery Speed

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.423	Not Significant	Not Significant
	Traditional	0.000	Significant	--

The following boxplot highlights a clear difference between Collaborative and Traditional delivery types. Higher variability was also noted in Collaborative methods.

Figure 164: Boxplot of Delivery Speed Against PDTs

Thus, the following table of descriptive statistics was created, which highlights differences between the medians of Collaborative and Traditional methods (36 and 10) and their means (39 and 11). Furthermore, Collaborative delivery methods exhibited higher variability (25) than Traditional ones (8).

Table 185: Descriptive Statistics for Delivery Speed Per PDT

Statistical Test	Collaborative	Traditional
Mean	39	11
Maximum	98	61
75% Percentile	49	14
Median	36	10
25% Percentile	24	6
Minimum	5	3
Standard Deviation	25	8

To validate the preceding suggestions, statistical tests were performed. Fligner-Killeen returned a *p*-value of 0.000, sufficient to reject the null hypothesis at 95% confidence level. This suggests a higher variability for Collaborative delivery types than Traditional ones.

Mann-Whitney U tests also returned a p-value of 0.000, sufficient to reject the null hypothesis of equal medians in favor of Collaborative projects, once again at the 95% confidence level.

Table 186: Significance of Variance & Distribution for Schedule Growth

Statistical Test	Parameter	p-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Mann–Whitney	Overall	0.000	Significant	--

The findings confirm the previous conclusion about the higher performance of Collaborative delivery types over Traditional ones when it comes to project overall speed and timely execution. However, Collaborative projects are on average more risky than Traditional ones such that a higher potential of deviating from the mean exists for Collaborative projects than Traditional ones.

5.3.6 Section Summary

In this section, special emphasis was given to compare the cost per gross square foot of eleven facility types, representing around 75% of the dataset, to the National Average reported by RS Means 2018. The results indicated that three out of eleven facility types experienced lower-than-average cost per gross square foot with an average difference of 14%. Meanwhile, the remaining 8 facility types experienced higher-than-average cost per gross square foot with an average difference of 23%. As a result, it was concluded that the cost of constructing the eleven facility types was 13% above the National Average.

This analysis also compared the performance of the different delivery methods based on eight performance metrics that spans four performance areas including schedule, communication, change management and spending. The results concluded that DB outperformed SP and MP in all eight metrics, whereas CMAR outperformed SP in 5 metrics and MP in 7 metrics. Moreover, it was determined that DB outperformed CMAR in one metric only, while SP outperformed MP in 3 metrics. These results suggest a higher performance of DB, followed by CMAR in regard to the eight identified performance metrics. SP also performed slightly better than MP in three metrics.

Table 187: Statistical Significance Summary of PDM Pairwise Performance

Compared PDSs	DB – Single	DB – Multiple	CMAR – Single	CMAR – Multiple	DB – CMAR	Single – Multiple
Schedule growth	**	**		**		**
Schedule Factor	**	**	**	**		**
RFI processing time	**	**		*	**	**
# of RFIs per $ Million	**	**	**	**		
# of change orders per $ Million	**	**	**	**		
Project percentage change	**	**				
Construction speed	**	**	**	**		
Delivery speed	**	**	**	**		

**statistically significant difference at the 95% confidence level (p-value <0.05)
*statistically significant difference at the 90% confidence level (p-value <0.10)

Further results focused on the global performance of delivery types whether Collaborative or Traditional. The analysis concluded that Collaborative delivery methods experienced significantly lower schedule growth at the 95% confidence level with an average of 1.5% compared to the Traditional ones with an average of 38.4% (1.85 times less[3]). Also, Collaborative delivery methods experienced significantly lower schedule factor with an average of 0.91 compared to the Traditional ones with an average of 1.27 (33%

[3] Percent Difference was computed as defined in Chapter 1 – Definition of Terms. Note the difference from Percent Error.

less[3]), which provide sufficient evidence of the higher performance of Collaborative projects in regard to maintaining minimum schedule overrun.

Similarly, in terms of communication among parties, the analysis concluded that Collaborative delivery methods experienced significantly lower RFI processing time at the 95% confidence level with an average of 2.1 days compared to the Traditional ones with an average of 8.5 days (1.2 times less[3]). Also, Collaborative delivery methods experienced significantly lower number of RFIs per million dollars with an average of 1.8 RFI for every million dollars compared to the Traditional ones with an average of 13.9 RFI for every million dollars (1.5 times less[3]). These findings provide sufficient evidence of the higher performance of Collaborative projects in regard to maintaining better communication among the project participants. It also provides further insights about the quality of the design generated with minimum ambiguity, which accordingly required minimum clarification using RFIs.

Although the number of RFIs can sometimes be an indicator of a good communication and high level of coordination among the parties, further analysis found that RFIs could be utilized as a trigger for change orders with a strong positive correlation (0.73) between the number of RFIs per million dollars and the number of change orders per million dollars.

As for change management, the research findings determined that Collaborative delivery methods experienced significantly lower number of change orders per million dollars with an average of 1.0 change order for every million dollars compared to the Traditional ones (10.3, which is 1.65 times less[3]). Also, Collaborative delivery methods experienced significantly lower project percentage change with an average of 3.9% compared to the Traditional ones with an average of 8.9% (78% less[3]).

Finally, the project spending was examined for each delivery type where it was proven that Collaborative delivery methods experienced significantly higher construction speed with an average spending of 64 thousand dollars per day compared to the Traditional ones (28, which is 78% more[3]). Also, Collaborative delivery methods experienced significantly higher delivery speed with an average spending of 39 thousand dollars per day compared to the Traditional ones with an average spending of 11 thousand dollars per day (1.1 times more[3]).

5.4 Project Performance Index (PPI) Analysis

5.4.1 Comparative Analysis of Delivery Methods

When quantitatively assessing the impact of each delivery method on the project performance index scores, it is essential to understand the data type and validate the normality assumption in order to select the applicable significance tests.

The following Q-Q plot is used to evaluate the normality assumption. The figure shows a clear departure from normality at the overall level of the dataset, as well as some levels of the delivery methods.

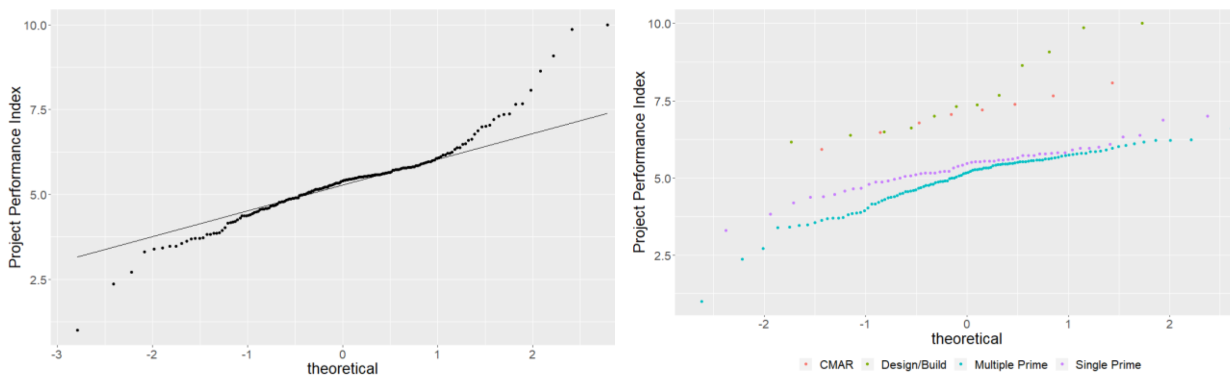


Figure 165: Q-Q Plot of Project Performance Index

As with the other analysis, the initial step was to determine normality. As evident from the Q-Q plots below, normality was not present. This was confirmed at the 95% confidence level by the Shapiro-Wilk test, as shown in the table below.

Table 188: Significance of Normality for Project Performance Index Scores

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.130	Not Significant	Not Significant
	CM at Risk	0.999	Not Significant	Not Significant
	Single-Prime	0.244	Not Significant	Not Significant
	Multiple-Prime	0.000	Significant	--

In light of the following comparative boxplot, a clear difference was identified between DB and CMAR on the one hand, and SP and MP on the other hand based on median values. Meanwhile, no clear differences were identified between DB and CMAR, or between SP and MP. In terms of variability, it is also noticed that DB experience the highest variability when compared to CMAR, SP and MP, highlighted by the compaction level of each boxplot.

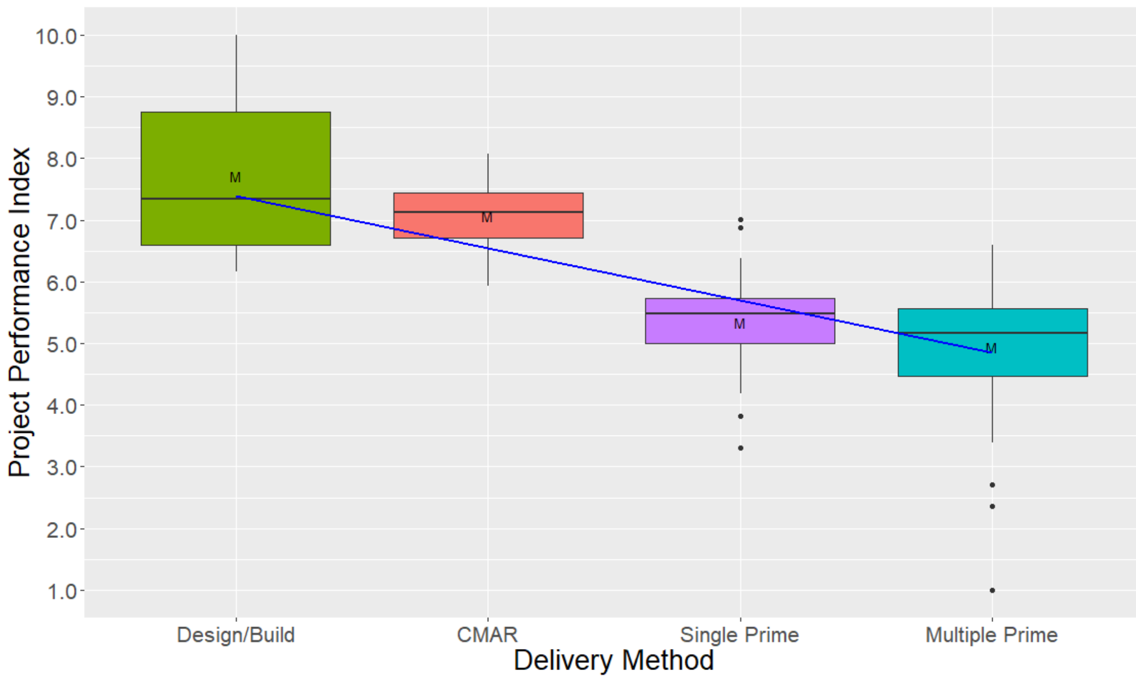


Figure 166: Boxplot of Project Performance Index Against PDMs

To put our previous observations into perspective, we refer to the following table of descriptive statistics, where a clear difference between the medians of each of DB at 7.3 and CMAR at 7.1 on the one hand, and SP at 5.5 and MP at 5.2 on the other hand is highlighted. The statistics also suggest no difference between DB and CMAR and a slight difference between SP and MP, which will require further statistical tests to confirm its significance.

As for the variability, DB experience the highest standard deviation of 1.2, while CMAR, SP and MP experienced relatively consistent and smaller values at 0.7, 0.7 and 0.9 respectively. However, it should be noticed that standard deviations can sometimes be influenced by outliers. Therefore, an additional statistical test was performed as to confirm this finding.

Table 189: Descriptive Statistics for Project Performance Index Scores Per PDM

Statistical Test	Design/Build	CM at Risk	Single-Prime	Multiple-Prime
Mean	7.6	7.1	5.3	4.9
Maximum	10.0	8.1	7.0	6.6
75% Percentile	8.7	7.4	5.7	5.6
Median	7.3	7.1	5.5	5.2
25% Percentile	6.6	6.7	5.0	4.5
Minimum	6.2	5.9	3.3	1.0
Standard Deviation	1.2	0.7	0.7	0.9

Furthermore, quantitative statistical tests were performed to confirm our previous findings. In this case, since the data did not follow a normal distribution, Fligner-Killeen is used to test the homogeneity of variances and Kruskal Wallis to test the significance across the four delivery methods.

With that, Fligner-Killeen test resulted in a *p*-value of 0.020, rejecting the null hypothesis of equal variance among the delivery methods, with a statistically significant difference at the 95% confidence level, and suggesting a higher variability of DB compared to CMAR, SP and MP.

In addition, Kruskal-Wallis test resulted in a *p*-value of 0.000, rejecting the null hypothesis of equal population medians among the delivery methods with a statistically significant difference at the 95% confidence level. Following that, Conover-Iman post-hoc test is performed to evaluate pairwise differences between each of the four delivery methods. The results show a statistically significant difference between DB and each of SP and MP, CMAR and each of SP and MP and SP and MP, with no statistically significant difference between DB and CMAR.

Table 190: Significance of Distribution for Project Performance Index Scores

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.020	Significant	--
Kruskal-Wallis	Overall	0.000	Significant	--
Conover-Iman	DB vs Single	0.000	Significant	--
	DB vs Multiple	0.000	Significant	--
	CMAR vs Single	0.000	Significant	--
	CMAR vs Multiple	0.000	Significant	--
	DB vs CMAR	0.968	Not Significant	Not Significant
	Single vs Multiple	0.017	Significant	--

The previous informative findings provide in-depth inference about the higher performance of DB over SP and MP and CMAR over SP and MP from a comprehensive performance perspective represented by project performance index scores.

5.4.2 Comparative Analysis of Delivery Types

Following that, another quantitative assessment is performed to study the impact of each delivery type on project performance index scores. In this assessment, similar strategy is followed to understand the data type and validate the normality assumption, which accordingly allow for selecting applicable significance tests.

Q-Q plot is used again to evaluate the normality assumption. The following figure shows a clear departure from normality at the level of each delivery type.

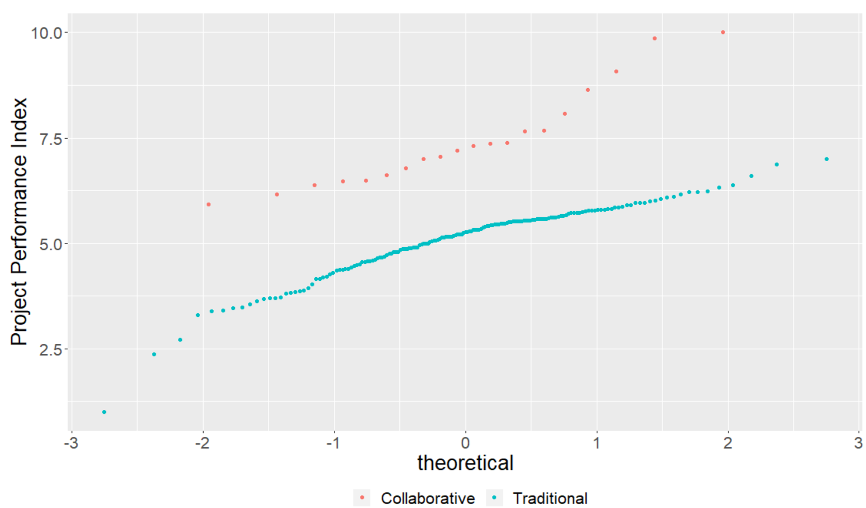


Figure 167: Q-Q Plot of Project Performance Index

As evident from the Q-Q plots below, normality was not present. This was confirmed at the 95% confidence level by the Shapiro-Wilk test, as shown in the table below.

