

# *Does Planning Quality Matter?* Results Frameworks, Adaptive Design, and Development Outcomes in World Bank Projects in South Asia

By

Sargam Kuckreja

[sa3230ku-s@student.lu.se](mailto:sa3230ku-s@student.lu.se)

**Abstract:** Does the quality of project planning documents predict development outcomes? This thesis examines that question empirically using 1,110 World Bank Investment Project Financing operations in South Asia approved between 1958 and 2021. Drawing on the full text of project appraisal documents, the thesis constructs quantitative scores operationalising three competing theoretical frameworks: Results-Based Management, Problem-Driven Iterative Adaptation, and the isomorphic mimicry hypothesis. These scores are then tested against the Independent Evaluation Group outcome ratings in a horse-race regression design across sixteen models. The results show that the results-based management infrastructure is the only framework that consistently and significantly predicts better outcomes, holding across specifications, sectors, and project types. Whereas, adaptive design language shows no predictive power in any model and context specificity is only significant when predicting outcomes in the Governance sector, where institutional diagnosis is most consequential. Tangibility only shapes baseline difficulty but does not moderate the design-quality relationship. The finding positions planning quality as a modest, robust and cross-sectoral contributor to development outcomes, and can be seen as necessary but not sufficient.

*Keywords: Aid Effectiveness, Project Design Quality, Results-Based Management, Problem-Driven Iterative Adaptation, Isomorphic Mimicry, World Bank, South Asia, Text-As-Data*

EKHS21

Master's Thesis (15 ECTS) in Economic Development

May 2026

Supervisor: Erik Green

Examiner: Tobias Axelsson

Word Count: 15,661

# Table of Content

|                                                                                     |           |
|-------------------------------------------------------------------------------------|-----------|
| <b>1. Introduction.....</b>                                                         | <b>1</b>  |
| 1.1. Aim and Scope.....                                                             | 3         |
| 1.2. Research Questions.....                                                        | 4         |
| 1.3. Contribution.....                                                              | 4         |
| 1.4. Outline of the Thesis.....                                                     | 5         |
| <b>2. Context.....</b>                                                              | <b>5</b>  |
| 2.1. The World Bank and Investment Project Financing.....                           | 5         |
| 2.2. South Asia as a Study Region.....                                              | 6         |
| 2.3. Evolution of Project Design Practices.....                                     | 7         |
| 2.4. Project Performance Patterns in South Asia.....                                | 8         |
| <b>3. Previous Research.....</b>                                                    | <b>9</b>  |
| 3.1. The Aid Effectiveness Debate.....                                              | 9         |
| 3.1.1. Macro-level explanations: institutions, governance, country context.....     | 9         |
| 3.1.2. The empirical record on aid and outcomes.....                                | 10        |
| 3.1.3. The shift toward project-level analysis.....                                 | 11        |
| 3.2. The Role of Project Design.....                                                | 11        |
| 3.2.1. Does design matter? The case for and against.....                            | 11        |
| 3.2.2. Evidence that context dominates over design.....                             | 12        |
| 3.2.3. Tangibility and task tractability: evidence from other datasets.....         | 12        |
| 3.2.4. Case study evidence on design and outcomes.....                              | 13        |
| 3.3. Planning Theory Beyond Aid.....                                                | 13        |
| 3.3.1. New Public Management and the rise of results-based management.....          | 13        |
| 3.3.2. The limits of NPM: compliance, mimicry, and performative accountability..... | 13        |
| 3.3.3. Context-conditioned performance: when does planning predict results?.....    | 14        |
| <b>4. Theoretical Framework.....</b>                                                | <b>15</b> |
| 4.1. Results-Based Management.....                                                  | 15        |
| 4.2. Problem-Driven Iterative Adaptation.....                                       | 15        |
| 4.3. Isomorphic Mimicry and the Specificity Hypothesis.....                         | 16        |
| 4.4. Tangibility as a Conditioning Factor.....                                      | 17        |
| 4.5. Hypotheses.....                                                                | 17        |
| <b>5. Data.....</b>                                                                 | <b>18</b> |
| 5.1. Primary Data Source.....                                                       | 18        |
| 5.2. Sample Construction.....                                                       | 19        |
| 5.3. Variable Definitions And Coding.....                                           | 20        |
| 5.4. Descriptive Statistics.....                                                    | 21        |
| <b>6. Methodology.....</b>                                                          | <b>27</b> |
| 6.1. Research Design.....                                                           | 27        |
| 6.2. Outcome Variable.....                                                          | 28        |
| 6.3. Text Scoring: Theory-Motivated Dimensions.....                                 | 29        |

|                                                                              |           |
|------------------------------------------------------------------------------|-----------|
| 6.3.1. RBM Score.....                                                        | 30        |
| 6.3.2. PDIA Score.....                                                       | 31        |
| 6.3.3. Specificity Score.....                                                | 31        |
| 6.3.4. Tangibility Variable.....                                             | 32        |
| 6.4. Regression Specifications.....                                          | 32        |
| 6.4.1. Isolation and Horse Race Models (A1-A3, B, C).....                    | 33        |
| 6.4.2. Interaction Models (D1-D3, E1-E3).....                                | 33        |
| 6.4.3. Tangibility Models (F, G1-G4).....                                    | 34        |
| 6.5. Identification and Limitations.....                                     | 34        |
| <b>7. Results and Findings.....</b>                                          | <b>36</b> |
| 7.1. Descriptive Overview.....                                               | 36        |
| 7.2. Isolation Models: Testing Each Theory Separately (Models A1-A3).....    | 36        |
| 7.3. The Horse Race: Competing Theories Simultaneously (Models B and C)..... | 38        |
| 7.4. Sector Interactions (Models D1-D3).....                                 | 40        |
| 7.5. Temporal Trends (Models E1-E3).....                                     | 41        |
| 7.6. Tangibility as a Moderating Factor (Models F and G1-G4).....            | 42        |
| 7.7. Summary of Findings Against Hypotheses.....                             | 45        |
| <b>8. Discussion.....</b>                                                    | <b>47</b> |
| 8.1. RBM Works - But Not for the Reasons the Reform Assumed.....             | 47        |
| 8.2. PDIA as a Management Practice, Not a Design Feature.....                | 47        |
| 8.3. Boilerplate as a Governance-Specific Problem.....                       | 48        |
| 8.4. The Tangibility Null - Rethinking What RBM Measures.....                | 49        |
| 8.5. Design as a Necessary but Insufficient Condition.....                   | 50        |
| <b>9. Conclusion.....</b>                                                    | <b>51</b> |
| 9.1. Future Research.....                                                    | 53        |
| <b>References.....</b>                                                       | <b>54</b> |
| <b>Appendix A.....</b>                                                       | <b>59</b> |
| <b>Appendix B.....</b>                                                       | <b>60</b> |
| <b>Appendix C.....</b>                                                       | <b>61</b> |
| <b>Appendix D.....</b>                                                       | <b>62</b> |
| <b>Appendix E.....</b>                                                       | <b>63</b> |
| <b>Appendix F.....</b>                                                       | <b>64</b> |
| <b>Appendix G.....</b>                                                       | <b>65</b> |
| <b>Appendix H.....</b>                                                       | <b>68</b> |

# List of Abbreviations

**ADB** - Asian Development Bank

**API** - Application Programming Interface

**CDN** - Content Delivery Network

**CI** - Confidence Interval

**DFID** - Department for International Development

**DPF** - Development Policy Financing

**FE** - Fixed Effects

**HC1** - Heteroskedasticity-Consistent Standard Errors (Type 1)

**IDA** - International Development Association

**IEG** - Independent Evaluation Group

**IPF** - Investment Project Financing

**M&E** - Monitoring and Evaluation

**NPM** - New Public Management

**NRM** - Natural Resource Management

**OECD DAC** - OECD Development Assistance Committee

**OLS** - Ordinary Least Squares

**PAD** - Project Appraisal Document

**PDO** - Project Development Objective

**PDIA** - Problem-Driven Iterative Adaptation

**PPAR** - Project Performance Assessment Report

**RBM** - Results-Based Management

**RCT** - Randomised Controlled Trial

**SAR** - Staff Appraisal Report

**USAID** - United States Agency for International Development

# List of Tables

|                                                                                       |    |
|---------------------------------------------------------------------------------------|----|
| Table 1. Overview Of Regression Models.....                                           | 32 |
| Table 2. Descriptive Statistics.....                                                  | 36 |
| Table 3. OLS Regressions: Each Theory Tested Separately (A1-A3).....                  | 37 |
| Table 4. Horse Race OLS (B) and Ordinal Logistic Regression (C).....                  | 39 |
| Table 5. Tangibility Models: Main Effect (F) and Interactions (G1-G4).....            | 44 |
| Table 6. Summary of Hypotheses and Findings.....                                      | 46 |
| Table 7. Scoring Dictionary For Dimension 1: Results-Based Management.....            | 60 |
| Table 8. Scoring Dictionary For Dimension 2: Problem-Driven Iterative Adaptation..... | 61 |
| Table 9. Scoring Dictionary For Dimension 3: Context Specificity.....                 | 62 |
| Table 10. Sector Classification by Tangibility.....                                   | 63 |
| Table 11. Theory $\times$ Sector Interactions, OLS (D1-D3).....                       | 65 |
| Table 12. Theory $\times$ Decade Interactions, OLS (E1-E3).....                       | 68 |

## List of Figures

|                                                                                         |    |
|-----------------------------------------------------------------------------------------|----|
| Figure 1. Analytical sample: number of IPF projects by country (N = 1,110).....         | 21 |
| Figure 2. Analytical sample: number of IPF projects by sector (N = 1,110).....          | 22 |
| Figure 3. Analytical sample: number of IPF projects by approval decade (N = 1,110)..... | 23 |
| Figure 4. Distribution of IEG outcome ratings, analytical sample (N = 1,110).....       | 24 |
| Figure 5. Mean IEG outcome score by country with 95% CI.....                            | 25 |
| Figure 6. Mean IEG outcome score by sector with 95% CI.....                             | 26 |
| Figure 7. Mean IEG outcome score by approval decade with 95% CI.....                    | 27 |
| Figure 8. Mean IEG outcome score by project tangibility with 95% CI.....                | 43 |
| Figure 9. Mean theory scores by project tangibility with 95% CI.....                    | 43 |

# 1. Introduction

For decades, the international development community has operated on a premise that better project planning leads to better outcomes (Kusek and Rist, 2004; OECD, 2019). This belief is so widely accepted that it has been institutionalized into the operational guidelines for every major donor (Carlsson, Köhlin and Ekbom, 1994; Cracknell, 2000). Donors invest heavily in project planning which requires implementing agencies to produce detailed appraisal documents which include objectives, indicators, baselines, risks and implementation details before the project begins (World Bank IEG, 2020; OECD DAC, 2010).