Table 191: Significance of Normality for Project Performance Index Scores

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.000	Significant	--
	Traditional	0.049	Significant	--

In light of the following comparative boxplot, a clear difference was identified between Collaborative and Traditional delivery types based on median and average values. In terms of variability, it is also noticed that both Collaborative and Traditional delivery types encountered very similar variability, highlighted by the compaction level of each boxplot.

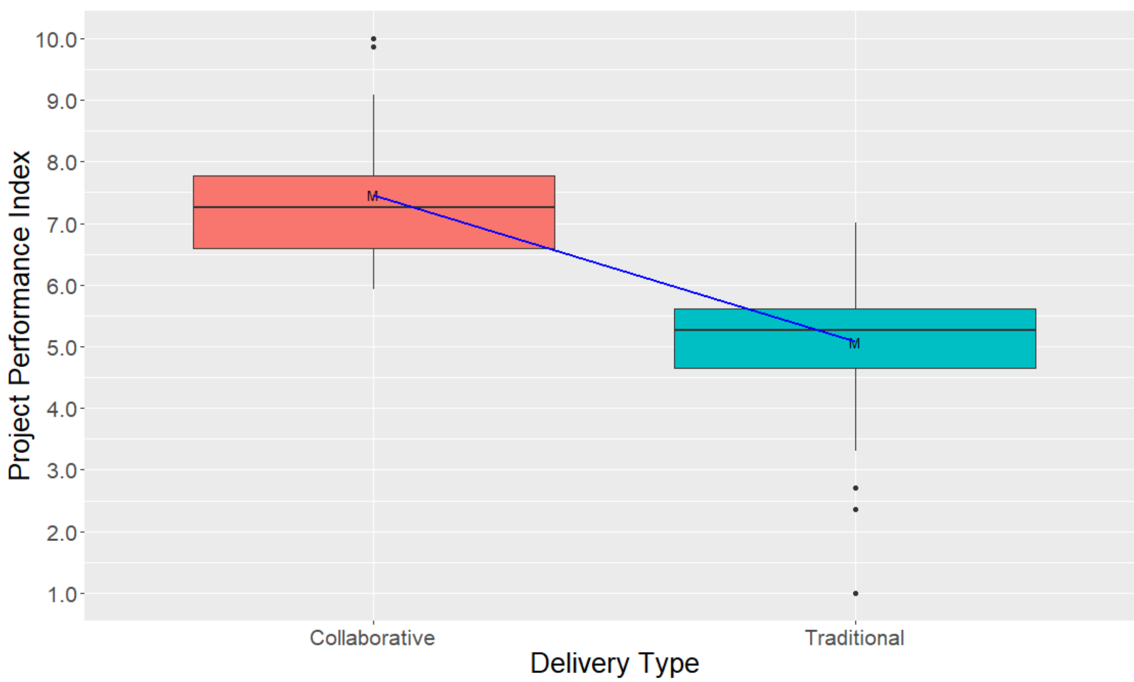


Figure 168: Boxplot of Project Performance Index Against PDTs

To put our previous observations into perspective, we refer to the following table of descriptive statistics where a clear difference is noticed between the medians of each of the Collaborative and Traditional delivery methods at 7.3 and 5.3 respectively. The same understanding is vivid when inspecting the mean values of the two types at 7.4 and 5.1 respectively.

As for the variability, both Collaborative and Traditional projects experience very similar standard deviations at 1.1 and 0.8 respectively. However, as noted earlier, variability should not be solely interpreted using standard deviations due to its ability to be distorted by outliers. Therefore, an additional statistical test was performed in order to confirm this understanding.

Table 192: Descriptive Statistics for Project Performance Index Scores Per PDT

Statistical Test	Collaborative	Traditional
Mean	7.4	5.1
Maximum	10.0	7.0
75% Percentile	7.8	5.6
Median	7.3	5.3
25% Percentile	6.6	4.6
Minimum	5.9	1.0
Standard Deviation	1.1	0.8

Moreover, quantitative appropriate statistical significance tests were performed so as to confirm the previous interpretations and claims. Since previous analysis has shown that the data did not follow a normal distribution, Fligner-Killeen is used to test the homogeneity of variances and Mann-Whitney to test the significance across the two delivery types.

With that, Fligner-Killeen test resulted in a p-value of 0.184, failing to reject the null hypothesis of equal variance among the delivery types, with a statistically insignificant difference at both the 95% confidence level.

In addition, Mann-Whitney U test resulted in a p-value of 0.000, rejecting the null hypothesis of equal medians among the two delivery types with a statistically significant difference at the 95% confidence level.

Table 193: Significance of Variance & Distribution for Project Performance Index Scores

Statistical Test	Parameter	p-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.184	Not Significant	Not Significant
Mann–Whitney	Overall	0.000	Significant	--

The findings conform with the previous conclusions about the better performance of Collaborative delivery types over Traditional ones when it comes to project performance index scores, which represents the collective output of schedule, communication, change management and efficiency.

5.4.3 PPI Scores vs. Performance Metrics Regression Assessment

The developed mathematical model aims at maximizing the correlation between the project performance index scores and each of the individual performance metrics. Therefore, this analysis serves as an internal validation to verify that the selected metrics provide substantial contribution to the performance index model by checking the correlation coefficient. In that sense, if a metric is found to have a negligible correlation coefficient with PPI scores, it can either be concluded that the metric is not representative of the project's performance, or the developed model was not successful in finding the optimum weights that could capture the maximum contribution of the introduced metrics.

Since the data experienced a significant departure from the normality assumption, Spearman's Rho ρ is used to compute the correlation strength. Note that the linear regression line is present for representation purposes only and does not contribute to the computation.

5.4.3.1 Schedule Growth

Considering the following figure, a moderate to strong negative correlation is found between the project performance index scores and the schedule growth. The correlation is statistically significant at the 95% confidence level with a *p*-value of 0.000 and a correlation coefficient of -0.61.

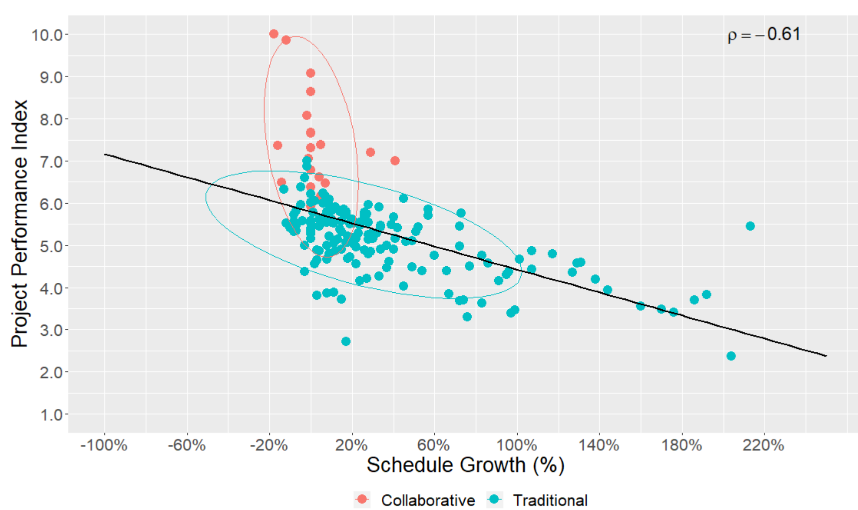


Figure 169: Plot of Project Performance Index vs Schedule Growth

The chart reveals a negative linear relationship for both Collaborative and Traditional projects. However, it can be noticed that Collaborative projects follow a sharper linear relationship as opposed to Traditional ones. Moreover, Collaborative projects experience on average higher PPI scores compared to Traditional ones with some overlap between PPI scores 5.0 and 6.5.

5.4.3.2 Schedule Factor

Considering the following figure, a moderate to strong negative correlation is found between the project performance index scores and the schedule growth. The correlation is statistically significant at the 95% confidence level with a *p*-value of 0.000 and a correlation coefficient of -0.59.

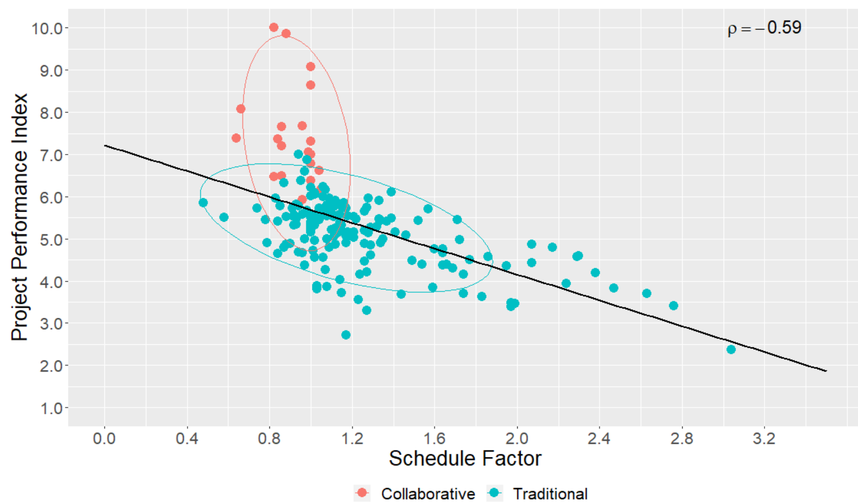


Figure 170: Plot of Project Performance Index vs Schedule Factor

The chart reveals a negative linear relationship for the both Collaborative and Traditional projects. However, it can be noticed that Collaborative projects follow a sharper linear relationship as opposed to Traditional ones. Moreover, Collaborative projects experience on average higher PPI scores compared to Traditional ones with some overlap between PPI scores 5.5 and 6.5.

It can also be noticed that while some Traditional projects experience the lowest schedule factor values in the dataset, their PPI scores are lower than most of the Collaborative projects, which can be explained by the poorer performance of those projects in other performance areas when compared to the Collaborative ones.

5.4.3.3 RFI Processing Time

Considering the following figure, a weak to moderate negative correlation is found between the project performance index scores and the RFI processing time. The correlation is statistically significant at the 95% confidence level with a *p*-value of 0.000 and a correlation coefficient of -0.34.

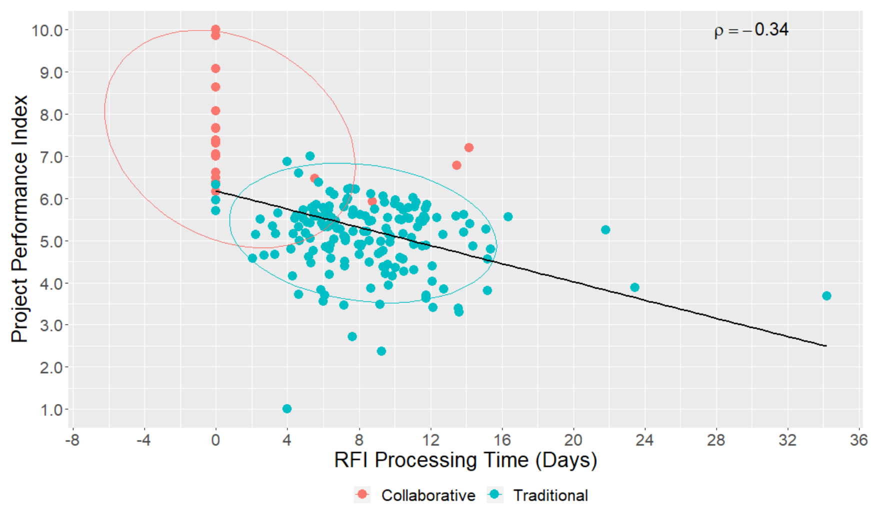


Figure 171: Plot of Project Performance Index vs RFI Processing Time

The chart reveals a negative linear relationship for the both Collaborative and Traditional projects. However, it can be noticed that Collaborative projects follow a much sharper linear relationship as opposed to Traditional ones, potentially due to the fact that RFIs are not processed by the owner under the DB approach and are thus presented with a zero from the owner perspective.

Moreover, Collaborative projects experience on average higher PPI scores compared to Traditional ones with some overlap between PPI scores 5.5 and 7.0.

5.4.3.4 Number of RFIs Per Million Dollars

Considering the following figure, a moderate to strong negative correlation is found between the project performance index scores and the schedule growth. The correlation is statistically significant at the 95% confidence level with a *p*-value of 0.000 and a correlation coefficient of -0.67.

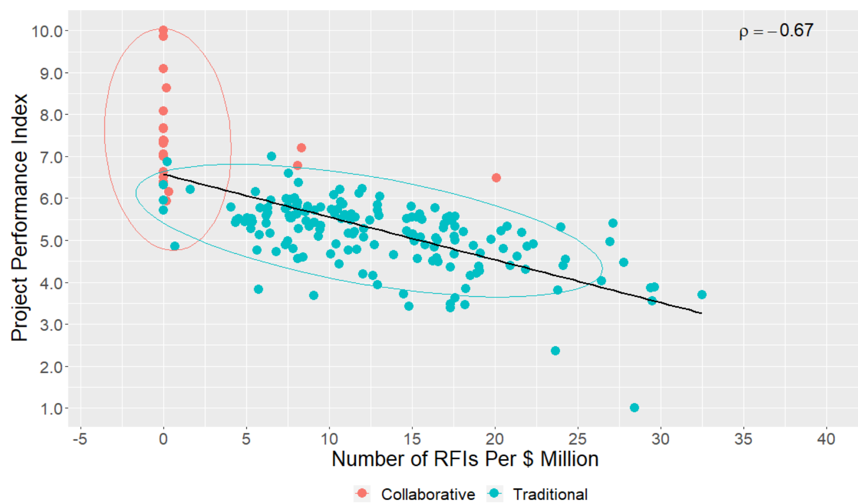


Figure 172: Plot of Project Performance Index vs Number of RFIs Per Million Dollars

The chart reveals a negative linear relationship for the both Collaborative and Traditional projects. However, it can be noticed that Collaborative projects follow a much sharper linear relationship as opposed to Traditional ones, potentially due to the fact that RFIs are not processed by the owner under the DB approach and are thus presented with a zero from the owner perspective.

Moreover, Collaborative projects experience on average higher PPI scores compared to Traditional ones with small overlap between PPI scores 6.0 and 7.0.

5.4.3.5 Number of Change Orders Per $ Million

Considering the following figure, a strong negative correlation is found between the project performance index scores and the number of change orders per million dollars. The correlation is statistically significant at the 95% confidence level with a *p*-value of 0.000 and a correlation coefficient of -0.78.

Figure 173: Plot of Project Performance Index vs Number of Change Orders Per Million Dollars

The chart reveals a negative linear relationship for the both Collaborative and Traditional projects. However, it can be noticed that Collaborative projects follow a much sharper linear relationship as opposed to Traditional ones. Moreover, Collaborative projects experience on average higher PPI scores compared to Traditional ones with small overlap between PPI scores 6.5 and 7.0.

5.4.3.6 Percentage Change

Considering the following figure, a moderate negative correlation is found between the project performance index scores and the project percentage change. The correlation is statistically significant at the 95% confidence level with a *p*-value of 0.000 and a correlation coefficient of -0.51.

Figure 174: Plot of Project Performance Index vs Project Percentage Change

The chart reveals a negative linear relationship for the Traditional projects in specific. It can be noticed that Collaborative projects follow a constant linear relationship as opposed to the negative relationship experienced by Traditional projects. Moreover, Collaborative projects experience on average higher PPI scores compared to Traditional ones with small overlap between PPI scores 6.0 and 7.0.

5.4.3.7 Construction Speed

Considering the following figure, a strong positive correlation is found between the project performance index scores and construction speed. The correlation is statistically significant at the 95% confidence level with a *p*-value of 0.000 and a correlation coefficient of 0.74.