The World Bank alone has developed elaborate planning standards over six decades of lending, in the form of the Project Appraisal Document (World Bank IEG, 2020; Geli, Kraay and Nobakht, 2014). They are central planning instruments for Investment Project Financing operations which run to dozens of pages of structured analytical content (World Bank IEG, 2020). The underlying logic is that better-designed projects would be better implemented which in turn would yield better outcomes (Kusek and Rist, 2004; OECD DAC, 2010). South Asia, home to some of the World Bank's largest and longest-standing lending programmes, spanning infrastructure, governance, health, and education across diverse institutional environments, offers a particularly revealing setting in which to examine whether this logic holds (World Bank, 2022; World Bank IEG, 2023). However, the empirical record offers little reassurance that it does (Denizer, Kaufmann and Kraay, 2013; Kilby, 2015).

The World Bank's Independent Evaluation Group (IEG), the Bank's autonomous unit responsible for independently assessing the development effectiveness of all Bank-financed operations (World Bank IEG, 2023), rates roughly one in three Bank-financed projects in South Asia as Moderately Unsatisfactory or worse and this pattern has persisted across decades, countries and sectors despite waves of design reform (World Bank IEG, 2023). This persistent underperformance of the projects creates tension with the enormous effort devoted to project planning. If the planning documents were used as intended, better designed plans should lead to better outcomes. Yet whether this is actually the case has rarely been tested systematically before.

Debates on why aid produces such varying results include mostly macro-level and contextual factors such as recipient country institutions, governance quality, donor-recipient alignment,

and the broader political economy of reform (Burnside and Dollar, 2000; Easterly, 2006; Rajan and Subramanian, 2008). Adjacent research has focused on implementation failures, such as coordination breakdowns, capacity constraints, and political obstacles (Pressman and Wildavsky, 1984; Andrews, Pritchett and Woolcock, 2017). One dimension which has received systematically less attention is the role of project design itself, the document produced before implementation begins, as a possible predictor of subsequent outcomes.

This is deemed as a notable gap. Planning documents need to be seen as more than simple administrative artefacts. They constitute theories of change, specify causal pathways, allocate resources, and create institutional expectations about what the project is supposed to achieve and how that success will be measured (Weiss, 1997; Rogers, 2008). If the documents consist of real and relevant planning content such as genuine engagement with the problem, context, and implementation logic, they can be used to predict possible outcomes (Kusek and Rist, 2004). If they are largely performative and produced solely to satisfy donor requirements rather than assisting in implementation, the planning document can at best be considered costly and at worst actively misleading (Pritchett, Woolcock and Andrews, 2013; Natsios, 2010). This gap is also particularly consequential given the public resources involved. World Bank investment project operations in South Asia represent hundreds of billions of dollars of public and concessional finance. This adds a democratic accountability argument to the systematic evaluation equation: if public money is being spent on planning processes then there is a reasonable expectation that those processes should demonstrably improve outcomes (Savedoff, Levine and Birdsall, 2005; Birdsall and Savedoff, 2011).

The objective of this thesis is to examine that question empirically. Using a sample of 1,110 World Bank IPF operations in South Asia approved between 1958 and 2021, it constructs quantitative scores from the full text of project appraisal documents, operationalizes three competing theoretical frameworks, and tests their predictive power against the outcome ratings. The analysis is structured in a horse race theory format. Rather than assuming that just one theory of planning correctly describes how planning works, the thesis allows the data to adjudicate between them.

A further dimension in the analysis concerns whether the relationship between planning quality and outcomes varies by the nature of the project itself. Development projects differ intrinsically in what they try to achieve. Some involve more physically countable outputs

such as building roads, training healthcare workers or expanding irrigation networks, while others target objectives like institutional change, governance reform or behavioral change, where the causal change from design to outcome is longer and harder to measure (Andrews, Pritchett and Woolcock, 2017; Pritchett and Woolcock, 2004). In this thesis, this distinction is referred to as tangibility and raises the question of whether results-based planning infrastructure is more informative in some sector contexts than others, and whether the right kind of planning discipline varies systematically with the nature of the problem a project is trying to solve.

### ***1.1. Aim and Scope***

This thesis examines whether the quality of the World Bank project planning documents, measured at the time of appraisal, can predict the development outcome that those projects eventually achieved. The thesis employs the patterns of results to test three competing theories of what makes the aid project planning effective. Three competing theoretical frameworks are operationalised: Results-Based Management, which predicts that quantified targets, baselines, and monitoring systems improve outcomes; Problem-Driven Iterative Adaptation, which predicts that flexible and adaptive design features matter more than upfront planning precision; and the isomorphic mimicry hypothesis, which predicts that generic donor boilerplate signals performative rather than genuine planning and should predict worse outcomes.

The scope of the thesis is defined by four boundaries. First, the analysis uses only World Bank Investment Project Financing operations, and excludes the Development Policy Loans and trust fund operations, due to their different design logics. Second, the geographic scope is confined to eight South Asian countries, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, Afghanistan, and Sri Lanka. This region is one of the Bank's most active lending regions and offers a diverse variety of governance environments, income levels and sector compositions. Third, the study period runs from 1958-2021, capturing both pre-Project Appraisal Documents era of Staff Appraisal Reports and the post-1995 era of Project Appraisal Documents, allowing comparison across two structurally different document regimes. Fourth, the unit of analysis in the thesis is the individual project, with outcomes measured by the evaluation unit's six-point rating scale applied at project completion.

Lastly, the thesis does not aim to establish causal identification in the sense of a randomised experiment or instrumental variable design. The goal is to identify the strongest feasible observational test of a question that has direct practical importance: whether features of planning documents which donors view as indicators of quality actually predict the outcomes those projects are intended to achieve.

## **1.2. Research Questions**

The analysis is structured around four research questions.

*RQ1: Does the quality of project design at appraisal predict development outcomes in World Bank investment projects in South Asia?*

*RQ2: Which features of project planning documents, if any, are associated with better development outcomes, and what does this reveal about competing theories of aid effectiveness?*

*RQ3: Does the relationship between planning quality and outcomes vary across sectors, and does project tangibility, whether a project involves physical infrastructure or institutional change, determine how much design quality matters?*

*RQ4: Has the relationship between planning quality and outcomes changed over time as World Bank design practices have evolved?*

## **1.3. Contribution**

The thesis contributes to aid effectiveness literature in two ways. Fundamentally, it shifts the focus from macro-level context and implementation processes to ex ante project design, treating planning documents as empirical material rather than administrative background. This is a perspective that evaluation agencies have long understood as important and weak design is routinely cited as a cause of underperformance in project completion reports, yet has rarely been tested systematically across large samples (OECD, 2019; World Bank IEG, 2020).

Methodologically, the contribution lies in the text-as-data approach. Instead of relying solely on manual quality ratings, which can be expensive, slow and subject to rater inconsistencies (Krippendorff, 2018; Lombard, Snyder-Duch and Bracken, 2022), this thesis constructs automated scores from the full text of appraisal documents using theory-motivated keyword dictionaries. This strategy is transparent, replicable, and scalable across the full historical portfolio, consistent with the broader turn toward computational text analysis in political science and economics (Grimmer and Stewart, 2013; Gentzkow, Kelly and Taddy, 2019). Additionally, the paper makes explicit the theoretical assumptions that are embedded in the scoring process, that is, what words and phrases are treated as signals of which framework and why. This allows those assumptions to be evaluated and contested, following the principle that quantitative text analysis should be theory-driven rather than purely inductive (Grimmer, Roberts and Stewart, 2022). The result of this approach is a horse race between three frameworks that each have substantial scholarly support but have rarely been tested against each other with the same data and the same outcome measure (Burnside and Dollar, 2000; Andrews, Pritchett and Woolcock, 2017; Pritchett, Woolcock and Andrews, 2013).

#### ***1.4. Outline of the Thesis***

This thesis is divided into 9 chapters. The following section provides the contextual background necessary to interpret the empirical analysis, covering the institutional context of World Bank lending in South Asia, the evolution of project design practices from the earlier planning documents through to the present, and the development context of the region. Chapter 3 reviews the relevant previous research. Chapter 4 develops the theoretical framework, setting out the three perspectives and deriving the testable hypotheses. Chapter 5 describes the data and document retrieval process. Chapter 6 sets out the measurement and regression methodology. Chapter 7 presents the empirical findings. Chapter 8 discusses their implications for theory, policy, and practice and finally, Chapter 9 summarises the findings for each research question and situates them within the broader debate on whether directed public investment can systematically improve development outcomes.

## **2. Context**

### ***2.1. The World Bank and Investment Project Financing***

The World Bank Group has played a central role in financing the development initiatives in South Asia since their earliest lending operations in the 1950s. Their primary instrument for

project-based lending are the Investment Project Financing (IPF), which supports discrete, time-bound interventions with defined outputs, structured implementation arrangements, and formal results frameworks. Unlike some other World Bank lending instruments that fund government reforms, IPF projects fund specific activities such as, building roads, training health workers, reforming land registries, and expanding irrigation networks. Each IPF project is designed around a stated Project Development Objective which underlines the outcomes that the project is intended to achieve and results frameworks that specifies how progress toward the objectives will be measured.

At project closure, the evaluation unit independently rates project outcomes on a six-point scale ranging from Highly Unsatisfactory to Highly Satisfactory, based on how well the projects achieved the stated development objectives. These ratings are the outcome variable used in this thesis. Although imperfect since they only reflect one organization's judgment about whether a stated objective was met or not and not a counterfactual estimate of development impact, they are the most systematic comparable outcome variable across the full historical IPF portfolio and they have been widely used in the aid effectiveness literature (Denizer, Kaufmann and Kraay, 2013; Honig, 2018).

## ***2.2. South Asia as a Study Region***

South Asia provides an analytically powerful setting for testing whether planning quality predicts project outcomes. The eight countries covered in this thesis collectively represent one of the World Bank's most active and longest-standing lending regions. Over the period covered in this thesis, the World Bank approved more than 1,300 lending operations across these eight countries, of which 1,110 have both a retrievable appraisal document and an outcome rating. India alone accounts for a substantial share of the global portfolio, both in terms of number of projects and money spent. South Asia is a particularly well-suited setting for this analysis for three reasons. First, the document retrieval rate is exceptionally high, with planning documents recoverable for over 84 per cent of the rated portfolio, which is critical for a text-based study of this kind. Second, the portfolio spans a long and continuous lending history from the 1950s through to 2021, allowing temporal comparison across both the pre-1995 and post-1995 appraisal document eras within a single coherent regional sample. Third, the combination of scale and institutional heterogeneity generates precisely

the variation needed to test competing theories of planning quality across meaningfully different implementation environments (World Bank, 2022; World Bank IEG, 2023).