Figure 175: Plot of Project Performance Index vs Construction Speed

The chart reveals a positive linear relationship for Collaborative and Traditional projects. However, it can be noticed that Traditional projects follow a sharper linear relationship as opposed to the Collaborative ones. Moreover, Collaborative projects experience on average higher PPI scores compared to Traditional ones with small overlap between PPI scores 6.0 and 7.0.

5.4.3.8 Delivery Speed

Considering the following figure, a moderate to strong positive correlation is found between the project performance index scores and construction speed. The correlation is statistically significant at the 95% confidence level with a *p*-value of 0.000 and a correlation coefficient of 0.63.

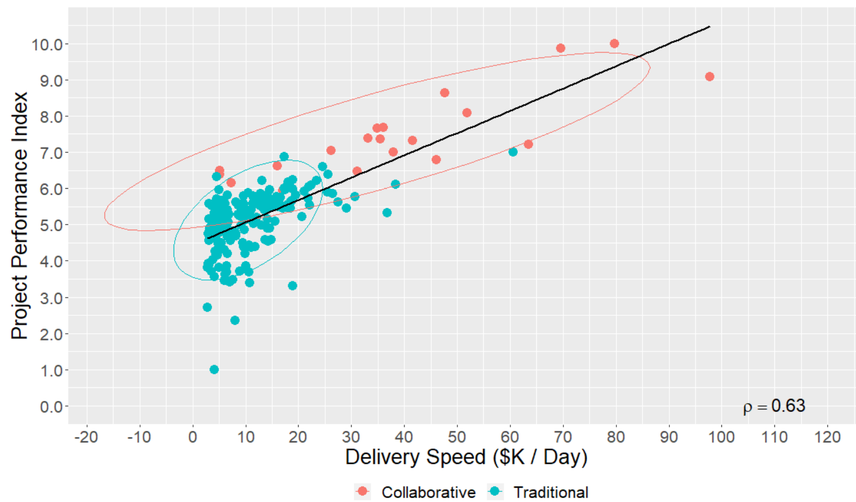


Figure 176: Plot of Project Performance Index vs Delivery Speed

The chart reveals a positive linear relationship for Collaborative and Traditional projects. However, it can be noticed that Traditional projects follow a slightly sharper linear relationship as opposed to the Collaborative ones. Moreover, Collaborative projects experience on average higher PPI scores compared to Traditional ones with small overlap between PPI scores 6.0 and 6.5.

5.4.4 Section Summary

This part of the analysis aimed at compiling the previous individual comparative performance assessment into a single score that captures most of the variance from the eight individual performance metrics. With that, the Project Performance Index model was developed for this purpose and then used to generate a combined project scores, which were then plotted against the different delivery methods.

One of key findings of this research was that at the 95% confidence level, DB outperformed each of SP and MP, CMAR outperformed each of SP and MP, and also SP outperformed MP. These findings provide substantial evidence of the higher performance of DB, followed by CMAR, then SP and MP when accounting for schedule, communication, change management and spending collectively.

On a similar note, the analysis concluded that Collaborative delivery methods experienced significantly higher PPI score at the 95% confidence level with an average score of 7.4 compared to the Traditional ones with an average score of 5.1, representing around 40% higher performance. As discussed earlier, the higher performance of Collaborative projects is believed to emerge from the teams' ability to coordinate various complex tasks in a collaborative environment that is project-driven rather than cost-driven. In addition to the delivery method, the selection process is a key factor in the existence of such environment by considering numerous factors such as past experience, past performance, and corporate's culture as criteria for the selection, which consequently reflects on the project performance.

Finally, an internal validation was conducted by plotting the developed PPI scores against each of the individual performance metrics to investigate their contribution to the score. Most of the contributions were moderate and strong with correlation coefficients higher than 0.6 and relative to the PPI weights, indicating an internally consistent and robust results.

5.5 Mathematical Model Validation Analyses

5.5.1 Confirmatory Factor Analysis

As part of the validation process, confirmatory factor analysis was utilized as a widely acceptable statistical technique for reducing dimensionality. The technique was applied twice, first using a single factor to replicate the results obtained by the mathematical model, and second using 3 factors to further explore the difference in performance across the four delivery methods.

It should be noted that at both analyses runs, an oblique rotation was favored over an orthogonal one, for the reasons previously explained in the Methodology chapter, using the oblimin algorithm. Further, the results have shown that for the case of a single factor, Generalized Least Squares (GLS) technique was able to capture the highest variance and lowest RMSR among the other techniques due to its ability to allocate smaller weights to variables with high uniqueness than those with low uniqueness (Aitken 1934). Whereas, in the case of generating three factors, the Principal Axis Factoring (PAF) and Maximum Likelihood (ML) provided comparably satisfactory results; however, PAF was selected due to its robustness to severely violated normality assumption, a data characteristic that was tested and highlighted in previous sections, unlike ML (Fabrigar et al. 1999).

5.5.1.1 Single Factor

The first part of the analysis returned a single factor that could resemble the developed PPI Index. The analysis produced the loadings as explained in the following table. The loadings produced from the factor analysis were rescaled to sum up to one in order to be compared to the original PPI weights.

Referring to the table, it can be noticed that the loadings distribution generated from the factor analysis is almost exact to that produced by the PPI model. Subtle differences were noticed with the biggest absolute difference at 1.62% and the smallest at 0.64%.

Performance Metrics	PPI Weights	FA Loadings	Rescaled FA Loadings	Difference (PPI – FA)
RFI Processing Time	8.65%	0.326	7.02%	1.62%
Percentage Change	9.10%	0.348	7.50%	1.60%
Schedule Growth	11.97%	0.619	13.33%	-1.37%
Schedule Factor	12.00%	0.624	13.44%	-1.44%
Number of RFIs / $Million	13.08%	0.566	12.19%	0.89%
Number of Change Orders / $M	14.70%	0.653	14.07%	0.64%
Delivery Speed	14.78%	0.729	15.70%	-0.92%
Construction Speed	15.72%	0.777	16.74%	-1.02%
Absolute Total				9.3%

The chart below also reveals another subtle difference with the order of metrics. Number of RFIs per Million Dollars was assigned a smaller weight than Schedule Growth and Schedule Factor using CFA, unlike the results from the PPI model. Nevertheless, this was the only ranking disagreement between the two models with a 1.25% difference (13.44% minus 12.19%).

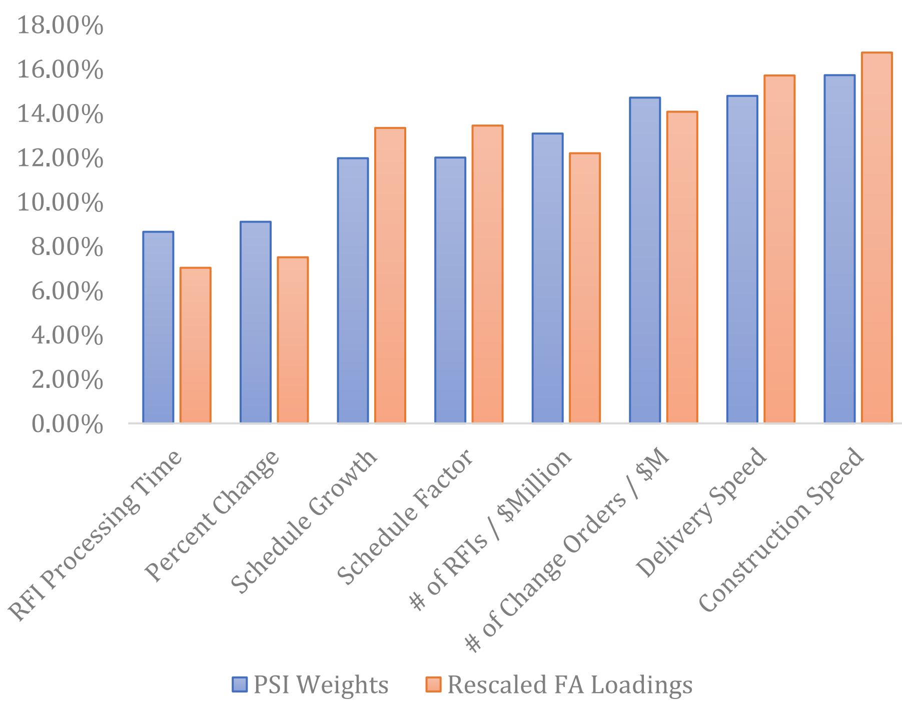


Figure 177: Comparative Bar Chart Between PPI Weights and FA Loadings

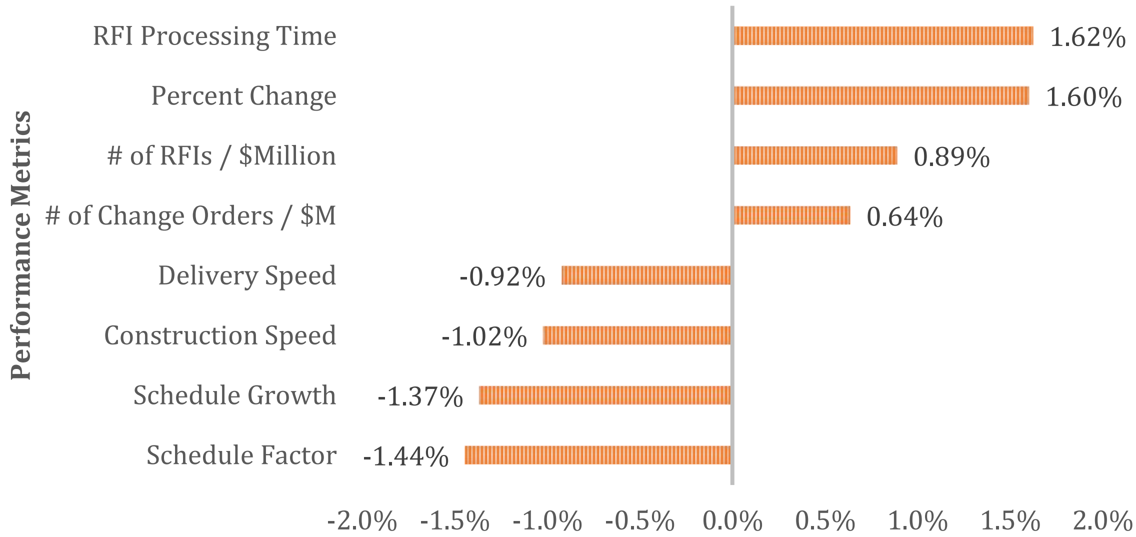


Figure 178: Difference Between PPI Weights and FA Loadings

The generated scores from the CFA were used to plot similar comparative boxplots for the various delivery methods, as well as the delivery types to validate the significance in performance.

5.5.1.1.1 <u>Comparative Analysis of Delivery Methods</u>

The following Q-Q plot is used to evaluate the normality assumption. The figure shows a clear departure from normality at the overall level of the dataset, as well as some levels of the delivery methods.

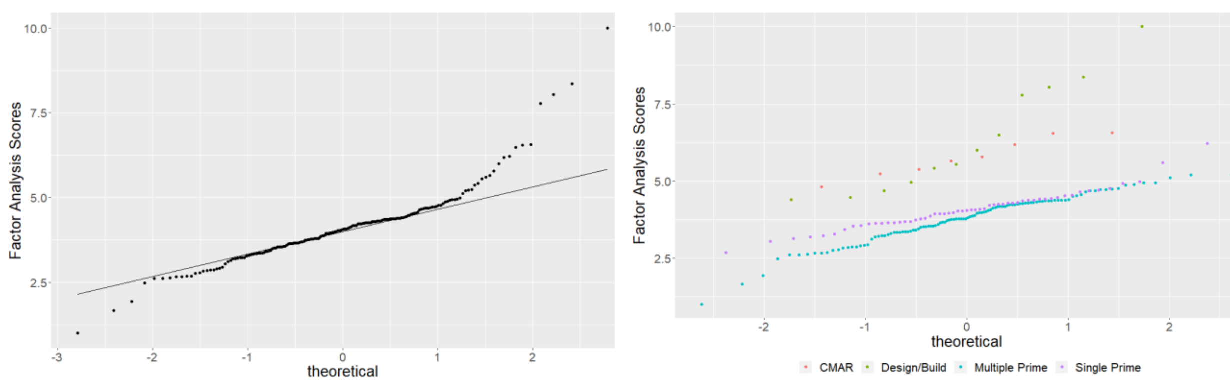


Figure 179: Q-Q Plot of Factor Analysis Scores

Q-Q plots below shows that normality was not present, which was confirmed at the 95% confidence level by the Shapiro-Wilk test, as shown in the table below.

Table 194: Significance of Normality for Factor Analysis Scores

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Design-Build	0.188	Not Significant	Not Significant
	CM at Risk	0.718	Not Significant	Not Significant
	Single-Prime	0.046	Significant	--
	Multiple-Prime	0.009	Significant	--

The following chart highlights the better performance of DB and CMAR delivery methods with average scores of 6.3 and 5.8 respectively, compared to SP and MP with average scores 4.1 and 3.8 respectively.

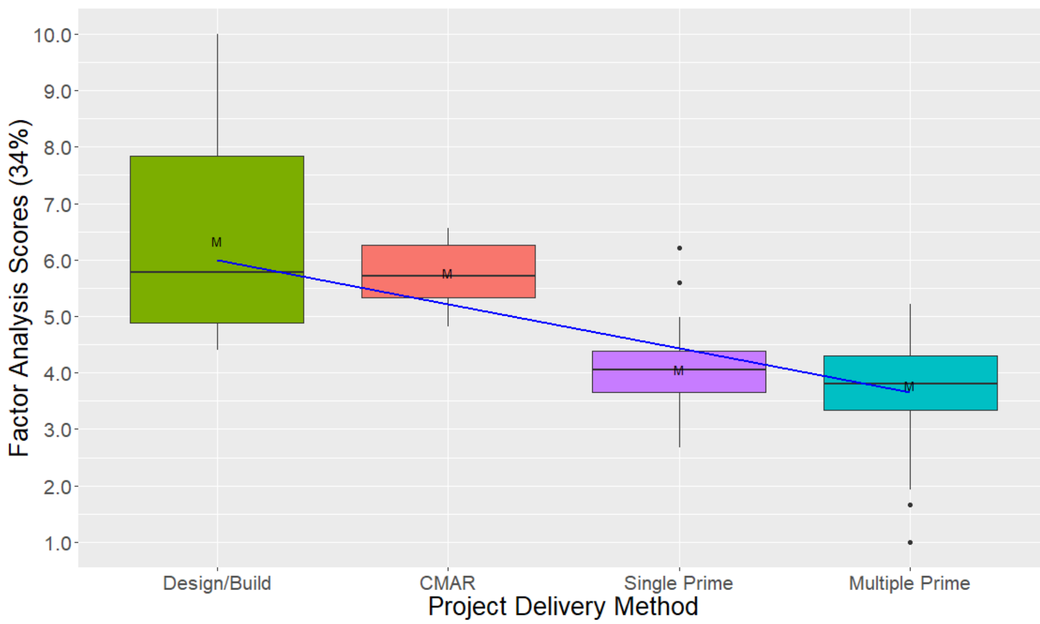


Figure 180: Comparative Boxplot of FA Scores by PDMs

Quantitative statistical tests were performed to confirm our previous findings. In this case, since the data did not follow a normal distribution, Fligner-Killeen is used to test the homogeneity of variances and Kruskal Wallis to test the significance across the four delivery methods.

With that, Fligner-Killeen test resulted in a *p*-value of 0.000, rejecting the null hypothesis of equal variance among the delivery methods, with a statistically significant difference at the 95% confidence level. This suggests a higher variability of DB compared to CMAR, SP and MP.

In addition, Kruskal-Wallis test resulted in a *p*-value of 0.000, rejecting the null hypothesis of equal population medians among the delivery methods. Following that, Conover-Iman post-hoc test resulted in a statistically significant difference between DB and each of SP and MP, CMAR and each of SP and MP and SP and MP, with no statistically significant difference between DB and CMAR.

Table 195: Significance of Distribution for Factor Analysis Scores

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.000	Significant	--
Kruskal-Wallis	Overall	0.000	Significant	--
Conover-Iman	DB vs Single	0.000	Significant	--
	DB vs Multiple	0.000	Significant	--
	CMAR vs Single	0.000	Significant	--
	CMAR vs Multiple	0.000	Significant	--
	DB vs CMAR	0.960	Not Significant	Not Significant
	Single vs Multiple	0.079	Not Significant	Significant

5.5.1.1.2 <u>Comparative Analysis of Delivery Types</u>

Q-Q plot is used again to evaluate the normality assumption. The following figure shows a clear departure from normality at the level of each delivery type.

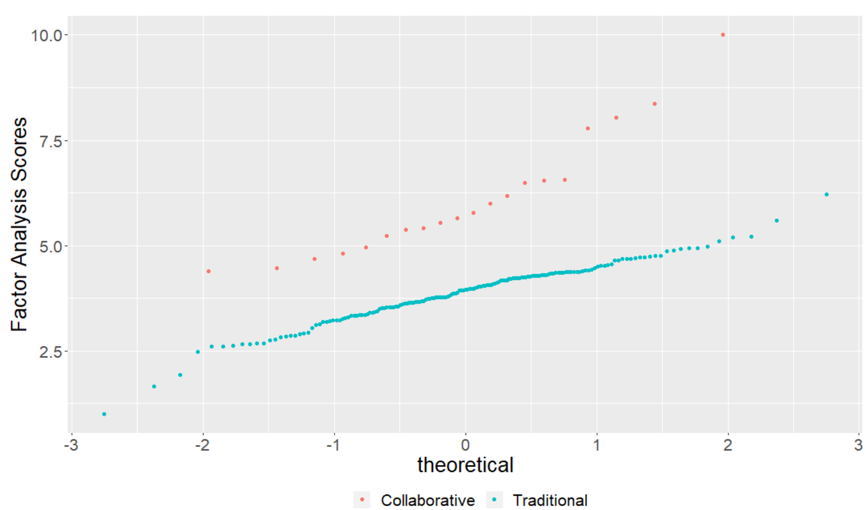


Figure 181: Q-Q Plot of Factor Analysis Scores

Shapiro-Wilk statistical test is conducted again to confirm our visual interpretation from the Q-Q plot. The table below highlights a statistically significant *p*-value at the overall level of the dataset, as well as the level of each delivery type, rejecting the null hypothesis of normal distribution at the 95% confidence level and confirming our interpretation from the Q-Q plot.

Table 196: Significance of Normality & for Factor Analysis Scores

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Shapiro-Wilk	Overall	0.000	Significant	--
	Collaborative	0.0355	Significant	--
	Traditional	0.004	Significant	--

Similarly, the following boxplot shows the higher performance of Collaborative projects with an average score of 6.1, as compared to Traditional ones with an average score of 3.9. The difference reveals that Collaborative projects outperforms Traditional projects by around 45%, based on the computations generated from the Factor Analysis.

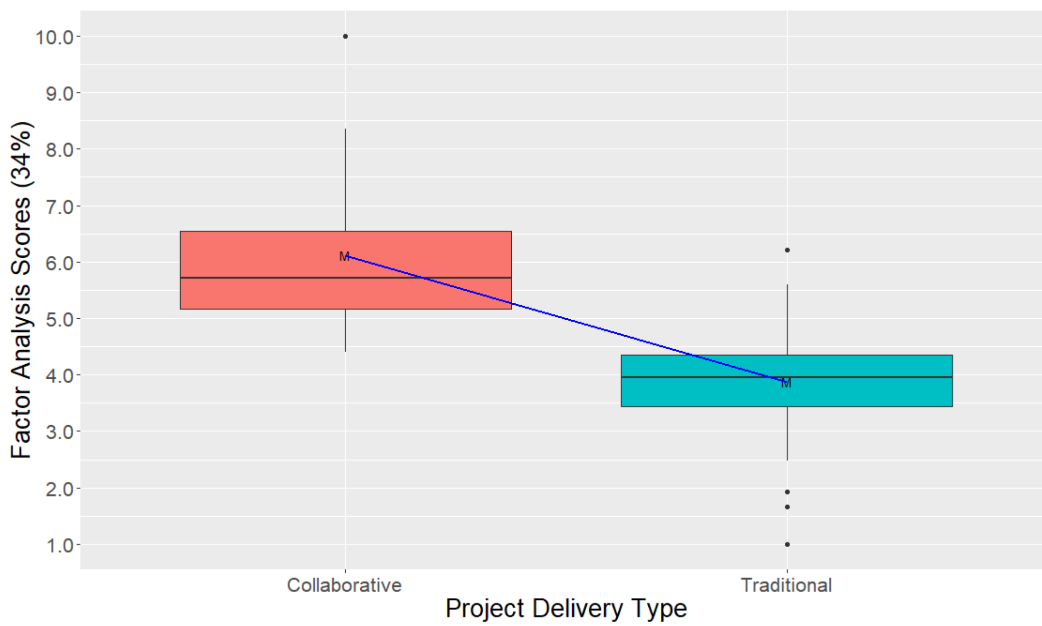


Figure 182: Comparative Boxplot of FA Scores by PDTss

Fligner-Killeen resulted in a *p*-value of 0.002, rejecting the null hypothesis of equal variance among the delivery types. This suggests that Collaborative delivery methods have higher variance than Traditional ones.

In addition, Mann-Whitney U test resulted in a *p*-value of 0.000, also rejecting the null hypothesis of equal medians among the two delivery types with a statistically significant difference at the 95% confidence level.

Table 197: Significance of Variance & Distribution for Factor Analysis Scores

Statistical Test	Parameter	*p*-value	Significant at 95% Level	Significant at 90% Level
Fligner-Killeen	Overall	0.002	Significant	--
Mann–Whitney	Overall	0.000	Significant	--

5.5.1.2 Triple Factors

This part of the analysis involved generating three factors using oblimin as an oblique rotation method and a Principal Axis Factoring (PAF) algorithm as an extraction method. The aim of this analysis was to provide validation, not only to the results previously generated, but also to our understanding of how factors are interrelated. It also provided additional insights regarding the performance of the various delivery methods and delivery types at the performance areas level, namely schedule, spending, etc.

The following table highlights that Schedule Growth and Schedule Factor (Factor 2), and Construction Speed and Delivery Speed (Factor 1) perform very similarly with each pair acquiring one factor, which will be called Schedule and Spending respectively. Factor 3, on the other hand, includes the remaining four metrics, which will be called Change & Communication. Two of the four metrics had very high loadings and two barely acceptable above the 0.3 rule-of-thumb (Hooper et al. 2008; Kline 2011).

The table however represents a well-established simple structure, with six of the metric loadings exceeding 0.8 for a single factor and negligible loadings for the rest. Two of the metrics were assigned small loadings; Percentage Change had negligible loadings for factors 1 and 2, which is acceptable, while RFI Processing Time experienced cross-loadings with less than 0.2 difference between factors 1 and 3. The overall structure is rationale, further testing was conducted to evaluate the goodness of fit.

Furthermore, it can be noticed that the three factors represent 65% of the variance in the data. The first factor contributes to 23%, second factor 22% and third factor 20%, which is a representative percentage given that premise of Factor Analysis is not to maximize the variance contribution, but rather explore performance similarities among the metrics.

Table 198: Factor Loadings for Confirmatory Factor Analysis

Performance Metric	Factor Loadings		
	Factor 1 *(Spending)*	Factor 2 *(Schedule)*	Factor 3 (*Change & Comm.*)
Percentage Change	--	--	0.32
Schedule Growth	--	0.96	--
Schedule Factor	--	0.93	--
Number of Change Orders / $M	--	--	0.82
Number of RFIs / $Million	--	--	0.83
RFI Processing Time	0.13	--	0.29
Construction Speed	0.93	--	--
Delivery Speed	0.93	--	--
% of variance explained	23%	22%	20%
Total % of variance explained	23%	45%	65%

Figure 183 highlights the performance of each of the three generated factors: Spending (x-axis), Schedule (y-axis) and Change & Communication (z-axis). Collaborative delivery methods such as DB and CMAR show significantly higher performance in regard to Spending and Change & Communication in comparison to Traditional ones such as SP and MP. These findings provide additional evidence that Collaborative projects could deliver projects much quicker during design and construction, facilitate to a larger extent the communication between the project parties, and provide better change management than their Traditional counterparts.

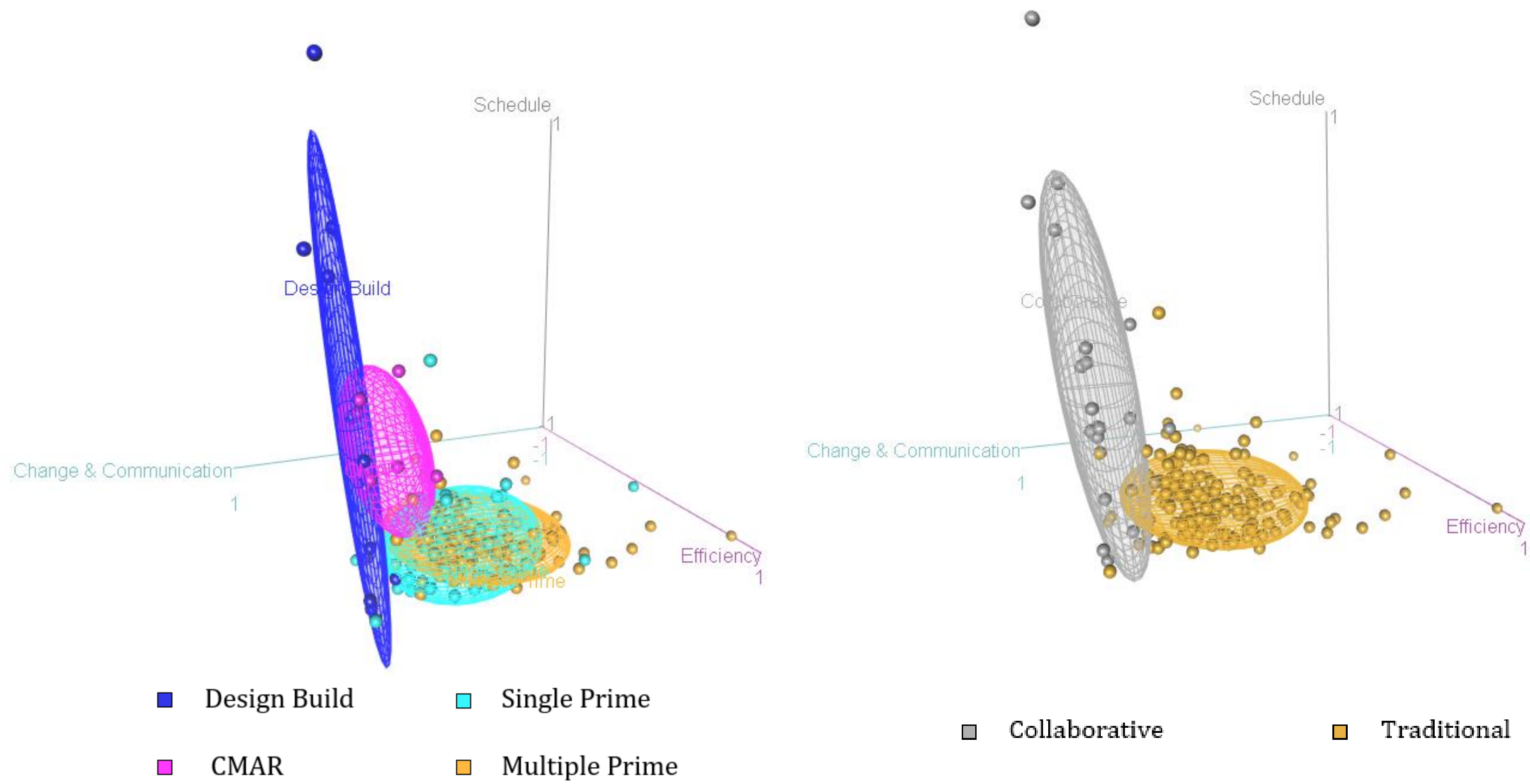


Figure 183: Factor Analysis with Triple Factors Against PDMs & Types

As explained earlier, PAF was selected as an extraction method due to its robustness to the violation of normality, unlike Maximum Likelihood (ML) that requires a relatively normally distributed data (Fabrigar et al., 1999). PAF also satisfied all the identified goodness-of-fit thresholds as presented in the following table:

Table 199: Factor Analysis Results Against Goodness-of-Fit Thresholds

Goodness-of-Fit Parameter	Threshold	Results Using PAF	Quality of Model Fit
Root Mean Square Error of Approximation (RMSEA)	<0.1	0.095	Acceptable
Root Mean Square Residuals (RMSR)	Approx. 0	0.02	Acceptable
Tucker Lewis Index (TLI)	>0.9	0.94	Acceptable
p-value of Empirical Chi-Square*	>0.05	0.66	Acceptable

* Empirical Chi-Square was utilized in lieu of likelihood Chi-Square as a nonparametric statistic as the data violated the normality assumption

The quantitative results in the previous table provide sufficient evidence of a good fitting model. The fitting validation was very essential to ensure that the interpreted results do in fact represent the original data, which translates to a high confidence in the findings and conclusions.

5.5.2 Partitioning Clustering Analysis

In addition to CFA, clustering analysis was used to validate the conclusions previously drawn from the PPI model. Using this technique, two sub-analyses were conducted: k-means algorithm to generate two clusters, which resembles the two delivery types, and PAM algorithm to generate four clusters, which resembles the four delivery methods.

First the dissimilarity matrix was create using the Manhattan distance, highlighting relatively similar projects in red and dark orange, and highly dissimilar projects in white and blue as shown in Figure 184. Each vertical column and horizontal row represent the dissimilarity of one specific observation (i.e. project) with the rest of the dataset.

With that, it is noticed that the first set of projects are highly similar to each other, and highly dissimilar to the rest of the dataset, indicating high potential of a good clustering. Also, the same trend can be seen towards the end of the matrix with a group of projects performing much similar as compared to their counterparts. However, 2 to 3 projects seem to be similar to a big portion of the dataset, which reduces their ability to be clustered efficiently.