This heterogeneity is both a methodological challenge and an analytical asset. Countries range from low-income fragile states such as Afghanistan and Nepal to lower-middle-income economies with substantial domestic institutional capacity such as India and Sri Lanka. Governance quality varies from relatively functional federal democratic systems to post-conflict environments with limited state presence beyond major urban centres. Sector composition differs across the region too, with some countries dominated by infrastructure, energy, transport, and water, and others rich in social sectors and governance reform. This heterogeneity demands careful control for country and sector effects, but it also generates the variation in conditions under which theories of planning quality can be meaningfully tested.

The development trajectory of the region over the study period adds a further layer of variation. India saw sustained growth following the 1990s liberalisation, Bangladesh pursued an export-led development path, and Sri Lanka made considerable improvements in human development indicators. Yet alongside these successes, persistent structural challenges remained. Poverty was heavily concentrated in specific geographical areas and demographic groups, human development indicators in Afghanistan, Nepal, and rural Bangladesh lagged significantly behind income levels, and governance constraints continued to affect public service delivery quality across many contexts (World Bank, 2022; UNDP, 2016). Taken together, these contrasts, between countries and within them over time, create the variation in implementation environments that makes it necessary to test whether the planning-outcomes relationship is conditional on context.

### ***2.3. Evolution of Project Design Practices***

World Bank project design practices have changed over the six decades that are covered in this study. Prior to 1995, projects were hegemonized by Staff Appraisal Reports (SAR) which were documents that followed a less standardized format and places less emphasis on formal results measurement. SAR-era projects typically have more specified objectives in qualitative terms, identified outputs rather than outcome-level indicators, and bare minimum of the monitoring and evaluation language, which is now standard. Results were assessed at

completion using criteria that entailed more physical completion and disbursement rates alongside development impact.

The introduction of the Project Appraisal Document (PAD) format in 1995, was a structural shift. The PAD format imposed a more standardised structure on appraisal documents, they have structured sections for development objectives, results frameworks, risk matrices, and implementation agreements. In the early 2000s, there was a shift towards the mandatory inclusion of the Results Framework, which included baseline data, target values, and specified means of verification. By the mid-2000s, PAD-era were significantly longer, more structured and more explicitly results-oriented than their SAR-era predecessors (Kusek and Rist, 2004; OECD DAC, 2010; World Bank IEG, 2020).

The consequences of this standardisation for actual project design quality, as opposed to document format, are contested, and resolving that contestation is one of the central ambitions of this thesis. Those that are for the reform argue that the Results Framework has genuinely improved planning by forcing teams to specify causal pathways, identify measurable outcomes and build monitoring systems into the project design from the outset (Kusek and Rist, 2004; Mayne, 2007; World Bank IEG, 2012). Those against it argue that the standardisation created new forms of gaming. Teams produced results frameworks because they had to and populated them with output-level proxies that were easier to achieve rather than genuine outcome indicators, and with baseline values that were aspirational rather than empirically grounded (Honig, 2018; Natsios, 2010; Pritchett, Woolcock and Andrews, 2013). A binary document-era variable is included in all the regression models, SAR-era and PAD-era, aimed at absorbing this structural shift and preventing it from confounding the theory-score coefficients.

## ***2.4. Project Performance Patterns in South Asia***

World Bank project performance in South Asia, as measured by the independent evaluators, has been historically below the global institutional average, though this gap has narrowed in recent cohorts (OECD, 2019; World Bank IEG, 2023). This trend can be attributed to a combination of factors: concentration of lending in sectors and countries where development is inherently more difficult to achieve, complex institutional environments (Pressman and

Wildavsky, 1984; Andrews, Pritchett and Woolcock, 2017), and lastly, as this thesis investigates, systematic differences in project design quality.

Sector-level variation in performance is substantial. Projects in infrastructure sectors such as transport, energy, water and sanitation tend to achieve higher outcome ratings than projects in sectors like governance, public sector reform and social policy (World Bank IEG, 2020; World Bank IEG, 2008). This pattern has been consistent across countries and time periods (World Bank IEG, 2023). Education and health projects tend to have both physical delivery components and behavioural change objectives, which places them in an intermediate position (Pritchett and Woolcock, 2004).

Country-level variation in performance adds a further layer of complexity. Outcome ratings differ meaningfully across the eight countries in the sample, reflecting differences in institutional capacity, political stability, and the composition of each country's lending portfolio (Denizer, Kaufmann and Kraay, 2013). Countries with stronger administrative systems and more stable implementation environments tend to perform better on average, while those facing fragility, conflict, or severe governance constraints show systematically lower ratings (World Bank IEG, 2023). Afghanistan represents the most pronounced case of the latter, given the exceptional institutional and security challenges of its post-2001 lending programme. These country-level patterns are not incidental to the analysis, they are a structural feature of the data that must be controlled for, which is why country fixed effects are included in all regression models. A detailed breakdown of outcome scores by country is presented in Chapter 5.

## **3. Previous Research**

### ***3.1. The Aid Effectiveness Debate***

#### *3.1.1. Macro-level explanations: institutions, governance, country context*

The dominant discourse in aid effectiveness research has sought to explain variation in aid outcomes using the country level argument. For decades, the central question has been whether aid, in aggregated form, promotes economic growth or not. The answer, as literature shows, has proved to be stubbornly ambiguous. Burnside and Dollar's (2000) influential research finds that aid promotes growth only in countries with sound macroeconomic policy.

This generated a wave of follow-on studies. Most of which did not manage to replicate the results across broader samples or longer time horizons (Easterly, Levine and Roodman, 2004; Rajan and Subramanian, 2008). The overarching lesson from this literature is not that aid does not work, instead that country-level institutional quality, governance capacity, and the political economy of reform are powerful confounders. When institutions are weak, even well-resourced interventions struggle to produce durable results and when they are strong, aid might be redundant.

Although this thesis operates at a different level of analysis, this macro level tradition has vital implications. If country-level context explains most of the variation in development outcomes, then project-level planning quality may have limited predictive power regardless of its theoretical merits. Country effects are therefore included in all the regression models to control for those nation-level differences, which allows the analysis to highlight how planning quality affects outcomes within counties rather than between them.

### *3.1.2. The empirical record on aid and outcomes*

A parallel strand of research looks beyond growth regressions to investigate project-level performance data directly. Studies that used World Bank IEG ratings as dependent variable have consistently found that sector, country, income level, and governance quality are amongst the strongest predictors of project outcomes, while project-specific design features have received less attention (Denizer, Kaufmann and Kraay, 2013; Chauvet and Collier, 2008; Kilby, 2015). (Denizer, Kaufmann and Kraay (2013) is a study that reinforces the primacy of implementation capacity over formal design by focusing on task manager characteristics. They are proxies for human capital and institutional quality within the Bank and it was found that they predict outcomes more strongly than observable project characteristics. Isham and Kaufmann (1999), in their study, find that civil liberties and participatory governance in recipient countries predict better project outcomes across the World Bank portfolio, once again pointing to context over plan.

What is interesting about this body of literature is that it does not examine the content of the planning document itself. IEG ratings are linked to project characteristics such as sector, size and country but does not mention what the appraisal document actually said at the moment of approval. This is the gap that is directly addressed in this thesis.

### *3.1.3. The shift toward project-level analysis*

A relatively newer strand of research has begun moving the trend downwards from country aggregates to the project level. Briggs (2017) demonstrated that World Bank projects locations within the country had predictive power in determining local development outcomes. This suggests that project-level decisions about targeting matter independently of the country context. Honig (2018) investigated the degree of managerial autonomy granted to field staff, rather than rule-following from headquarters. The study showed better outcomes in complex projects, a finding directly connected to the adaptive design framework (PDIA) examined in this thesis. The project level studies collectively show that the variation in outcomes achieved after controlling for a country is not purely random but is rather structured by decisions made at the project design and management level. This inference motivates the core research design of this thesis.

## **3.2. The Role of Project Design**

### *3.2.1. Does design matter? The case for and against*

Even though the claim that better-designed projects achieve better outcomes is the premise of modern development projects management, it has been subjected to surprisingly little systematic empirical scrutiny. The case for this claim is based on a straightforward theory of change: projects that clearly specify objectives, identify causal pathways, allocate resources rationally, and establish monitoring systems give implementing agencies a coherent roadmap which reduces coordination failures and allows for an early detection of implementation problems (Kusek and Rist, 2004; OECD DAC, 2010). This logic is compelling in stable and technically tractable environments, where the assumptions embedded in the design are more likely to hold.

The case against is also equally coherent. Easterly (2006) argues that top-down planning in development consistently fails because it does not take into account local knowledge which actually determines whether the project will work in practice. For example, the conditions under which an irrigation scheme functions, a health system reforms, or a governance programme produces compliance, are all deeply content-dependent in ways that are not anticipated in the pre-implementations documentation. This argument deems design quality as irrelevant since outcomes are determined by implementation conditions and/or actively

harmful due to elaborate planning processes creating false confidence in the predictability of complex social change.

### *3.2.2. Evidence that context dominates over design*

Empirical literature that directly examines whether project design independently affects development outcomes remains relatively limited. The few studies that do exist generally find that contextual factors such as country conditions and sector characteristics explain a larger share of project outcome variation. Denizer, Kaufmann and Kraay (2013) analyzed a large sample of World Bank projects and their conclusion shows that country and sector fixed effects account for substantially more outcome variation than the project-level covariates. These results hint to the fact that the project's fate is largely determined before the first disbursement and by the governance environment it is being deployed in. Similarly, qualitative case evidence documents how formal planning compliance diverges systematically from the actual implementation. Andrews, Pritchett and Woolcock (2017) show that agencies produce technically sophisticated appraisal documents to satisfy donor needs while they continue to operate in pre-existing routines and political incentives. In cases like these, design documents become performative artefacts rather than genuine guides to action. Whether this characterization holds for the full World Bank South Asia portfolio, is precisely what the tangibility conditioning analysis in this thesis is designed to answer.

### *3.2.3. Tangibility and task tractability: evidence from other datasets*

There is evidence from outside the World Bank dataset that suggests that depending on how straightforward or measurable a task is affects how much planning quality matters for outcomes. Research on public sector reform in OECD contexts has found that management practices predict performance more strongly in organisations delivering standardized, measurable services than in those that are engaged in complex regulatory or behavioral change functions (Andrews, 2013). Pritchett and Sandefur (2015) in their research found that the predictive validity of randomised controlled trial evidence is higher for technically simple interventions with short causal chains than for complex, multi-actor reform programs. The findings align with the theoretical prediction advanced in this thesis which states that results-based planning infrastructure is most informative where outputs are physically countable and the causal chain from input to outcome is short.

#### *3.2.4. Case study evidence on design and outcomes*

Qualitative evidence on the relationship between specific design features and project outcomes is mixed but suggestive. Valters and Whitty (2017) shows how the introduction of results frameworks in the UK development programming improved monitoring discipline in some instances but also generated perverse incentives since the program managers started to focus more on activities with measurable outcomes rather than harder-to-observe development goals. Ramalingam (2013) reviews a range of case evidence which highlights how the projects that operate in complex adaptive systems consistently outperform the blueprint approaches when they incorporate learning and feedback loops. The drawback of this literature is that case selection is rarely random and then evidence is vulnerable to confirmation bias. Systematic quantitative tests of the design-outcome relationship that use large samples and independent outcome measures, remain scarce, which is the core contribution this thesis attempts to make.

### ***3.3. Planning Theory Beyond Aid***

#### *3.3.1. New Public Management and the rise of results-based management*

The intellectual origins of results-based management lie not in development economics but in public administration theory. New Public Management (NPM) got prominence in OECD governments during the 1980s and 1990s and argues that public sector performance can be improved by improving private sector management skills such as explicit performance targets, output measurement, and accountability mechanisms linking rewards to results (Hood, 1991; Pollitt, 1993). This way of thinking was rapidly absorbed by multilateral development institutions. The World Bank's formal adoption of Results Frameworks in 2000 and USAID's subsequent shift to performance-based contracting were direct institutional expressions of NPM theory applied to development project management. The results framework approach score devised for this thesis intends on capturing how deeply the NPM principles are embedded in individual project appraisal documents.

#### *3.3.2. The limits of NPM: compliance, mimicry, and performative accountability*

NPM, too, comes with its own set of challenges. Public administration research documents dysfunctions that emerge when results measurements are mandated from above. Such as,

gaming of indicators, displacement of effort towards measurable outputs at the expense of mission-critical but harder-to-quantify outcomes, and the multiplication of formal compliance procedures that absorb administrative resources without improving performance (Power, 1997; Bevan and Hood, 2006).

In a development context, these drawbacks take a specific form described by Pritchett, Woolcock and Andrews (2013) as isomorphic mimicry. They formulate it as organisations adopting external forms of effective institutions such as results frameworks, M&E (Monitoring and Evaluation) systems, logical frameworks, without internalising the underlying management logic. The consequence is that planning documents become sophisticated performances of donor-required compliance than genuine instruments for project management. This theoretical possibility is what the specificity score in this thesis is designed to measure. This will help in deciphering whether the generic boilerplate density negatively predicted outcomes.

### *3.3.3. Context-conditioned performance: when does planning predict results?*

Public administration literature provides a meaningful theoretical bridge between the monitoring-oriented design (RBM) and adaptive design (PDIA) perspectives. Mintzberg (1994) critiques formal planning by arguing that it adds value in stable and predictable environments, but becomes counterproductive when there is turbulence and plans decay rapidly. Snowden and Boone's (2007) Cynefin framework makes a related distinction between complicated problems, where expertise and structured planning yield reliable solutions, and complex ones, where outcomes are emergent and adaptive management is required.

When applied to development projects, this framework predicts a heterogeneous relationship between planning quality and outcomes. Rigorous upfront design should predict success in tangible, technically tractable projects such as infrastructure and service delivery, but may be irrelevant or even harmful in institutionally complex ones such as governance reform and behaviour change. This theoretical prediction is formalised in the tangibility conditioning hypothesis tested in Models G1 through G4.

## 4. Theoretical Framework

This chapter describes the three theoretical frameworks that structure the empirical analysis. Each framework generates a distinct prediction about what makes aid project planning effective and each is operationalised as a quantitative score derived from the text in the project planning documents. The analysis then puts the three frameworks in direct competition with each other which allows the data to adjudicate between them rather than assuming any single theory is correct.

### ***4.1. Results-Based Management***

Results-Based Management (RBM) puts forth the notion that strong planning infrastructures at appraisal, i.e. quantified targets, baseline measurements, time-bound measurements, and formal monitoring and evaluation systems, achieve better development outcomes because these features aid in making the implementation problems more visible and in time be corrected. The underlying logic is informational which explains that without baseline data, progress cannot be measured, without milestones you can measure schedule slippage, and without specified verification sources, indicator data cannot be collected. Planning compliant with RBM therefore gives managers the tools that are necessary to track progress and intervene before problems compound (Kusek and Rist, 2004; OECD DAC, 2010).

The prediction based on this theory hypothesis that projects with higher RBM score at appraisal should achieve higher outcome ratings at completion. Given that the World Bank formally adopted the Results Frameworks from 2000 onwards, RBM scores are expected to be higher in the PAD era than the SAR era, and the document era control is included in all models to account for this structural shift.

### ***4.2. Problem-Driven Iterative Adaptation***

Problem-Driven Iterative Adaptation (PDIA), developed by Andrews, Pritchett and Woolcock (2017) and extended empirically by Honig (2018), is built on the observation that many development challenges, governance reform, institutional capacity building, behaviour change, are complex problems for which no reliable technical blueprint exists at the time of design. In these conditions, rigid upfront planning is not merely helpful but instead counterproductive. They instead lock in assumptions that will prove wrong and create accountability pressures that discourage honest reporting of problems.

This theory predicts that projects designed with explicit uncertainty acknowledgment, phased implementation, contingency provisions, and formal learning mechanisms will perform better in complex environments because they are capable of adapting when initial assumptions fail. From a critical perspective, this prediction is conditional on project type. Adaptive design language should matter most for institutionally complex, intangible projects where problems are emergent and solutions cannot be specified upfront. For straightforward infrastructure projects, where the path from design to delivery is more predictable, the theory's predictive value is likely to be limited. This conditioning prediction is tested directly in the tangibility interaction models.

#### ***4.3. Isomorphic Mimicry and the Specificity Hypothesis***

The isomorphic mimicry hypothesis, drawn from Pritchett, Woolcock and Andrews (2013) and rooted in DiMaggio and Powell's (1983) institutional theory, predicts that organisations under external compliance pressure will adopt the visible forms of effective institutions without the underlying substance. DiMaggio and Powell (1983) identify three mechanisms through which this occurs: coercive isomorphism, where organisations conform due to external pressure from dominant actors such as donors; mimetic isomorphism, where organisations copy the forms of peers perceived as successful; and normative isomorphism, driven by professionalisation and shared training across organisations. In the World Bank context, it is primarily coercive and mimetic isomorphism that are relevant, implementing agencies adopt results frameworks and M&E plans because donors require them (coercive) and because these forms have become the recognised signal of a well-designed project across the multilateral system (mimetic).

The most observable sign of mimicry is the use of generic language. This includes documents that rely heavily on standard donor vocabularies, which could be applied to any project in any context, rather than language which reflects genuine engagement with specific problems and setting in hand. The specificity hypothesis therefore estimates that higher generic boilerplate density should be associated with worse outcomes, while greater use of context-specific language, named institutions, geographic references, identified beneficiary groups, should predict better outcomes.

#### ***4.4. Tangibility as a Conditioning Factor***

Tangibility is also a factor deemed important in this thesis as it moderates the relationship between planning quality and outcomes. Drawing on Andrews, Pritchett and Woolcock's (2017) distinction between technically tractable problems with short causal chains and institutionally complex problems with long, contested ones, the thesis classifies projects as tangible (physical infrastructure, direct service delivery) or intangible (governance reform, policy change, institutional capacity building).

The claim is that tangible projects, with results framework planning should be the most informative since the targets are physically meaningful, baselines can be established, and monitoring systems can detect problems early. Whereas, for the intangible ones, adaptive design should add the most value, since the outcomes depend on political will and social dynamics that lie outside the project manager's control.

#### ***4.5. Hypotheses***

The frameworks above result in six testable hypotheses:

- H1: RBM predicts that stronger results infrastructure at appraisal leads to better outcomes.
- H2: PDIA predicts that adaptive design language improves outcomes in complex environments.
- H3: The specificity hypothesis predicts that generic boilerplate is negatively associated with outcomes.
- H4: RBM effects are expected to be stronger for tangible projects than intangible ones.
- H5: PDIA effects are expected to be stronger for intangible projects than tangible ones.
- H6: RBM's predictive power is expected to have grown after the formal adoption of Results Frameworks in 2000, and the specificity penalty to have deepened as boilerplate spread through the post-1995 appraisal document portfolio.

All six hypotheses are tested across sixteen regression models described in Chapter 6.

## 5. Data

This study draws on a dataset of 1,110 World Bank Investment Project Financing (IPF) projects in South Asia, spanning eight countries and 15 sectors with approval years from 1958 to 2021. Each project is linked to an IEG outcome rating, the primary dependent variable, and to a project planning document from which three theory-based text scores are derived. A tangibility classification is additionally assigned to each project based on its sector. The following sections describe the data sources underlying this dataset (Section 5.1), the filtering decisions that produced the final sample (Section 5.2), the coding of all variables (Section 5.3), and the distributional features of the sample (Section 5.4).

### 5.1. *Primary Data Source*

This study employs two linked data sources. Firstly, the outcome ratings of projects are drawn from the World Bank IEG Project Performance Ratings Database, available through the World Bank Data Catalog (World Bank, 2026a), accessed in March 2026. The IEG is the World Bank's autonomous evaluation unit, required to assess the development effectiveness of all the Bank-financed operations, independently of Bank management. For each completed project, IEG assigns an outcome rating on a six-point ordinal scale: Highly Unsatisfactory (1), Unsatisfactory (2), Moderately Unsatisfactory (3), Moderately Satisfactory (4), Satisfactory (5), and Highly Satisfactory (6). These ratings are derived through systematic desk reviews, and for larger operations, through direct field based evaluations (Independent Evaluation Group, 2023).

A central methodological advantage in this study is the independent nature of the IEG ratings from the project teams that were involved in the documents that are being studied. Self-reported completion reports produced by the teams that were involved in the projects are subjected to attribution bias and strategic framing (Michaelowa and Borrmann, 2006). IEG ratings, on the other hand, are assigned by staff who have had no operational responsibility for the project under review. This independence allows for a fair testing of whether the planning document characteristics carry information about the outcomes. The use of IEG ratings as the dependent variable follows an established practice in development effectiveness literature (Kilby, 2015; Denizer, Kaufmann and Kraay, 2013; Chauvet and Collier, 2008).