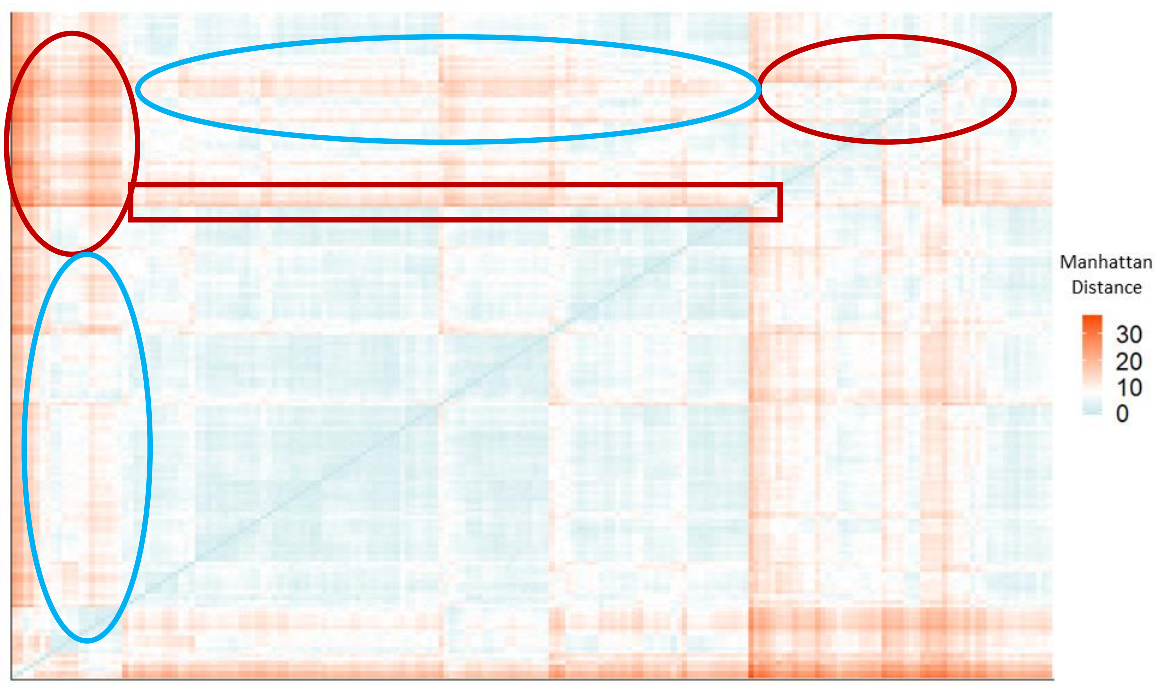


Figure 184: Heat Map of Dissimilarity Matrix

To explore the dissimilarity matrix further, the following figures are generated to reveal dissimilarities between the different delivery methods and delivery types. The higher intensity of the green, the larger the distance between an observation (i.e. group) or a group of observations; whereas, less intense lines represent smaller distances between observations, and hence, higher similarity.

Referring to the following Figure 185, nodes belonging together are placed towards the center while those with higher dissimilarity are placed away from the center. The figure shows that Collaborative projects overall experience smaller distances and much similar traits as opposed to Traditional ones.

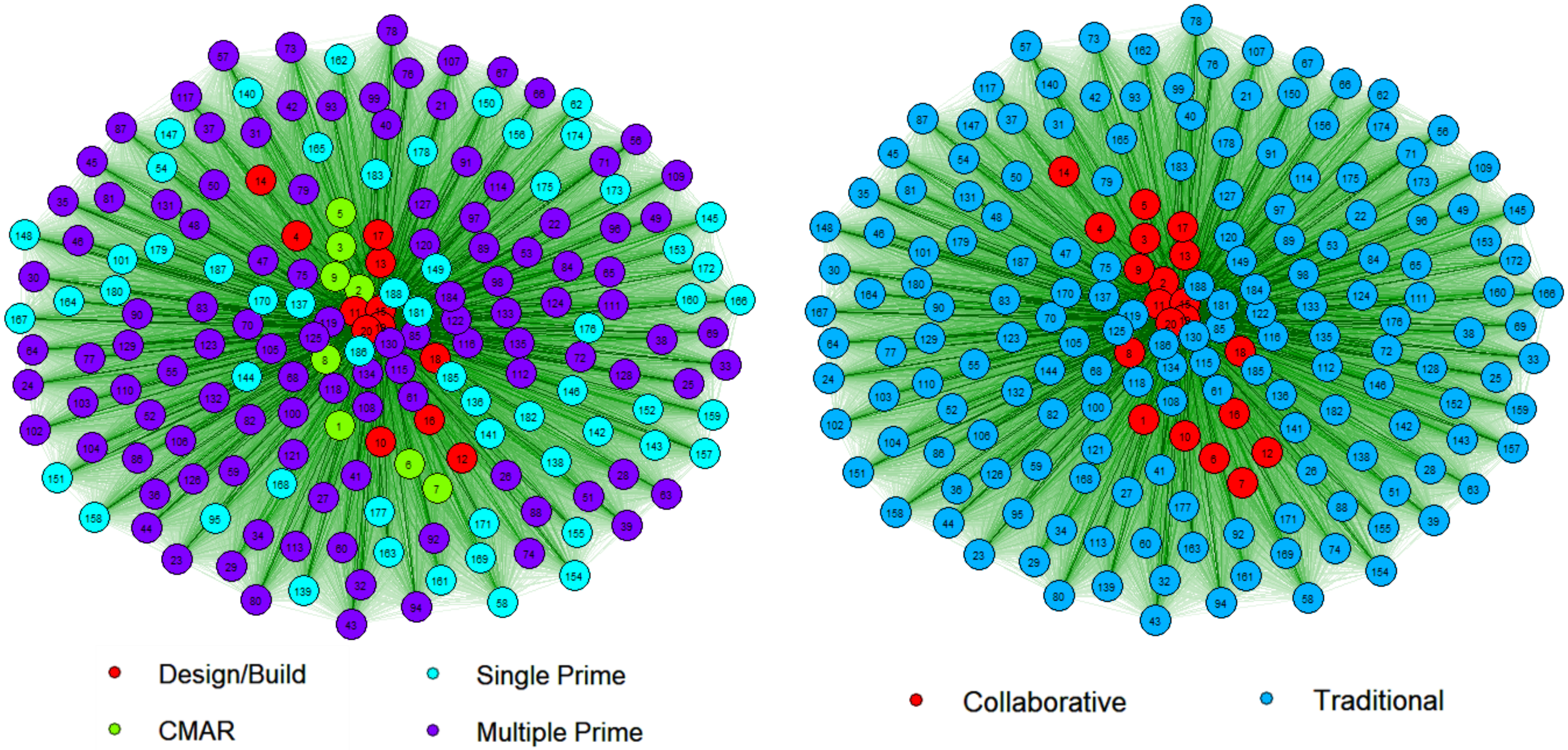


Figure 185: Graphical Representation of the Dissimilarity Matrix

The below Figure 186, provides additional insights about the existing dissimilarity between delivery methods. It can be noticed that overall DB and CMAR on the one hand experience intense green color when compared to SP and MP delivery methods. Whereas, the intensity of the color is significantly less when comparing DB and CMAR, or SP and MP, indicating higher level of similarity between each of those pairs.

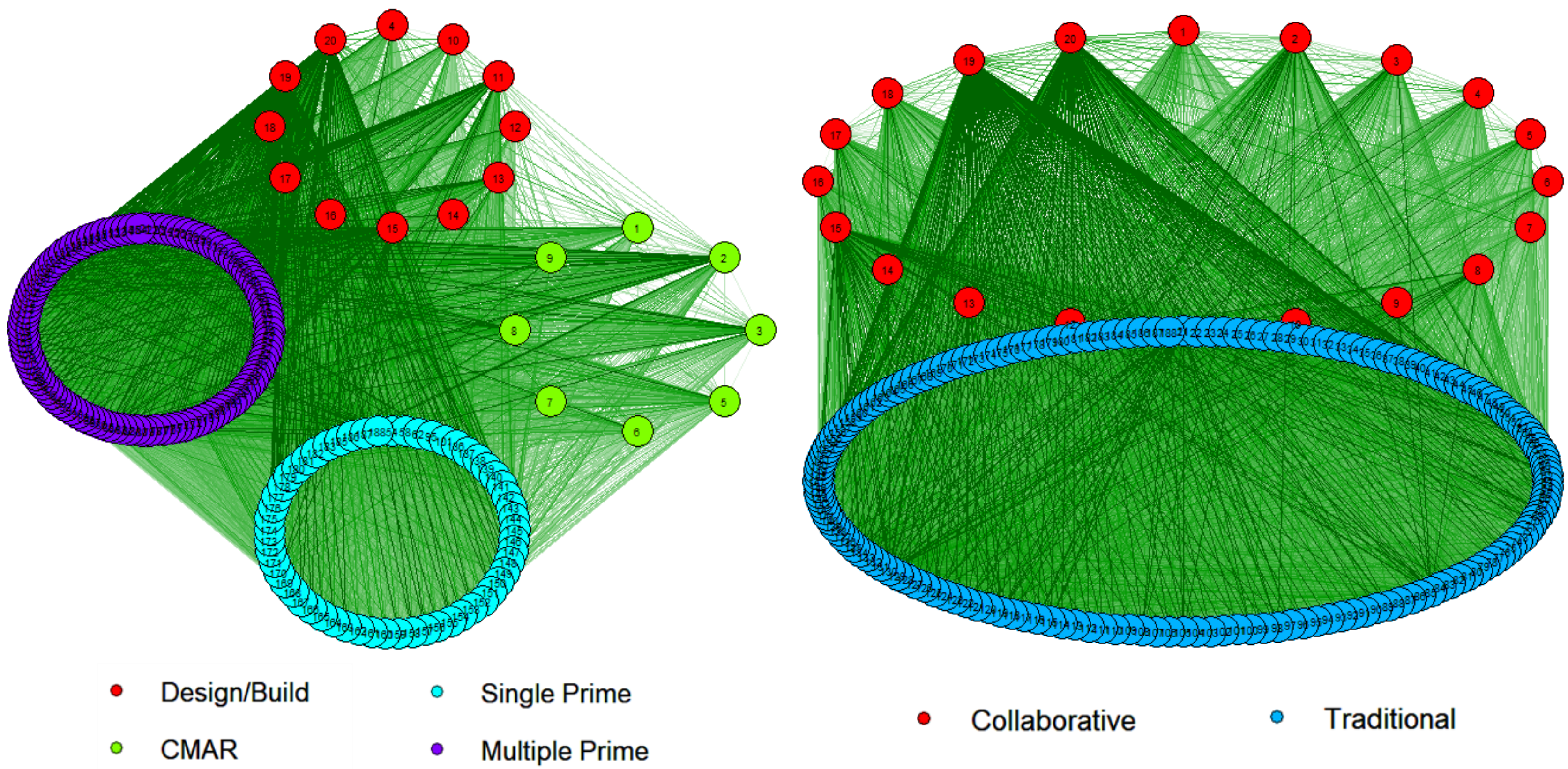


Figure 186: Graphical Representation of the Dissimilarity Matrix

Finally, Figure 187 confirms our understanding as it sorts the projects by their average distance compared to other projects. While the figure shows some overlap between DB and CMAR projects, as well as SP and MP projects, both Collaborative and Traditional types are highly dissimilar overall. This validates our understanding from previous plots and provides informative insights about the dissimilarity in performance of the different delivery methods in specific and delivery types in general.

Following that, Partitioning Clustering Analyses are conducted to provide statistical evidence of these findings.

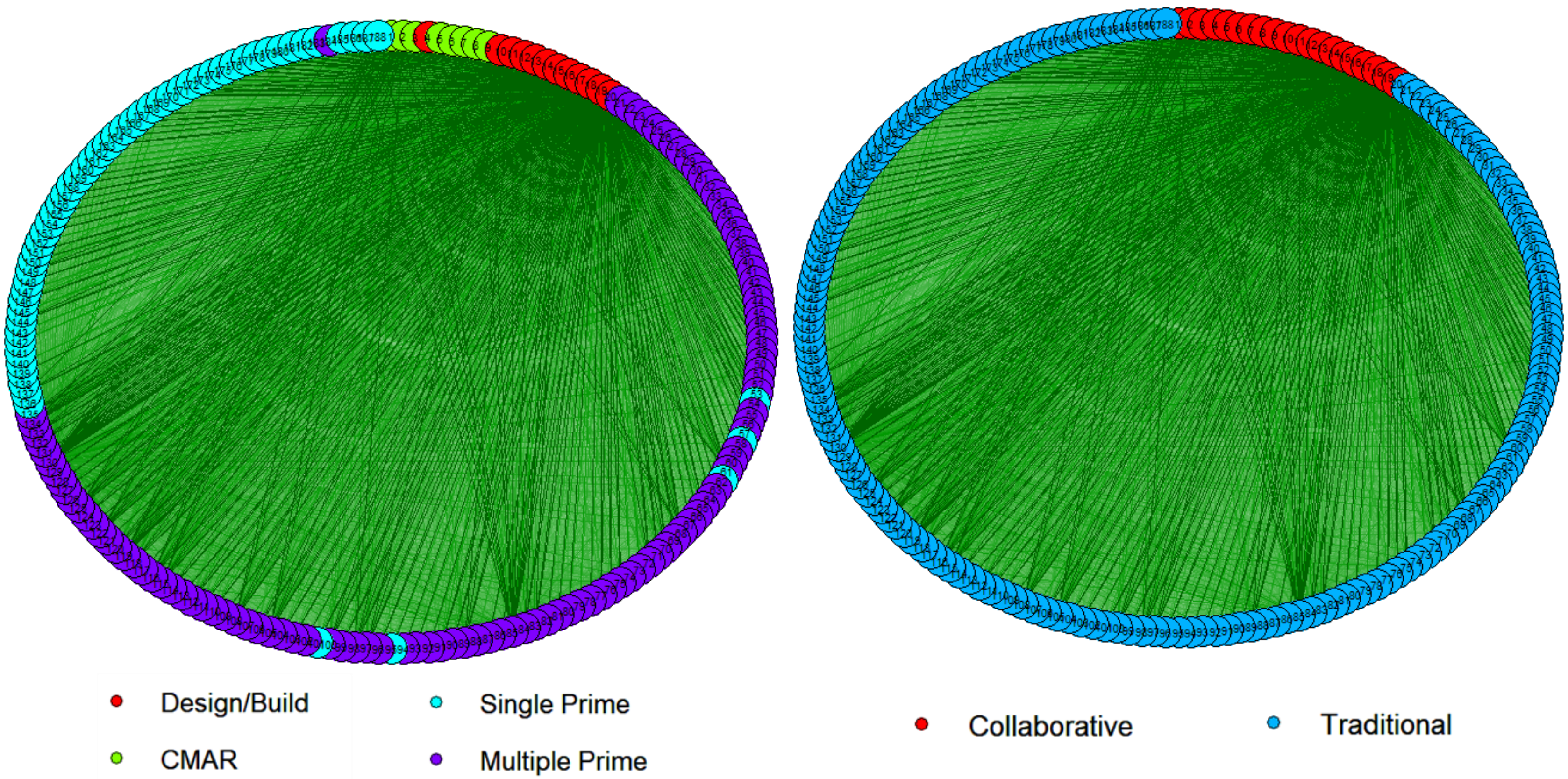


Figure 187: Graphical Representation of the Dissimilarity Matrix

5.5.2.1 K-Means Clustering Algorithm

The k-means clustering algorithm reported the grouping shown below in Figure 188 and Table 200. The analysis reported that 100% of the DB projects and more than 85% of the CMAR projects were assigned to cluster 1. Conversely, only around 7% of the SP projects and 4.5% of the MP projects were assigned to cluster 1 with the remaining majority of projects assigned to cluster 2.

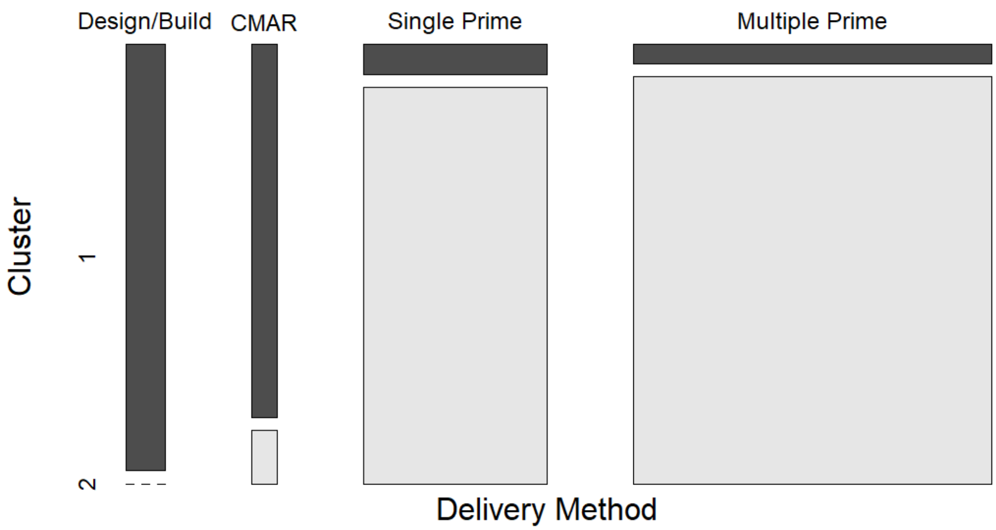


Figure 188: Mosaic Plot Showcasing Two Clusters Against PDMs

Table 200: Contingency Table of Two Clusters & PDMs

	Project Delivery Method				
	Design/Build	**CMAR**	**Single-Prime**	**Multiple-Prime**	***Row Total***
Cluster 1	12	7	4	5	**28**
Cluster 2	0	1	53	106	**160**
Column Total	**12**	**8**	**57**	**111**	**188**

To summarize the results, a similar analysis is performed on the delivery types and further confirms the same understating, 95% of the Collaborative projects were assigned to cluster 1, and only 5% of the Traditional projects performed similarly and were assigned to the same cluster. Conversely, the remaining 95% of Traditional projects were classified as tier-2 (cluster 2).