Secondly, the planning documents, Project Appraisal Documents (PADs) and Staff Appraisal Reports (SARs), were retrieved from the World Bank Documents and Reports portal (World Bank, 2026b). SARs were the standard format before 1995 and PADs have been successor to the post-1995 collection. Both of these documents contain project objectives, implementation design, and monitoring arrangement. These are publicly available for a majority of the portfolio. Document URLs were retrieved pragmatically using the World Bank Documents API and PDFs were downloaded using browser-compatible HTTP requests to handle Content Delivery Network (CDN) redirections in the document delivery system.

## **5.2. *Sample Construction***

**Instrument Filter.** The sample consists of the IPF instruments. IPF was chosen because they are the only WB financed development projects where the objectives are clearly defined, components and beneficiary groups are mentioned, and the only instrument type where the PADs and SARs are produced in a comparable manner. Contrarily, the Development Policy Financing (DPF) operations, accounting for 157 observations in the initial South Asia database, are budget-support instruments tied to policy reform conditionality. They have a fundamentally different planning logic and are evaluated against a different criteria. Including the DPF sample would conflate two distinct organizational processes and render the text-based theory scores theoretically incoherent for that sub-sample.

After filtering for IPF, the eligible population comprised 1,334 projects. Planning documents for the same projects were successfully retrieved for 1,127 projects (84.5 per cent), using a multi-stage download process that handled Bank's CDN redirect architecture and automatically retried failed downloads with updated URL formats. After linking the project id's from the IEG ratings database to the available planning documents, and removing 17 projects rated as 'Not Rated', the final analytical sample came down to 1,110 projects spanning eight countries and 15 sectors with approval years from 1958 to 2021.

The 207 projects for which no planning documents could be located were not randomly distributed. Older projects, approved before the mid-1980s are overrepresented, reflecting an incomplete digital coverage of the World Bank's earliest operational period. There were 10 documents that were classified under Official Use Only and were not publicly accessible.

### 5.3. *Variable Definitions And Coding*

**Dependent Variable.** The IEG outcome rating is coded as an integer from 1 to 6. In the OLS specifications this is treated as an approximately continuous variable, following standard practice in development effectiveness literature (Denizer, Kaufmann and Kraay, 2013; Isham and Kaufmann, 1999). The preferred specification uses ordinal logistic regression, which treats the variable as ordered categorical and does not impose equal interval spacing between categories. The mean outcome score in the analytical sample is 4.11 (SD = 1.20).

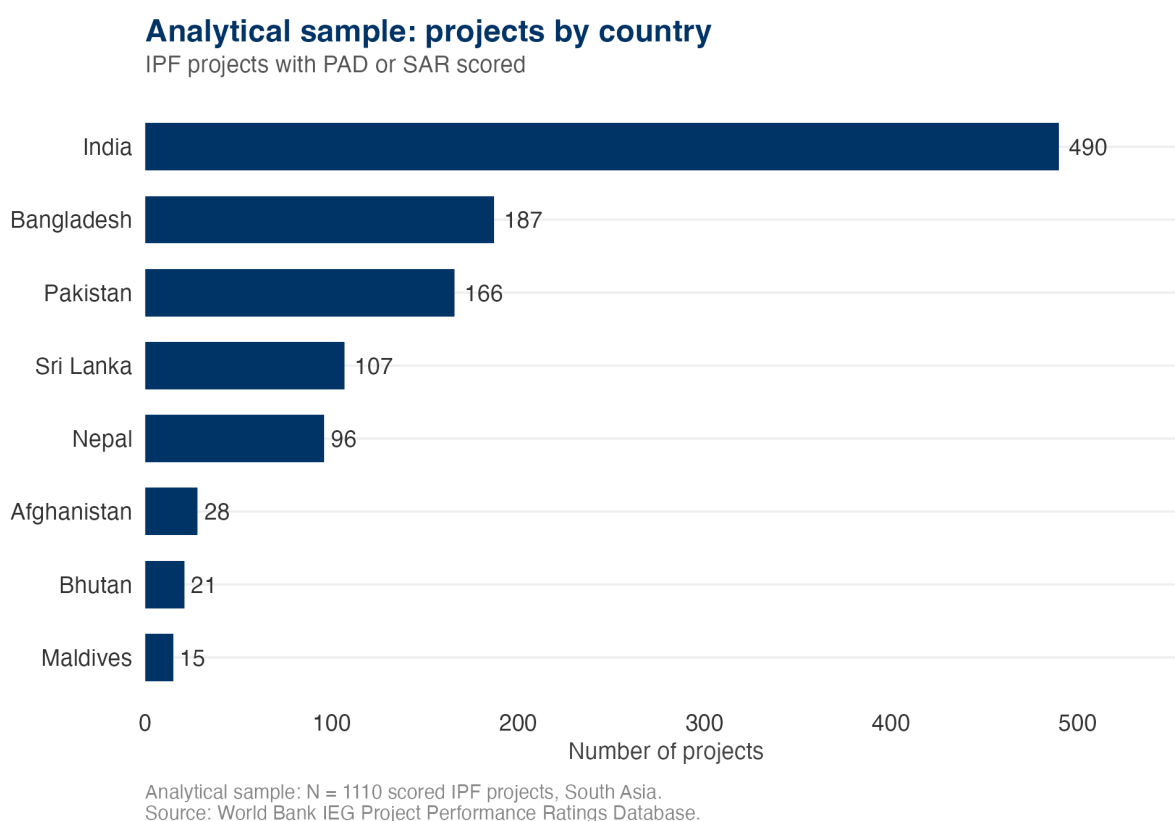
**Theory Scores.** The three theory-based scores, RBM, PDIA, and Specificity, are derived from an automated content analysis of the planning documents. Their full construction is described in Section 5.2. All three are standardized to mean zero and unit standard deviation for regression analysis.

**Tangibility.** A binary variable which classifies each project's primary objective, by looking at the sector, as tangible or intangible was created. It was based on Andrews, Pritchett and Woolcock (2017) distinction between 'easy' reforms with short, predictable causal chains and 'hard' reforms requiring institutional change and political will. Sectors with elements of both physical delivery and institutional reform, Education, Health, Urban, Environment, are coded as Mixed and treated as a separate baseline category in regressions rather than forced into either of the binary groups.

**Control Variables.** Four control variables are included in all regression specifications. *Sector* is a categorical variable with 15 categories based on the World Bank's Global Practice classification, with Education as the reference category. There were 617 fields where the Global Practice field was empty, predominantly older projects approved before the practice was formalised before 2014. This issue was fixed using a keyword classifier applied to project names, built and validated against the 874 observations that had a value in the Global Practice field. *Country* is a categorical variable with India as the reference category, capturing country-level fixed effects. *Approval decade* is a categorical variable with the 1950s as reference. *Document era* is a binary variable distinguishing PAD-era projects (post-1995) from SAR-era projects (pre-1995), which controls for structural differences between the two document formats.

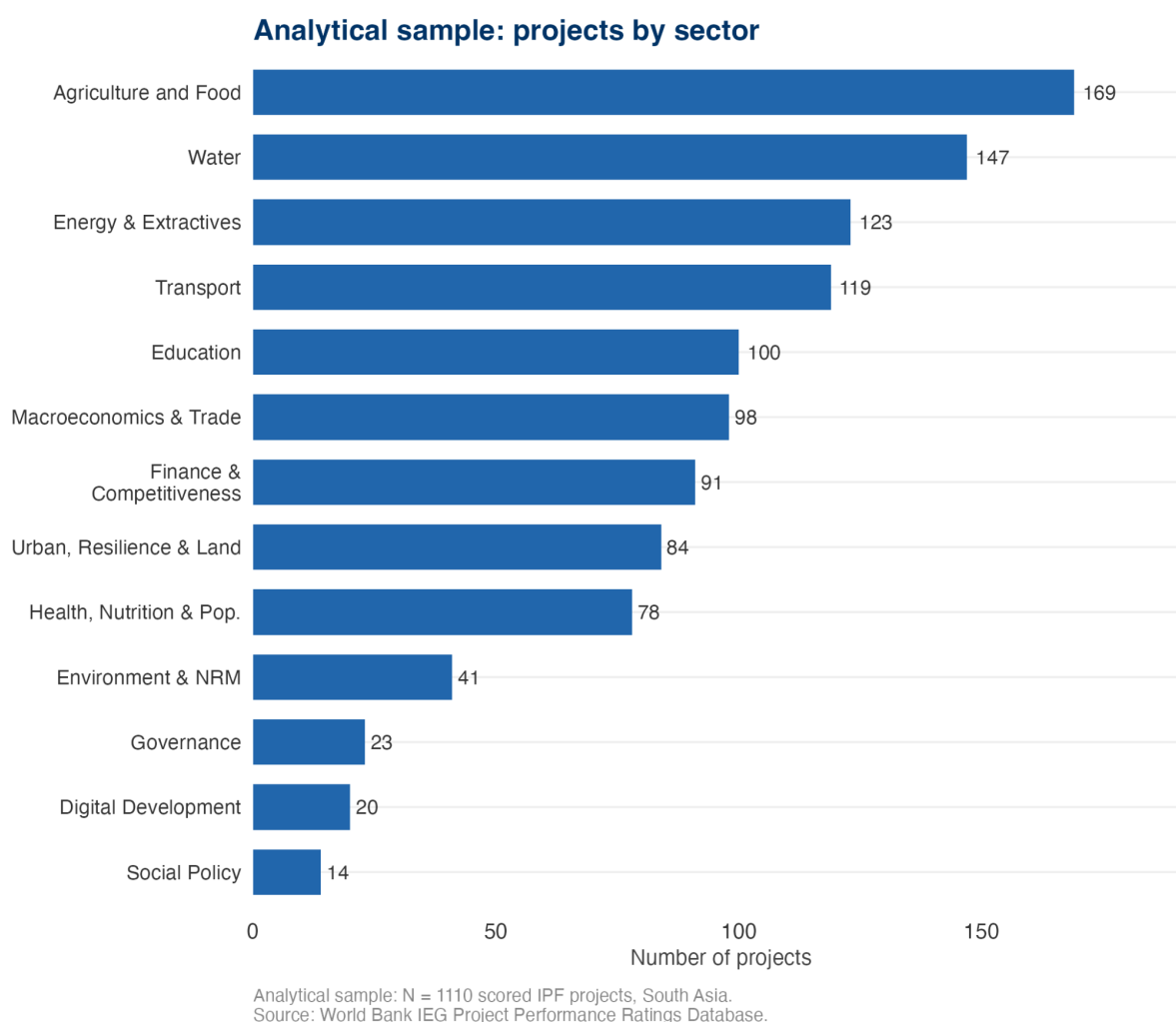
#### 5.4. Descriptive Statistics

Figure 1 shows the distribution of projects across countries. India accounts for 44.1 per cent of the sample ( $n = 490$ ), consistent with its dominant position as the World Bank's largest South Asian borrower. Bangladesh ( $n = 187$ ), Pakistan ( $n = 166$ ), Sri Lanka ( $n = 107$ ), and Nepal ( $n = 96$ ) constitute the next largest sub-samples. Afghanistan ( $n = 28$ ) and Bhutan ( $n = 21$ ) are the smallest, with Maldives ( $n = 15$ ) the smallest overall.



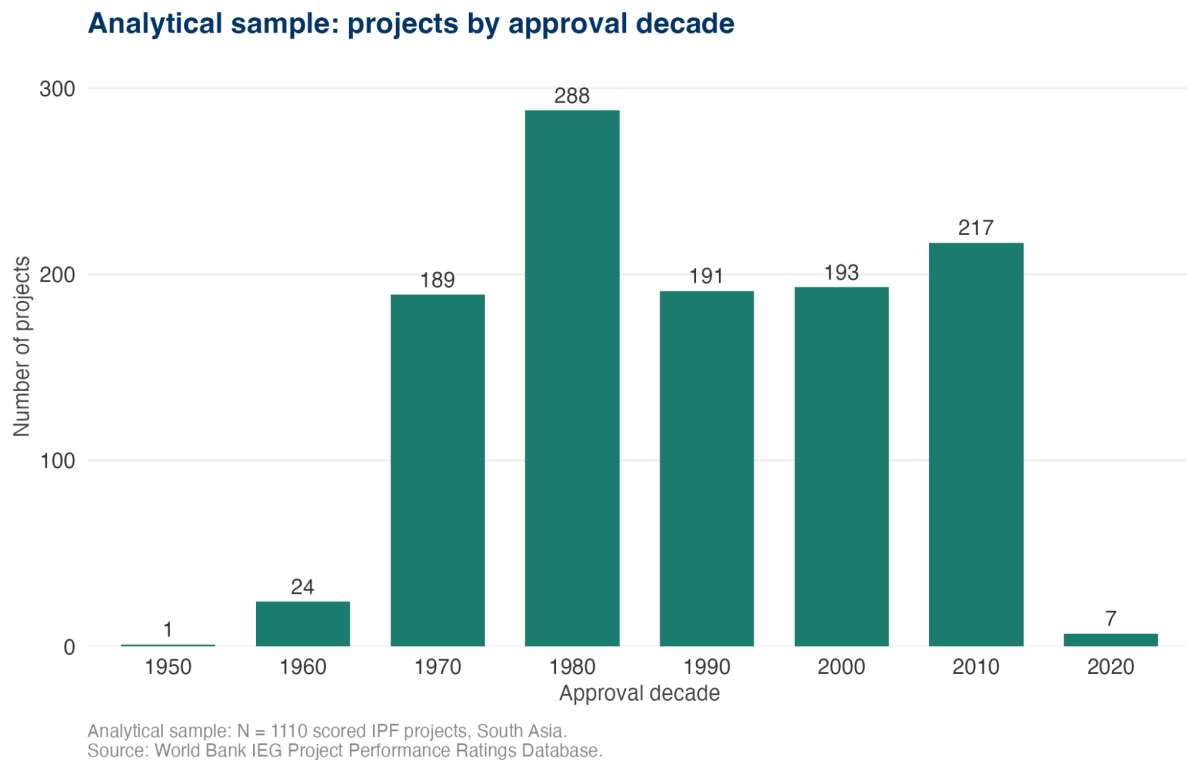
*Figure 1. Analytical sample: number of IPF projects by country ( $N = 1,110$ ).*

Figure 2 shows the sectoral composition. Agriculture and Food (n = 169) and Water (n = 147) are the largest sectors, reflecting the historical centrality of irrigation and rural development in South Asia. Energy and Extractives (n = 123) and Transport (n = 119) are the next largest. Governance (n = 23) and Digital Development (n = 20) are small; their regression coefficients carry wider confidence intervals.



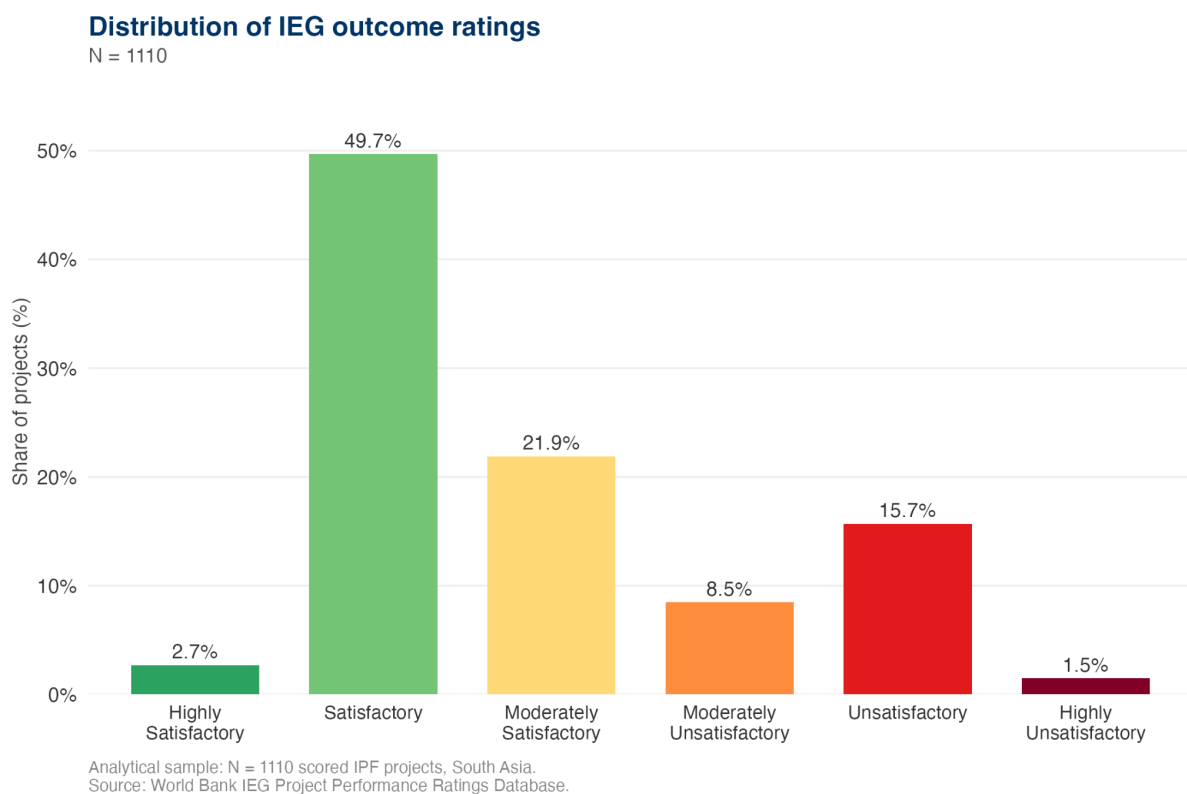
*Figure 2. Analytical sample: number of IPF projects by sector (N = 1,110).*

Figure 3 shows the temporal distribution by approval decade. The portfolio peaked in the 1980s ( $n = 288$ ), reflecting the World Bank's large structural adjustment and infrastructure lending surge. Lending volumes stabilized across the 1990s, 2000s, and 2010s ( $n = 191$ , 193, and 217 respectively). The 2020s cohort contains only seven projects and its decade coefficient is not interpreted substantively.



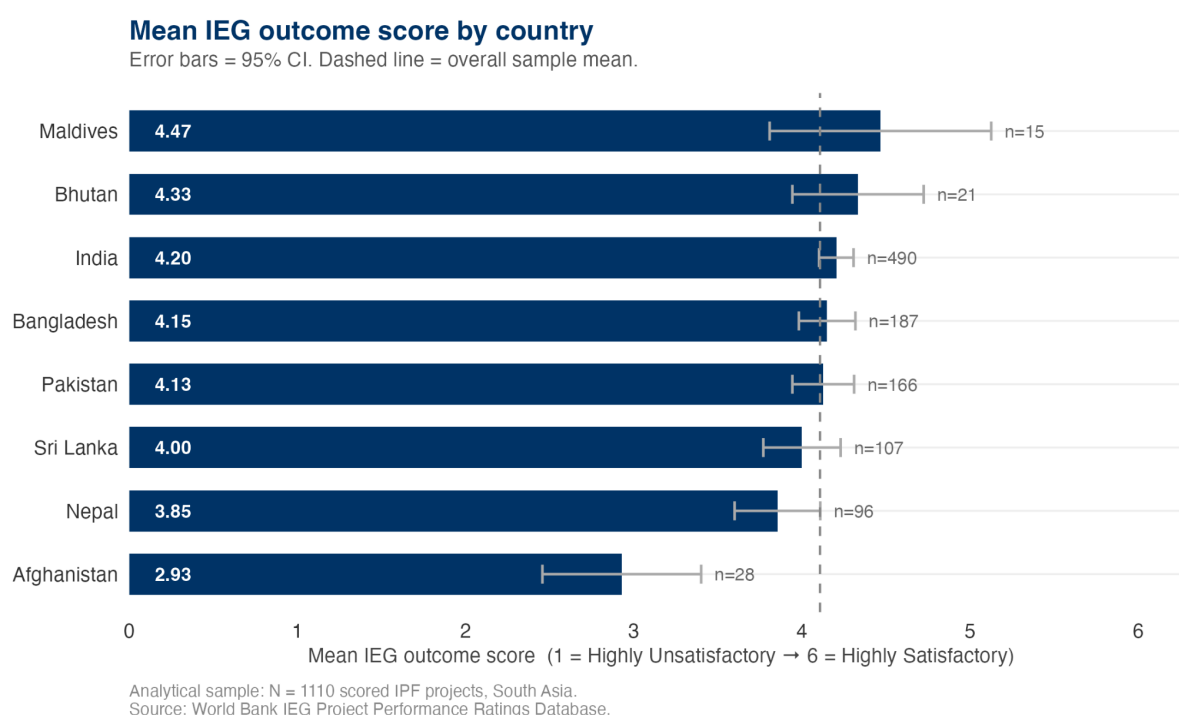
*Figure 3. Analytical sample: number of IPF projects by approval decade ( $N = 1,110$ ).*

Figure 4 shows the distribution of IEG outcome ratings. Satisfactory is the modal category (49.7 per cent), with 74.3 per cent of projects rated at or above Moderately Satisfactory. This positive skew is a well-documented feature of World Bank evaluations attributed to both genuine portfolio improvements and grade inflation in self-reported completion reports (Denizer, Kaufmann and Kraay, 2013; Michaelowa and Borrmann, 2006). The skew reduces variance in the dependent variable, compressing the range within which planning quality can explain cross-project differences, a limitation returned to in Section 6.5.