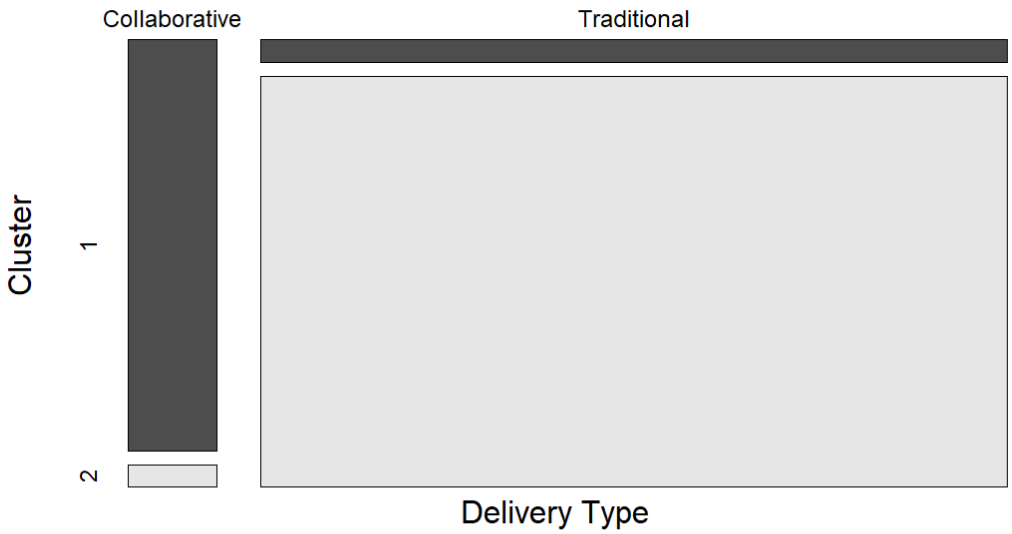


Figure 189: Mosaic Plot Showcasing Two Clusters Against PDTs

Table 201: Contingency Table of Two Clusters & PDTs

	Project Delivery Method		
	Collaborative	**Traditional**	***Row Total***
Cluster 1	19	9	**28**
Cluster 2	1	159	**160**
Column Total	**20**	**168**	**188**

The observations matrix was initially set so that all the vector metrics have positive direction (i.e. the higher the metric the better). With that, negative metrics like schedule growth were assigned a negative value to reverse their direction to the-higher-the-better direction. As a result, the centroid (mean) of each cluster pertains the same methodology (i.e. the higher the better) as well. Cluster 1 performed better than cluster 2 (centroid values of 8.2 and -1.4 respectively).

These findings are significant as they provide statistical evidence of the better performance of DB and CMAR (mostly assigned to cluster 1), against SP and MP (mostly assigned to cluster 2).

Similarly, Principal Component Analysis (PCA) was used to visually represent the clustering distribution with first component on the x-axis and second component on the y-axis. The plot reveals the distinct classification of projects with limited overlap between the clusters.

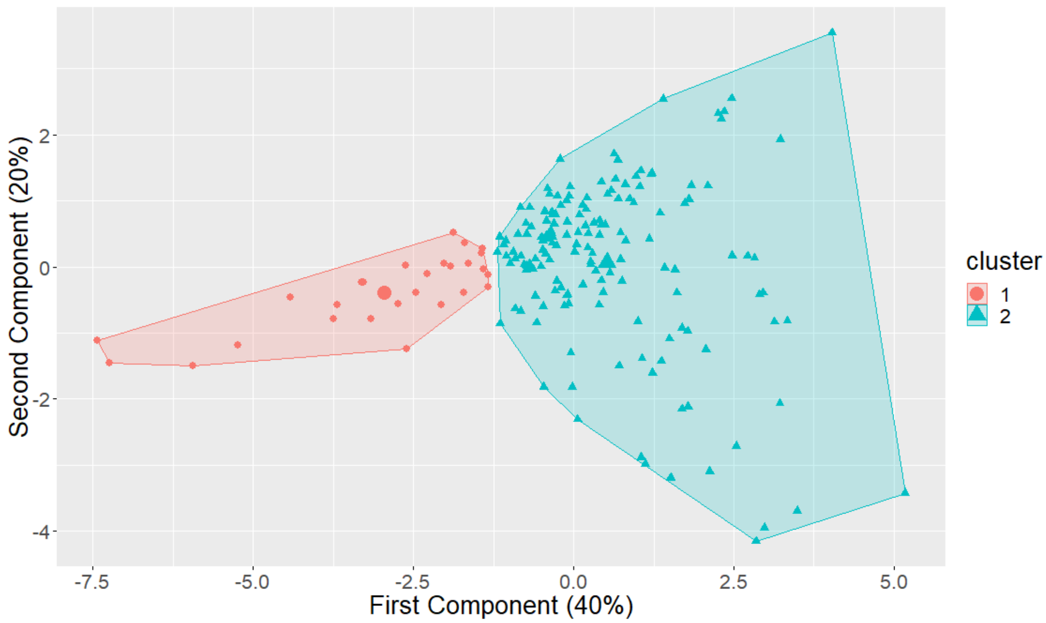


Figure 190: Cluster Plot Showcasing Two Data Clusters

As explained earlier, the higher the silhouette width, the higher evidence of a well-clustered model is present. This heuristic approach can be interpreted as follows:

- Large silhouette Si (almost 1) are very well clustered.
- Small Si (around 0) means that the observation lies between two clusters.
- Negative Si are probably placed in the wrong cluster.

Figure 191 shows an average silhouette width of 0.35, which is acceptable given that the objective of this analysis is to provide validation not to solely drive conclusions. Moreover, it is noticed that a limited number of projects were potentially incorrectly classified as cluster 1, while no incorrectly classified projects from cluster 2 are present.

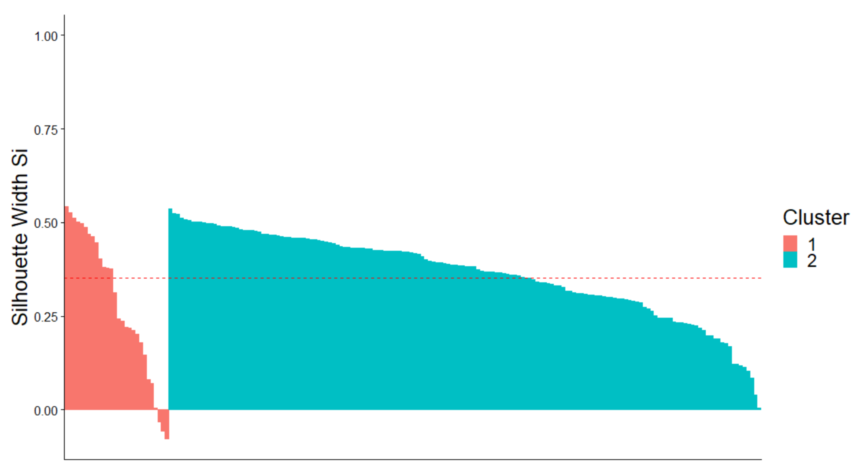


Figure 191: Silhouette Plot of Two Data Clusters

5.5.2.2 Partitioning Around Medoids (PAM) Clustering Algorithm

The PAM clustering algorithm reported the grouping shown below in Figure 192 and Table 202. The analysis reported that more than 90% of the DB projects and 75% of the CMAR projects were assigned to cluster 1. Conversely, only around 5% of the SP projects and 0% of the MP projects were assigned to that cluster. Moreover, it can be noticed that 100% of DB projects and 100% of CMAR projects were assigned to the highest 2 clusters (clusters 1 & 2) with no projects classified as clusters 3 or 4. On the other hand, around 30% of SP projects and around 50% of MP projects were assigned to the lowest two clusters (clusters 3 & 4). This reveals the significant difference in performance between the Traditional vs Collaborative delivery methods.

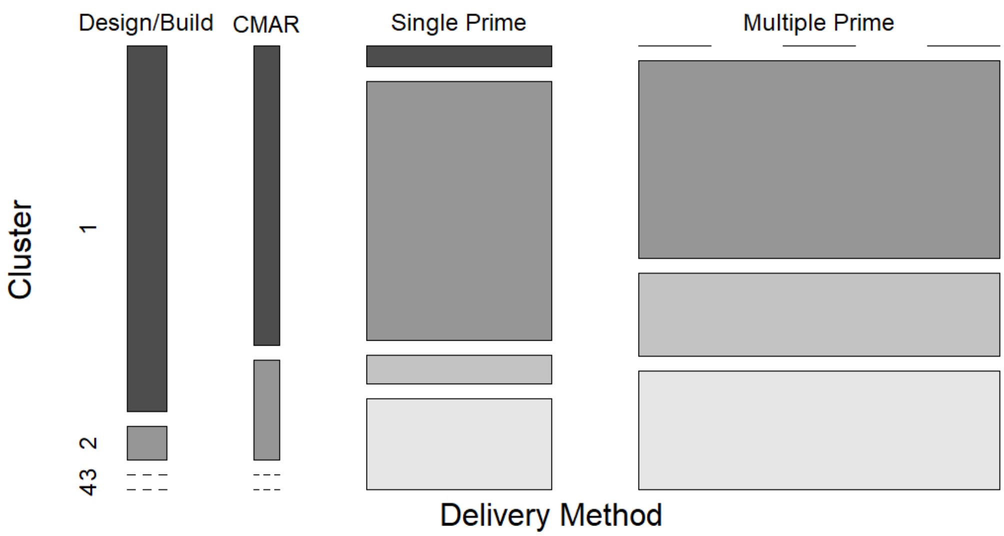


Figure 192: Mosaic Plot Showcasing Four Clusters Against PDMs

Table 202: Contingency Table of Four Clusters & PDMs

	Project Delivery Method				***Row Total***
	Design/Build	**CMAR**	**Single-Prime**	**Multiple-Prime**	
Cluster 1	11	6	3	0	**20**
Cluster 2	1	2	37	55	**95**
Cluster 3	0	0	4	23	**27**
Cluster 4	0	0	13	33	**46**
Column Total	**12**	**8**	**57**	**111**	**188**

To summarize the results, a similar analysis is performed to the delivery types. Results confirm same understating - 85% of the Collaborative projects were assigned to cluster 1, while only 1.7% of the Traditional projects performed similarly and were assigned to that highest cluster. Moreover, clusters 3 and 4 were fully populated by Traditional projects only. On the other hand, cluster 2 was split between Traditional and Collaborative projects.

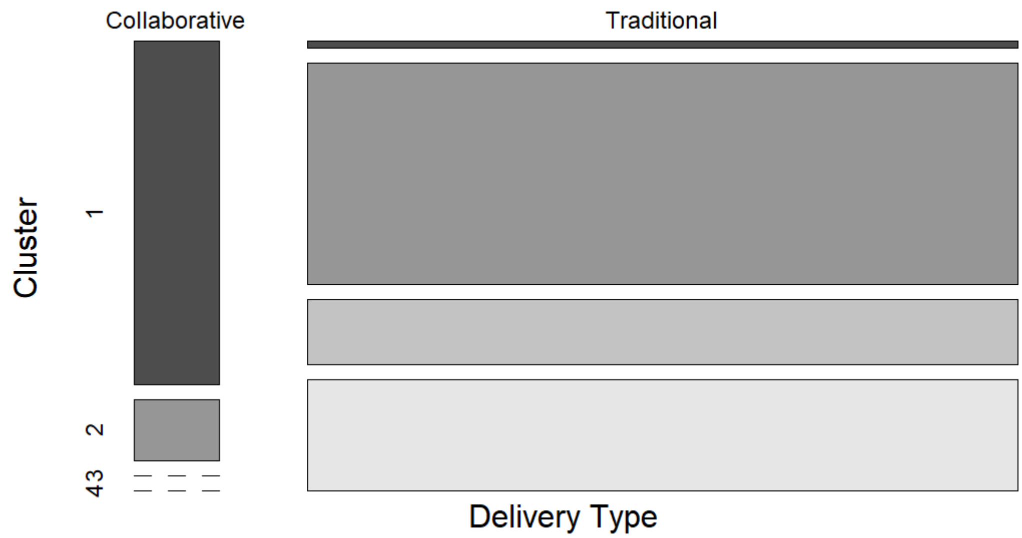


Figure 193: Mosaic Plot Showcasing Four Clusters Against PDTs

Table 203: Contingency Table of Four Clusters & PDTs

	Project Delivery Method		
	Collaborative	**Traditional**	***Row Total***
Cluster 1	17	3	**20**
Cluster 2	3	92	**95**
Cluster 3	0	27	**27**
Cluster 4	0	46	**46**
Column Total	**20**	**168**	**188**

These findings provide statistical evidence of the better performance of DB and CMAR (mostly assigned to cluster 1), against SP and MP (mostly assigned to clusters 2, 3 and 4).

Similarly, Principal Component Analysis (PCA) was used to visually represent the clustering distribution with first component on the x-axis and second component on the y-axis. The plot reveals distinct classification of projects in clusters 1 and 4, with some overlap between clusters 3 and 4.

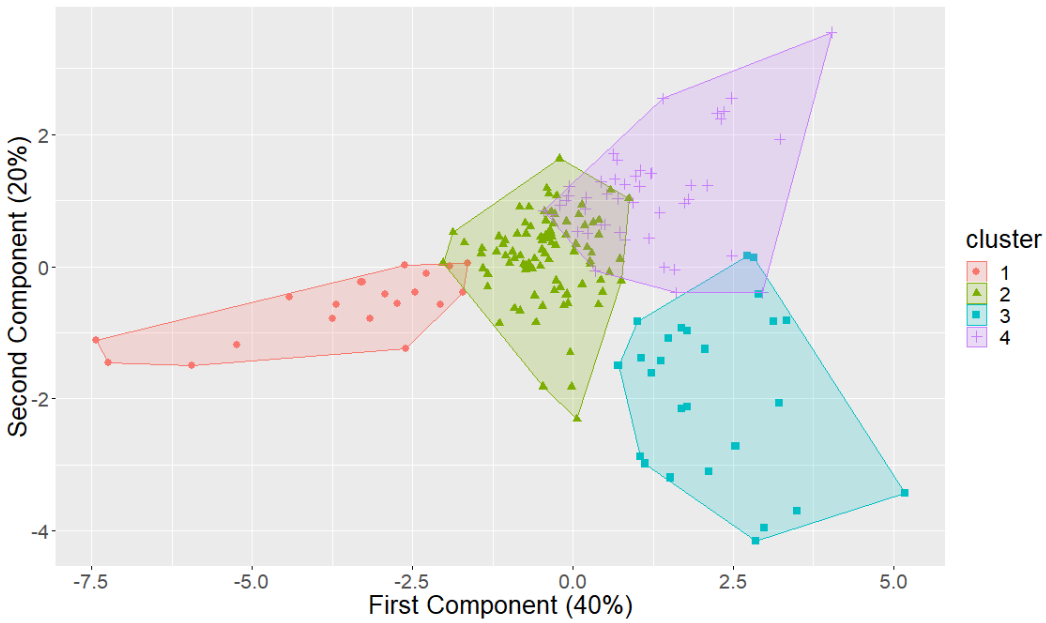


Figure 194: Cluster Plot Showcasing Four Data Clusters

Figure 195 below reveals an average silhouette width of 0.25, which is acceptable given that the objective of this analysis is to provide validation not to solely drive conclusions. Moreover, a limited number of projects were potentially incorrectly classified as clusters 1, 2 and 3, while higher number of potentially incorrectly classified projects were present at cluster 4.