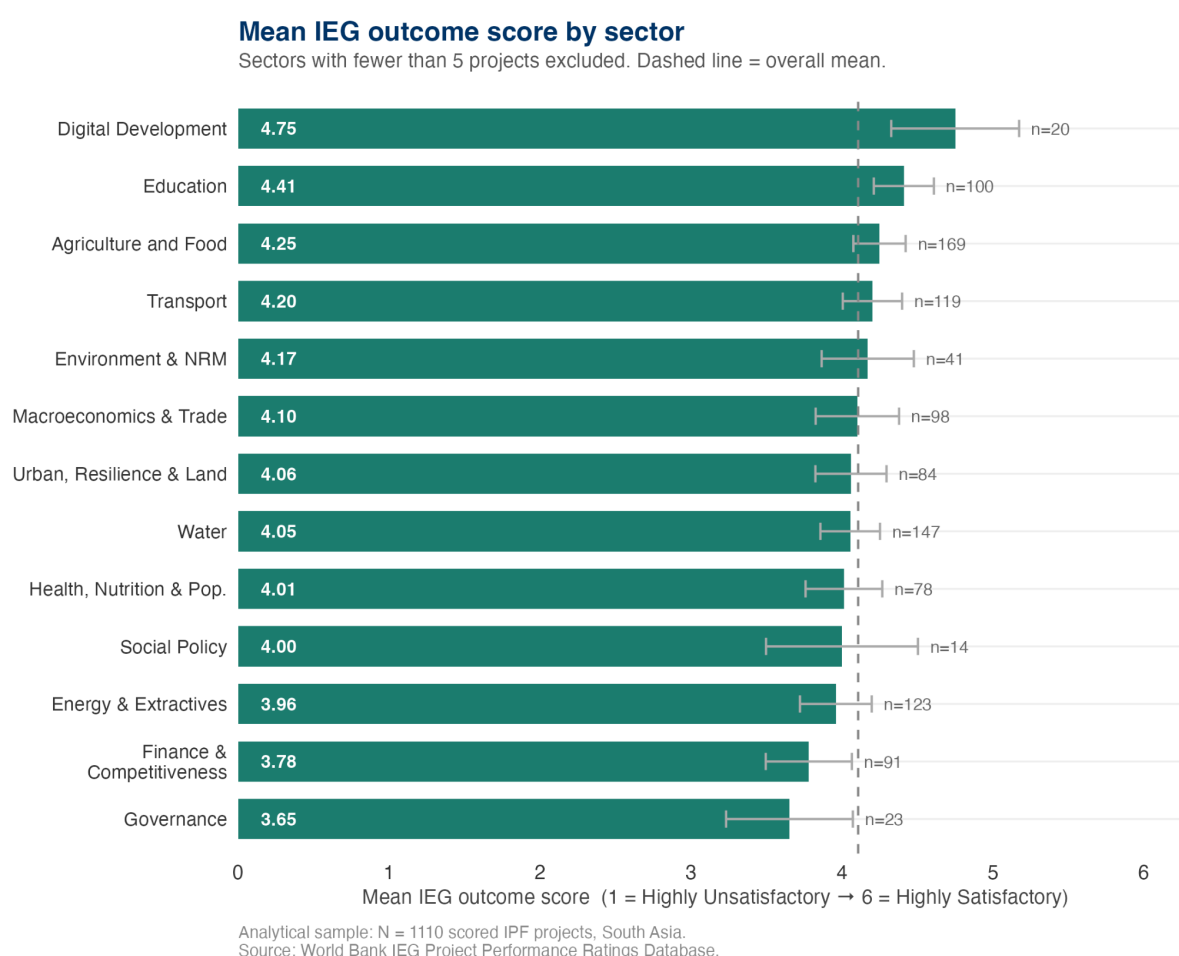
*Figure 4. Distribution of IEG outcome ratings, analytical sample (N = 1,110).*

Figure 5 shows mean IEG outcome scores by country with 95 per cent confidence intervals. Maldives (4.47) and Bhutan (4.33) record the highest means, though both carry wide intervals. India (4.20), Bangladesh (4.15), and Pakistan (4.13) cluster near the overall mean of 4.11. Afghanistan records the lowest mean by a substantial margin (2.93), reflecting the extraordinary implementation context of post-2001 reconstruction rather than systematic differences in planning quality, confirmed by the large Afghanistan country fixed effect in the regression models.



*Figure 5. Mean IEG outcome score by country with 95% CI.  
Dashed line = overall mean (N = 1,110).*

Figure 6 shows mean outcome scores by sector. Digital Development (4.75) and Education (4.41) record the highest sector means. Governance (3.65) and Finance and Competitiveness (3.78) are consistently at the lower end, a range of more than one outcome point across sectors, which is substantial relative to the standard deviation of 1.20. This cross-sectoral variation motivates the inclusion of sector fixed effects and the sector interaction models (D1-D3) described in Section 6.4.



*Figure 6. Mean IEG outcome score by sector with 95% CI.*

*Sectors with  $n < 5$  excluded. Dashed line = overall mean.*

Figure 7 shows mean outcome scores over time. The 1980s cohort records the lowest mean (3.91), consistent with the documented difficulties of structural adjustment era operations (World Bank, 1992). Scores recovered to 4.0 - 4.2 across the 1990s through 2010s with no strong secular trend. This stability across decades, despite the introduction of PADs in 1995, the adoption of results-based management in the early 2000s, and successive operational reforms, is itself suggestive of context-level determinism over planning quality, foreshadowing the main regression findings.

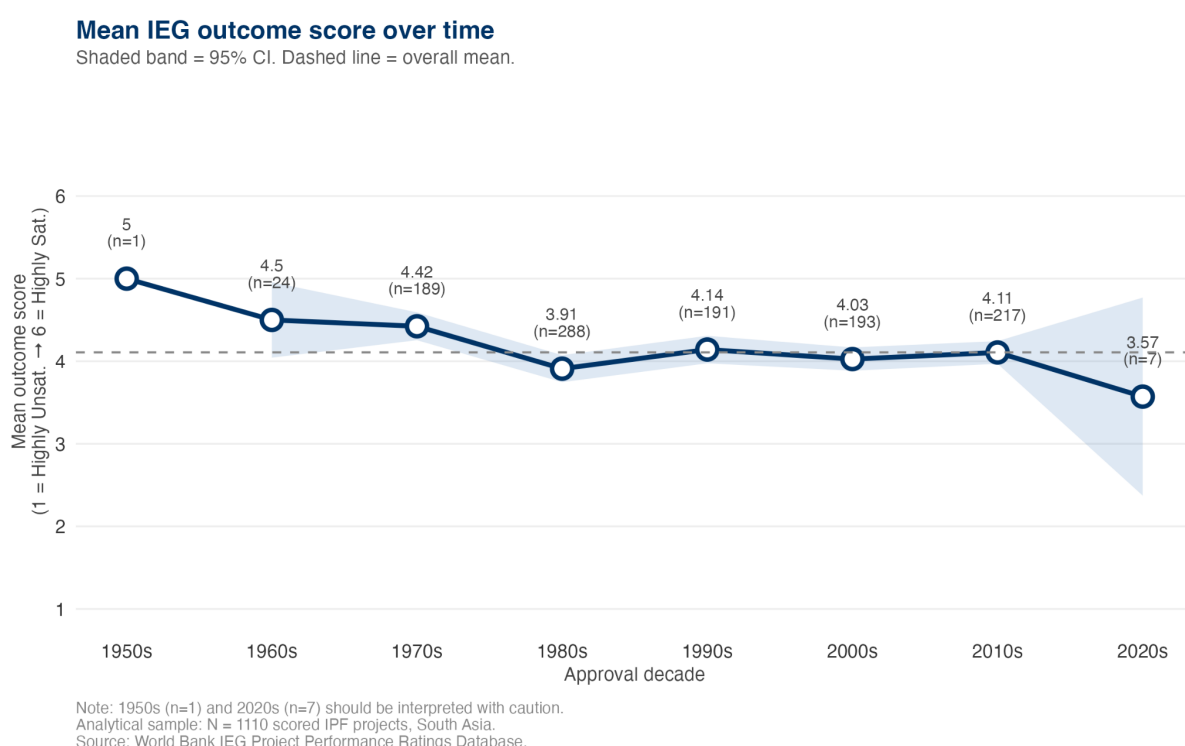


Figure 7. Mean IEG outcome score by approval decade with 95% CI.

Shaded band = 95% CI. Note: 1950s (n = 1) and 2020s (n = 7) should be interpreted with caution.

## 6. Methodology

### 6.1. Research Design

This thesis adopts a quantitative approach with a document-based research design. The core analytical strategy is to score the textual content of the project planning documents using theory-motivated keyword dictionaries, then estimate whether those scores predict independent evaluations of development outcomes recorded years later. This structure creates a predictive test: *does the quality of the plan, as characterised at the moment of appraisal, forecast the eventual outcome?* Since the planning quality is measured before the evaluation

and the IEG evaluators who produced the outcome scores did not use the planning document scores, the design avoids the most obvious forms of common-method bias.

The unit of analysis is the individual project. The analytical sample consists of 1,110 World Bank Investment Project Financing projects in South Asia with both a retractable planning document and an IEG outcome rating, approved between 1958 and 2021. Three theory-motivated text scores and one sector-level tangibility variable are the primary independent variables and the IEG outcome rating, recorded on a six-point ordinal scale, is the dependent variable.

The design is more observational than experimental. Projects are not randomly assigned to planning quality conditions. Higher-quality plans may be associated with more capable implementing agencies, or more technically tractable problems or more politically stable operating environments, all of which could independently predict better outcomes. The regression specifications include fixed effects in the form of sector, country, approval decade and a document-era control (PAD versus SAR) to absorb the most important sources of confounding variation. Nonetheless, residual confounding cannot be ruled out and is acknowledged as a limitation throughout. All code used for text extraction, keyword scoring, and regression analysis was written by the author with assistance from AI-based tools; a full account of AI use in this thesis is provided in Appendix A.

## **6.2. *Outcome Variable***

The dependent variable is the project outcome score recorded in the World Bank IEG Project Performance Ratings Database, accessed in March 2026 via the World Bank Data Catalog. The autonomous unit evaluates the results of the completed project typically within two to three years of the project closure. The assessments are conducted by independent evaluators who review project documents, interview stakeholders, and visit implementation sites. Hence, the outcome rating used in this study reflects the IEG's assessment of the extent to which the project's development outcomes were achieved.