One potential explanation of this phenomenon is that some projects performed very different than the rest of those assigned to cluster 4. Further analysis with more clusters should be performed to investigate the loosely fitting observations

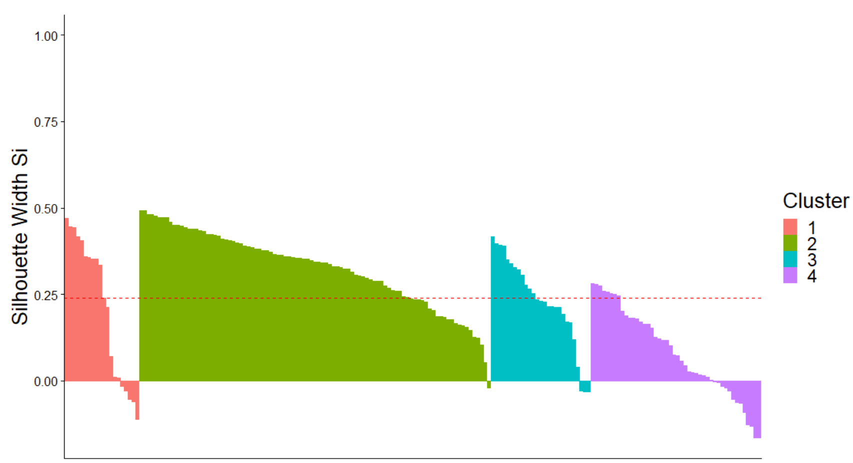


Figure 195: Silhouette Plot of Four Data Clusters

5.5.3 Section Summary

In this section two separate and well-stablished statistical techniques were used to validate the results previously generated using the developed Project Performance Index mathematical model. Confirmatory Factor Analysis (CFA) as well as Partitioning Clustering Analysis techniques confirmed earlier findings about the better performance of DB and CMAR on the one hand, over SP and MP on the other hand. Nevertheless, it should be noted that the results could have been impacted due to a set of limitations that include: type and quality of the collected data, as well as the selected statistical algorithms. Future research could provide different perspectives to the conclusions made in this section.

First, CFA was conducted twice using a single factor, for comparison with PPI scores, followed by another assessment using three factors, to capture higher variance and be visually presentable. When generating a single factor, the results were almost identical to those obtained from the PPI model, having negligible difference between the weights generated from both models with the highest absolute difference being 1.62% and the lowest 0.64%. The results have aligned with the previous conclusion as Collaborative projects were found to significantly outperform Traditional ones at the 95% confidence level by 45% on average, versus 40% performance difference generated from the PPI model.

Furthermore, when generating three factors, representing Spending, Schedule and Change & Communication performance areas, the results provided visual evidence that Collaborative projects are delivered much quicker, with higher communication levels among project participants and better performance over project changes. The results revealed also that Collaborative projects have, on average, better schedule performance compared to Traditional ones, with some overlap between the two types.

Second, Partitioning Clustering Analysis was utilized as means of validating the conclusions previously drawn from the PPI model and the CFA. Using this technique, two sub-analyses were conducted using two clusters, to resemble the two delivery types, and four clusters, to resemble the four delivery methods. When splitting the dataset into two clusters using k-means algorithm, it was noticed that 95% of the Collaborative projects were allocated to the higher performance cluster (cluster 1), whereas, 95% of the Traditional projects were allocated to the lower performance cluster (cluster 2) with 5% of overlap in each case.

In addition, a second assessment was conducted to drive more in-depth conclusions by splitting the dataset into four clusters using PAM algorithm. The results indicated that 85% of the Collaborative projects and only 1.7% of Traditional projects were allocated to the highest cluster (cluster 1), and that the lowest two clusters (clusters 3 and 4) were fully populated by Traditional projects. The second highest cluster (cluster 2) contained the remaining portions - 15% of Collaborative projects and ~50% of the Traditional projects.

Although the results suggested that Traditional projects can in some situations be on par with Collaborative ones (i.e. be allocated to the same cluster), the likelihood of such similarity in performance was very low at only 5% when considering two clusters and 1.7% when considering four clusters.

5.6 Comparative Performance of Project Participants

This part of the analysis was developed with the motivation of exploring additional factors that could impact the performance of State projects. In addition to previously exploring the delivery method as a potential significant factor, this section studies the performance of the University of Wisconsin-System campuses.

5.6.1 Performance of UW-System Campuses

In this section, the data included 13 UW-System campuses as shown in the following figure. UW-Madison comprised the highest percentage of 41% (57 projects), while UW-Parkside comprised the lowest at 2% (3 projects) with a total number of 140 UW projects.

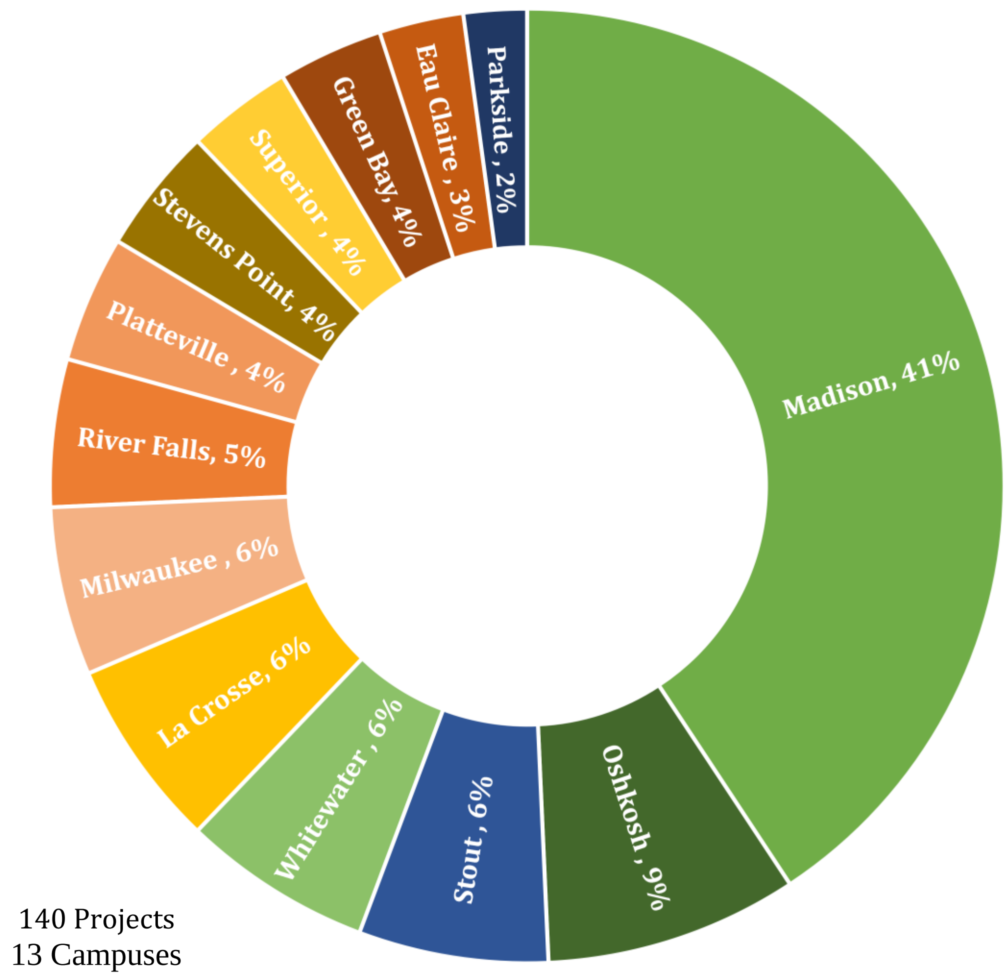


Figure 196: Distribution of Projects by UW-System Campuses

PPI scores were used to rank the campuses based on the average score of the projects executed under their supervision. The intent was to explore whether a noticeable difference exists between projects executed under different supervisions.

As shown in the figure below, the highest average performance was achieved by UW-River Falls at 5.7 points, whereas the lowest average performance is associated with UW-Milwaukee at 4.6 points. The overall average PPI score was 5.2, represented by UW-Platteville. It is also worth mentioning that UW Madison was ranked 3rd with an average score of 5.6, which is substantial given its large portfolio of executed projects.

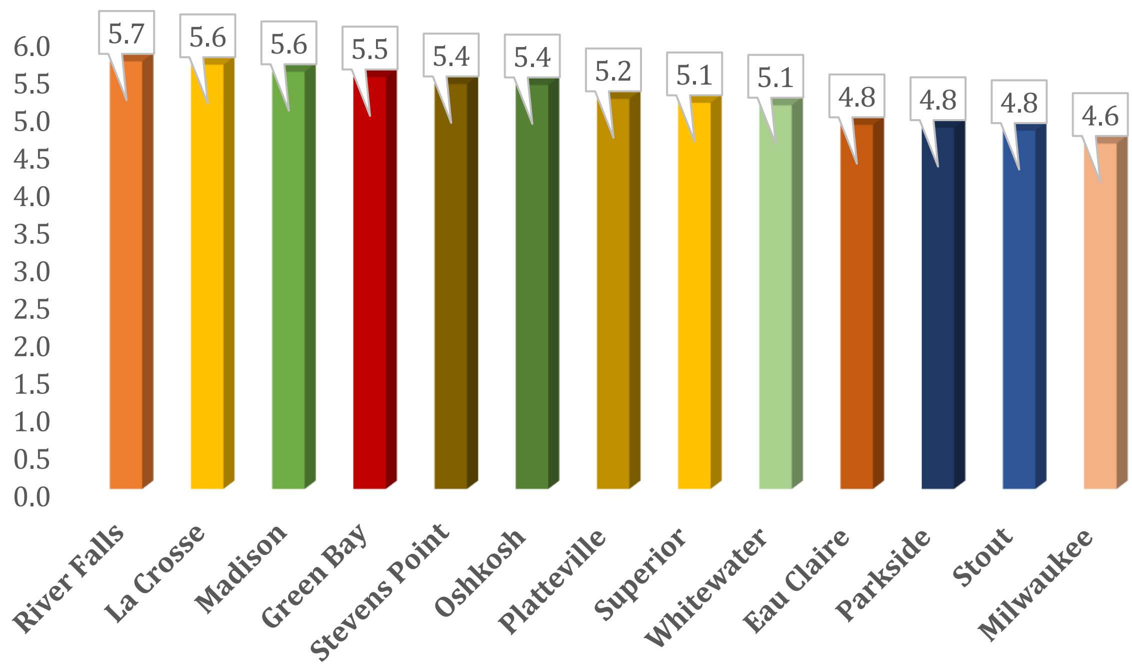


Figure 197: Ranked Average PPI Scores by UW-System Campuses

Table 204 provides additional information about the performance of the 13 campuses in each of the eight performance metrics, upon which the PPI score depends.

Considering Madison and Milwaukee campuses for example. The analysis highlighted that the performance of the Madison campus exceeded the average in 5 metrics, met the average in 1 (schedule factor) and was below than average in 2 (schedule growth and percentage change). Milwaukee campus, on the other hand, exceeded the average in 2 metrics only (RFI processing time and construction speed), met the average in 1 (delivery speed) and was lower than average in the remaining 5.

These findings are very informative to public owners as it sheds light on areas with potential for performance enhancement. For instance, the Milwaukee campus experienced higher than average construction speed, yet average delivery speed, which could imply a longer-than-average design and pre-construction planning duration compared to its neighboring campuses.

The interpretation of these analysis can vary significantly. However, the baseline in this research was to provide additional inputs to decision makers, project managers and campus management teams to consider, as well as initiate a collaborative discussion about the current practices.

Table 204: Descriptive Statistics of Performance Metrics Per UW-System Campuses

Parameter	Schedule Growth	Schedule Factor	RFI Processing Time	# of RFIs per $M	# of Cos per $M	Percentage Change	Construction Speed	Delivery Speed	PPI Score
River Falls	7.2%	1.1	6.7	9.4	6.6	4.7%	29	11	5.7
La Crosse	13.5%	1.1	9.6	9.2	6.1	5.9%	34	14	5.6
Madison	36.9%	1.2	7.7	12.0	8.4	9.8%	43	20	5.6
Green Bay	11.3%	1.1	6.3	13.7	9.6	6.9%	31	13	5.5
Stevens Point	12.5%	1.1	9.6	11.3	10.3	6.4%	31	13	5.4
Oshkosh	27.0%	1.3	7.4	11.6	8.2	5.3%	28	11	5.4
Platteville	39.3%	1.3	11.1	11.4	7.6	5.7%	31	10	5.2
Superior	31.0%	1.2	8.1	19.7	7.6	8.2%	27	12	5.1
Whitewater	28.5%	1.2	8.3	16.4	12.9	7.2%	31	14	5.1
Eau Claire	45.1%	1.4	8.5	13.5	16.6	9.8%	33	14	4.8
Parkside	31.3%	1.1	10.0	29.5	10.9	9.1%	33	14	4.8
Stout	31.9%	1.3	9.8	15.9	15.1	9.5%	27	10	4.8
Milwaukee	81.6%	1.6	6.7	20.9	14.7	8.1%	36	13	4.6
Average	30.6%	1.2	8.4	15.0	10.3	7.4%	32	13	5.2

Although the results suggest higher performance of specific campuses over others, the dispersion of PPI scores is very tight, given that the difference between the highest and lowest index values is only 1.1 score points (around 9%). This outcome can be intuitive since campuses have limited control over key project variables such as delivery methods and contracting firms' selection. The small difference in performance is likely associated with each campus's ability to provide higher level of interaction during design, and preconstruction phases, as well as ensure more onsite supervision during construction, which consequently translates to a higher project performance.

6 Chapter 6: Conclusion

6.1 Research Summary & Findings

This study considered the performance of State construction projects that were executed between 2000 and 2017. Data was collected from 189 total projects with a total cost of $5.2 billion in 2018 adjusted dollars. This represents over 50% of the funding allocated by the state for public construction during the research period. The distribution of projects in this research is similar to the total population of projects and provides a representative sample.

Two principal goals of this research: investigating the overall performance of State projects and evaluating the relative performance of four delivery methods: Design Build (DB), Construction Manager at Risk (CMAR). Single-Prime (SP) and Multiple-Prime (MP).

The results showed that 97% of studied projects exceeded their initial contractual delivery (design and construction) cost by 10.3%, on average. Collaborative projects (DB and CMAR delivery) were found to be less likely to experience construction and delivery cost growth as compared to Traditional (SP and MP delivery).

Furthermore, 69% of projects overran their initial contractual time duration, on average by 39%, with projects using Collaborative methods less likely to experience schedule overrun than projects using Traditional methods.

The research studied eight performance metrics related to schedule, communication, change management, and spending to compare the four delivery methods under study. It was found that DB statistically outperformed SP and MP in each of the eight metrics, while CMAR outperformed SP in five and MP in six metrics. DB further outperformed CMAR in only one metric, and SP outperformed MP in three metrics.

On average, the analysis concluded that projects using Collaborative methods experienced less schedule overrun, when both including and excluding time extensions, less RFI processing time, fewer RFIs per million dollars, fewer change orders per million dollars, lower percentage change, higher construction and delivery speeds.

Using the eight performance metrics, the Project Performance Index (PPI) was developed to summarize the performance of a project and encapsulate it in a single numerical score on a range of 1 to 10 (with 10 resembling the highest performance). The results of this mathematical model showed that Collaborative delivery methods outperformed Traditional delivery methods by 40% on average. This indicates that Collaborative delivery is more likely to be capable of controlling schedule, maintaining better communication among the project parties, controlling project changes and delivering quicker execution.

As the PPI model is newly developed, it was validated by two independent statistical tests: Confirmatory Factor Analysis (CFA), and Partitioning Clustering Analysis. Confirmatory Factor Analysis produced weights that were almost identical to those produced by the PPI model, concluding that Collaborative projects outperformed Traditional ones by 45%, confirming the PPI model's findings. Partitioning Clustering Analysis was also used to create four clusters that represent a 4-tier scale of performance. This analysis found that Collaborative projects have a 75-95% chance of performing at the tier-1 level. Traditional methods have only a 2-5% chance of performing at this level.

The final portion of the analysis examined a potential factor that can be impactful to project success. It was found that the performance of UW-System campuses was not a significant factor, as only 1.1 points (~9%) separated the highest and lowest performing campuses. The conclusions of this research can be insightful to public owners as it highlights key performance indicators that can be relied on to maximize projects' success.

6.2 Discussion & Recommendations

The ultimate intention of this research is to present for the first time a quantitative data analysis of the performance of Wisconsin State building projects executed in public interest. The results should be considered by decision makers at the state level when evaluating the state's contracting practices.

As was stated in the previous section of this chapter, 97% of projects experienced cost overrun and 69% experienced schedule slippage. While these findings align with overall trends in construction, research has shown that using Collaborative delivery methods like DB and CMAR can result in better cost and schedule performance, on average, and particularly in the case of complex or technologically advanced projects Furthermore, the current bidding system accounts only for the lowest upfront cost, without capturing other performance areas during construction such as on-time delivery, customer satisfaction, latent defects and others.

As a result, this research recommends that the State of Wisconsin pilot other Collaborative delivery methods and compare their performance to Traditional ones, particularly in the case of time-sensitive or complex projects. These methods allow an earlier and higher level of involvement of project stakeholders and provide more effective cost and time performance. This research also quantitatively found that these methods outperform DBB Single and DBB-Multiple Prime contracting by 40-45% based upon the metrics used in the study.

Furthermore, this research recommends using the PPI model to track contractor performance and investigate utilizing a performance-based, rather than cost-based selection system. Such a system can permit and incentivize innovation. Contractors and project team members will likely seek to improve performance by all means available if performance is correspondingly rewarded via the aforementioned selection system. While innovative methods can experience slightly higher initial cost, and higher risks in some areas (such as execution speed), it can, however, eliminate unnecessary changes and provide an overall better performance.

It should be understood that the results presented in the preceding section are based on a set of assumptions and limitations that were discussed in Section 1.7. These include: contingency, complexity, unforeseen conditions, etc. Future research could perhaps gather data which would permit exploration of these factors.

Finally, it is recommended that an integrated data system be created that can serve as a single source of project data. Implementing such a system would allow project managers to contrast performance of their current project with historical projects and utilize multiple performance metrics to track the performance of their projects via dashboards. Employing such practices can enhance data-driven decision making and generate value.

Appendix A: Bibliography

Aitken, A. C. (1934). "Note on selection from a multivariate normal population." Proceedings of the Edinburgh Mathematical Society, 4, 106-110.

Agarwal, J., Chandrasekaran, S., and Sridhar, M. (2016b). “Imagining construction’s digital future.” McKinsey & Company.

Changali, S., Mohammad, A., and Nieuwland, M. V. (2015). “The construction productivity imperative.” McKinsey & Company.

B20 Australia. (2014). “B20 infrastructure & Investment taskforce policy summary.”

Barbosa, D., Woetzel, J., Mischke, J., Ribeirinho, M. J., Sridhar, M., Parsons, M., Bertram, N., and Brown, S. (2017). “Reinventing construction: a route to higher productivity.” McKinsey & Company.

Blanco, J. L., Janauskas, M., and Ribeirinho, M. J. (2016a). Beating the low productivity trap: how to transform construction operations.” McKinsey & Company.

Bryant, F. B., & Yarnold, P. R. (1995). “Principal-components analysis and confirmatory factor analysis.” In L. G. Grimm & P. R. Yarnold (Eds.), Reading and understanding multivariate statistics (pp. 99-136). Washington, DC: American Psychological Association.

Capital Budget. (2001-2017). State of Wisconsin Building Commission (SBC). Department of Administration.

Carpenter, N., and Bausman, C. D. (2016). “Project delivery method performance for public school construction: Design-bid-build versus CM at risk.” Journal of Construction Engineering & Management, 142(10).

Col Debella, D., and Ries, R. (2006). “Construction delivery systems: A comparative analysis of their performance within school districts.” Journal of Construction Engineering & Management, 132(11).

Construction Industry Institute (CII). “Performance metric formulas and definitions.”

Construction Management Institute of America (CMAA). (2012). “An owner's guide to project delivery methods.”

Design-Build Institute of America (DBIA). (2014). “What is Design-Build?”

Design-Build Institute of America (DBIA). (2018a). “DBIA State Maps.”

Dobbs, R., Pohl, H., Yi Lin, D., Mischke, J., Garemo, N., Hexter, J., Matzinger, S., Palter, R., and Nanavatty, R. (2013). “Infrastructure productivity: how to save $1 trillion a year.” McKinsey & Company.

El Asmar, M., and Hanna, A., and Loh, W. (2015). "Evaluating integrated project delivery using the project quarterback rating." Journal of Construction Engineering & Management.

Environmental and Energy Study Institute (EESI). (2014). "How Better Recycling Can Minimize Waste and Boost the Economy." Washington DC: Environmental and Energy Study Institute.

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). "Qgraph: Network visualizations of relationships in psychometric data." Journal of Statistical Software, 48(4), 1-18.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). "Evaluating the use of exploratory factor analysis in psychological research." Psychological Methods, 4(3), 272-299.

Flyvbjerg, B. (2017). "Introduction: The Iron Law of Megaproject Management." The Oxford Handbook of Megaproject Management, Oxford University Press, Chapter 1, pp. 1-18.

Franz, B., R. Leicht, K. Molenaar, and J. Messner. (2017). "Impact of team integration and group cohesion on project delivery performance." Journal of Construction Engineering & Management, 143(1).

Global Construction Perspectives and Oxford Economics. (2013). "Global construction 2025." London: Global Construction Perspectives Limited.

Griffith, A. F., Gibson, E., Hamilton, M. R., Tortora, A. L., & Wilson, C. T. (1999). "Project success index for capital facility construction projects." Journal of Performance of Constructed Facilities, 13(1), 39-45.

Hale, D., Shrestha, P., Gibson, G., and Migliaccio, G. (2009). "Empirical comparison of design/build and design/bid/build project delivery methods." Journal of Construction Engineering & Management, 135(7).

Hanna, A. S. (2001). "Quantifying the cumulative impact of change orders for electrical and mechanical contractors." Rep. No. 158-11, Construction Industry Institute (CII), University of Texas at Austin.

Hanna, A. S., Camlic, R., Peterson, P. A., and Lee, M. J. (2004). "Cumulative effect of project changes for electrical and mechanical construction." Journal of Construction Engineering & Management, 130(6), 762–771.

Hill, J., Foldesy, J., Ferrer, S., Freedman, M. Loh, A., Plaschke, F., (2015). "Solving the Construction Industry's Productivity Puzzle." The Boston Consulting Group (BCG).

Hooper, D., Coughlan, J., and Mullen, M. R. (2008). "Structural equation modeling: Guidelines for determining model fit." Electronic Journal of Business Research Methods, 6(1).

Horn, R. A., and Johnson, C. (2005). "Matrix Analysis." Cambridge University Press, Cambridge, United Kingdom.

Hughes, N., Wells, M., Nutter, C. L., and Zack, J. G. (2013). "Impact & control of RFIs on construction projects." Navigant Construction Forum.

Ibbs, W., Kwak, Y., Ng, T., and Odabasi, A. (2003). "Project delivery systems and project change: Quantitative analysis." Journal of Construction Engineering & Management, 129(4), 382–387.

Ibrahim, M., Hanna, A., and Kievet, D. (2018). "comparative analysis between project delivery systems through quantitative assessment of project performance." Proc., Construction Research Congress 2018, American Society of Civil Engineers, New Orleans, Los Angeles, 670–681.

IMS Health. (2015). "IMS health finds global cancer drug spending crossed $100 billion threshold in 2014."

Kats, G. (2003). "The costs and financial benefits of green buildings: a report to California's sustainable building task force." Boston: Massachusetts Technology Collaborative.

Kline, R. B. (2011). "Principles and practice of structural equation modeling." New York: Guilford Press.

Kaufman, L., and Rousseeuw, P. (1990). "Partitioning around medoids (Program PAM). In Finding groups in data: An introduction to cluster analysis." Hoboken, New Jersey: John Wiley Sons Inc., 68-125.

Konchar, M., and Sanvido, V. (1998). "Comparison of U.S. project delivery systems." Journal of Construction Engineering & Management, 124(6).

KPMG International. (2011). "Confronting Complexity: Research Findings and Insights."

Lam, E. W. M., Chan, A. P. C., and Chan, D. W. M. (2007). "Benchmarking the performance of design-build projects: Development of project success index." Benchmarking: An Int. J., 14(5), 624–638.

McCluskey, L. (2013). "Using construction manager at risk to save time and money." University Center Sioux Falls, South Dakota Board of Regents.

MacQueen, J. (1967). "Some Methods for Classification and Analysis of Multivariate Observations." Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281-296.

Osei-Kyei, R., and Chan, A. P. C. (2017). "Developing a project success index for public–private partnership projects in developing countries." Journal of Infrastructure Systems

Owen, A. (1988). "Empirical likelihood ratio confidence intervals for a single functional." Biometrika, 75, 237–249.

Owen, A. (1990). “Empirical likelihood ratio confidence regions.” The Annals of Statistics 18, 90–120.

Pishdad-Bozorgi, P., and Garza, J. M. (2012). “Comparative analysis of design-bid-build and design-build from the standpoint of claims.” Construction Research Congress, American Society of Civil Engineers, Reston, Virginia, 21–30.

Riley, D. R., Diller, B. E., and Kerr, D. (2005b). “Effects of delivery systems on change order size and frequency in mechanical construction.” Journal of Construction Engineering & Management, 131(9), 953–962.

Riley, D. R., Varadan, P., Thomas, H. R., and James, J. (2005a). “Benefit-cost metrics for MEP coordination in buildings.” Journal of Construction Engineering & Management 131(8).

Rojas, E. M., and Kell, I. (2008). “Comparative analysis of project delivery systems cost performance in Pacific Northwest public schools.” Journal of Construction Engineering & Management, 134(6).

Rojas, E. M. (2008). “Single versus Multiple-Prime contracting.” Journal of Construction Engineering & Management, 134(10).

Shrestha, P. P., and J. D. Fernane. (2016). “Performance of design-build and design-bid-build projects for public universities.” Journal of Construction Engineering & Management 143(3).

State Building Program. (2001-2017). Wisconsin Legislative Fiscal Bureau.

Sullivan, J., M. El Asmar, J. Chalhoub, and H. Obeid. (2017). “Two decades of performance comparisons for design–build, construction manager-at-risk, and design–bid–build: Quantitative analysis of the state of knowledge on project cost, schedule, and quality.” Journal of Construction Engineering and Management, 143(6).

Sustainable Building Task Force. Boston: Massachusetts Technology Collaborative.

Tabachnick, B. G., & Fidell, L. S. (2007). “Using multivariate statistics.” Boston: Pearson/Allyn & Bacon.

Teicholz, P. (2013). “Labor productivity declines in the construction industry: Causes and remedies (another look).” AECbytes.

Teicholz, P., Goodrum, P. M., and Haas, C. T. (2001). “US Construction Labor Productivity Trends, 1970–1998.” J. Constr. Eng. Mgmt., 127(5).

United Nations Environment Programme (UNEP). (2007). “Buildings and climate change: status, challenges and opportunities.” Paris: UNEP DTIE.

University of North Carolina (UNC). (2009). “Construction management at risk within the university of North Carolina.” University of North Carolina.

U.S. Census Bureau. (2017). "Value of construction put in place at a glance." U.S. Department of Commerce.

U.S. Bureau of Labor Statistics. (2017). "Current employment statistics highlights." Department of Labor.

U.S. Bureau of Labor Statistics. (2018). "Measuring productivity growth in construction." Department of Labor.

UW Madison Website. "Approval process." Capital Planning & Development. Division of Facilities Planning & Management.

Wenning, C. J. (2004-2014), "Student Laboratory Handbook." Illinois State University (ISU).

West L, Hankin R.K.S. (2008). "Exact tests for two-way contingency tables with structural zeros." Journal of Statistical Software, 28(11).

Wisconsin Act 20 (2013). "2013 Assembly Bill 40." Wisconsin State Legislature.

World Economic Forum (WEF). (2014). "Strategic Infrastructure: Steps to Operate and Maintain Infrastructure." Geneva.

World Economic Forum. (2015). "Strategic infrastructure: mitigation of political & regulatory risk in infrastructure projects." Geneva.

World Economic Forum (WEF) and The Boston Consulting Group. (2016). "Shaping the Future of Construction: A Breakthrough in Mindset and Technology." Geneva.

World Steel Association. (2015). "World Steel in Figures 2015." Brussels: World Steel Association.

Appendix B: City Multipliers

Year / Location	National Average	Milwaukee	Kenosha	Racine	Madison	Green Bay	Beloit	Wausau	La Crosse	Eau Claire	Superior	Oshkosh	Lancaster	Portage	New Richmond
2018	215.8	217	216	216	210	209	212	203	205	212	210	207	192	207	205
2017	209.4	215	214	214	207	207	210	201	203	210	207	205	190	205	203
2016	207.7	214	213	212	207	207	208	200	202	208	206	204	189	204	202
2015	204	209	205	205	202	201	201	193	195	201	199	197	183	197	195
2014	203	210	205	205	202	199	201	193	195	201	199	197	183	197	195
2013	196.9	205	201	201	198	194	196	188	190	196	194	192	178	192	190
2012	194	202	196	195	191	190	191	183	185	191	189	187	173	187	185
2011	185.7	192	188	187	183	181	183	176	178	183	181	180	166	180	178
2010	181.6	187	185	184	181	179	181	173	175	181	179	177	164	177	175
2009	182.5	187	183	183	180	176	179	172	174	179	177	175	163	175	174
2008	171	176	172	173	168	166	170	163	164	170	168	166	154	166	164
2007	165	169	165	165	161	159	161	155	156	161	160	158	147	158	156
2006	156.2	159	157	157	153	153	154	148	149	154	152	151	140	151	149
2005	146.7	148	148	148	145	144	145	139	141	145	143	142	132	142	141
2004	132.8	134	132	133	130	129	130	125	126	130	129	127	118	127	126
2003	129.7	131	131	132	128	127	129	124	125	129	128	126	117	126	125
2002	126.7	128	127	127	125	123	124	119	120	124	123	121	113	121	120
2001	122.2	124	124	123	121	119	121	116	117	121	120	118	110	118	117
2000	118.9	121	119	119	117	115	116	112	113	116	115	114	106	114	113
1999	116.6	117	116	116	116	112	113	109	110	113	112	111	103	111	110
1998	113.6	113	113	113	111	110	110	106	107	110	109	108	100	108	107