The six-point rating scale, 1 through 6 as follows: Highly Unsatisfactory (1), Unsatisfactory (2), Moderately Unsatisfactory (3), Moderately Satisfactory (4), Satisfactory (5), and Highly Satisfactory (6). In the analytical sample, the mean outcome score is 4.11 (SD = 1.20), with Satisfactory the most common rating. Since the scale is ordered rather than continuous, the most appropriate statistical model would be an ordinal logistic regression, which is

estimated as Model C. However, given that Model C is only estimated for the horse race specification and not for the full set of sixteen models, OLS is used as the primary estimation throughout. The OLS results are also reported for all other models because OLS coefficients are more straightforward to interpret, they can be read directly as point changes on the 1-6 scale, and in practice procedure similar conclusions.

### **6.3. *Text Scoring: Theory-Motivated Dimensions***

Each planning document is scored on three dimensions derived from competing theoretical frameworks. Scores are derived from the raw PDF text using keyword density functions on R: weight counts of theoretically motivated terms and phrases, normalised by document length and mapped to the interval [0, 1] via a sigmoid transformation ( $1 - \exp(-\text{density}/k)$ ), where the scale parameter  $k$  is calibrated separately for each dimension based on the observed density range. Scoring is performed on full documents with no page-count truncation; this ensures that results frameworks and possible annexes are also captured in the analysis. To avoid rescoring due to possible interruptions, a checkpointed loop was added to every 100 documents in the pipeline which allowed the scoring to keep going and not be disturbed by possible anomalies. All code and data are publicly available in the project GitHub repository at <https://github.com/sargamkuckreja/EAEGD-Thesis-VT2026>.

The keyword dictionary uses differential weights rather than treating all matches equally. The logic applied is straightforward. Some terms are more diagnostically significant than others. They signal more genuine engagement with the theoretical framework rather than the routine planning language that appears in almost every document regardless of quality, and the weights reflect this hierarchy.

For the RBM dimension, the existence of a formal monitoring and evaluation framework received the highest weight (4) because it represents an institutional structure which requires a deliberate design effort, rather than a superficial reference (Kusek and Rist, 2004). Results-chain terminology and baseline phrasing receive the next highest weights (3), since they both need structured analytical effort to produce. Generic target mentions, by contrast, have been allocated the lowest weight (1) because the word ‘target’ alone appears in every planning document and carries little diagnostic value. This hierarchy follows OECD (2019) guidelines, which treats baseline measurements and formal M&E architecture as the most essential requirement of results-based planning. In short, the RBM score rewards documents

that have done the analytical work of measurement, not just those that use the right vocabulary.

For the PDIA dimension, uncertainty and complexity language receives the lowest weight (0.5) due to its use of ubiquitous words like ‘challenging’ and ‘complex’ in the, for example, risk section. This does not indicate genuine adaptive design intent. Political economy engagement receives higher weight (2), following Andrews, Pritchett and Woolcock (2017), who explain that explicit engagement with political constraints is the the most distinctively PDIA-aligned feature of a planning document. The weighting reflects a simple logic: words like 'complex' are less impactful, while explicit engagement with political constraints requires deliberate analytical effort.

For the Specificity dimension, all categories of the specific content are weighted equally, reflecting Pritchett, Woolcock and Andrews' (2013) argument that what matters is the overall ratio of contextually grounded to generic content, not the particular type of specific reference used. What the score is ultimately capturing is whether a document reads like it was written for this project, or for any project.

### *6.3.1. RBM Score*

The Results-Based Management framework is operationalised as the RBM score. Articulated by Kusek and Rist (2004) and OECD (2019), the RBM theory predicts that projects with strong results infrastructure at appraisal tend to achieve better development outcomes because their infrastructure allows the managers to track progress and correct problems. This infrastructure includes the presence of quantified targets, baseline data, time-bound milestones, explicit M&E frameworks and specified verification sources in the project planning. The score captures nine elements weighted by their theoretical diagnosticity, as set out in Table 7 (Appendix B).

The weighted raw count is normalised by the document length and transformed via the sigmoid  $1 - \exp(-\text{density}/30)$ , which helps in saturating the high densities and prevents outlier documents from dominating the distribution. In the analytical sample, the RBM scores range from near zero to 0.85, with a mean of 0.36 (SD = 0.13). These scores are significantly higher

for the PAD-era documents, which reflects the World Bank's formal adoption of the Results Frameworks from 2000 onwards.

### 6.3.2. *PDIA Score*

Developed by Andrews, Pritchett and Woolcock (2017) and extended by Honig (2018), the Problem-Driven Iterative Adaptation framework is operationalised by the PDIA score. PDIA theory states that projects designed with flexibility uncertainty acknowledgment, phased or pilot approaches, and explicit learning mechanisms will perform better in complex implementation environments because they are better equipped to adapt when the initial assumptions in the project are proven wrong. The full scoring dictionary is presented in Table 8 (Appendix C).

In the analytical sample, PDIA scores range from near zero to 0.55, with a mean of 0.11 (SD = 0.08), reflecting the rarity of explicit adaptive management language in World Bank planning documents.

### 6.3.3. *Specificity Score*

The specificity score operationalises the isomorphic mimicry hypothesis advanced by Pritchett, Woolcock and Andrews (2013) and Easterly (2014). Their hypothesis predicts that planning documents that are rich in generic donor boilerplate, language that appears in every project regardless of the context, tend to be more performative rather than genuine planning and should predict worse outcomes. This score measures how context-specific a document is in relation to how much generic boilerplate language it contains, with a small bonus for documents that are dense in specific references overall. The full scoring dictionary is presented in Table 9 (Appendix D).

The ratio-based specificity score was found to have a ceiling problem in the initial testing where approximately 49% of all the documents scored exactly 1.0 because they contained no generic boilerplate language at all, leaving insufficient variation for a regression analysis. As a robustness check, the specificity construct was split into two independent density components, SPECIFIC\_density and GENERIC\_density, each normalised via its own sigmoid function. In the analytical sample, SPECIFIC\_density has a mean of 0.30 (SD = 0.15) with no ceiling effect; GENERIC\_density has a mean of 0.19 (SD = 0.16).

#### 6.3.4. *Tangibility Variable*

Based on Andrews, Pritchett and Woolcock (2017), projects are classified by tangibility based on their intended outputs. Tangible projects involve physical infrastructure or a measurable service delivery, usually interventions with short predictable causal changes where the outcomes are not dependent primarily on technical execution. Intangible projects are those that require institutional change, behavior change, or improvements in governance; projects with interventions that are long, contested causal chains where outcomes depend heavily on political will, social norms, and implementation capacity. Including this distinction motivates a prediction about the conditioning role of tangibility: planning quality should be more relevant where the causal chains are predictable and targets are physically meaningful.

The classification is made at the sector level, as presented in Table 10 (Appendix E). Mixed actors combine physical delivery with institutional objectives and are kept in the full analytical sample as their own category.

In the analytical sample, 578 projects are classified as Tangible, 226 as Intangible and 306 as Mixed. The tangibility classification is used in two ways: as a main-effect control in Model F, and as a modelling variable in the theory-by-tangibility interactions models G1 through G4. Tangibility is used as a factor with Mixed as the baseline category in the main-effect specifications, and as a Binary (Tangibility versus Intangible) in the interaction specifications, which restrict the sample to these two groups.

#### 6.4. *Regression Specifications*

Sixteen regression models are estimated across seven groups, full specifications are provided in Appendix F, and Table 1 summarises their estimators and research question mapping, but the logic connecting them follows a simple sequence: establish whether any planning quality signal exists at all (Models A1-C), then ask where and when it matters (Models D1-G4).

**Table 1. Overview Of Regression Models**

| Model | Estimator | Key Variables                 | Answers  | Purpose                                                      |
|-------|-----------|-------------------------------|----------|--------------------------------------------------------------|
| A1-A3 | OLS       | One theory + controls         | RQ1      | Isolation test: each theory separately                       |
| B     | OLS       | All three theories + controls | RQ1, RQ2 | Horse race: all theories simultaneously (main specification) |

|       |                  |                                          |          |                                                                  |
|-------|------------------|------------------------------------------|----------|------------------------------------------------------------------|
| C     | Ordinal Logistic | All three theories + controls            | RQ1, RQ2 | Preferred specification respecting the ordinal DV                |
| D1-D3 | OLS              | Theory $\times$ sector interactions      | RQ3      | Does the theory-outcome relationship vary by sector?             |
| E1-E3 | OLS              | Theory $\times$ decade interactions      | RQ2      | Has the theory-outcome relationship changed over time?           |
| F     | OLS              | All three + tangibility (main effect)    | RQ3      | Does tangibility predict outcomes independently of plan quality? |
| G1-G4 | OLS              | Theory $\times$ tangibility interactions | RQ3      | Does tangibility condition the effect of planning quality?       |

#### 6.4.1. *Isolation and Horse Race Models (A1-A3, B, C)*

Model A1, A2, and A3, yield an isolated estimation uncontaminated by other scores by testing each theory separately with the full controls. Model B is the horse race where all the theories are entered simultaneously. This allows all the coefficients to be interpreted as the partial effect of each theory while controlling for the others. If theories are collinear, for example, because ‘well-written’ documents tend to score highly on all dimensions, the isolation models should show the apparent effects that would disappear in the horse race and conversely the suppressor variables can cause effects to only appear when all theories are entered together.

In Model C, the horse race is re-estimated using an ordinal logistic regression. It is the proportional odds model, estimated using the polar function in R's MASS package. This ordinal specification is methodologically more advantageous because the six-point IEG rating scale is explicitly ordered but not demonstrably interval-scaled, i.e. the distance between Moderately Satisfactory and Satisfactory need not equal the distance between Satisfactory and Highly Satisfactory. The proportional odds coefficients represent the change in the log-odds of being in a higher outcome category associated with a one-unit increase in the predictor. Because interpretation requires exponentiation, OLS coefficients from Model B are used for the primary discussion of effect sizes.

#### 6.4.2. *Interaction Models (D1-D3, E1-E3)*

Models D1 to D3 interact each theory score with the sectors, testing whether the theory-outcome relationship varies across the sectors. Models E1 to E3 interact each theory with the decade variable, testing if the relationship has changed over the study period. Only statistically significant interaction terms ( $p < 0.10$ ) are reported in the findings, given that

these models involve a large number of interaction cells. The D and E models are estimated with OLS and include the same full set of controls as Model B.

#### *6.4.3. Tangibility Models (F, G1-G4)*

Model F used tangibility as the main-effect control alongside the three theory scores for the full analytical sample. For the three levels in the tangibility factor, Tangible is the baseline and this model tests whether tangible projects systematically outperform intangible ones regardless of planning quality, after controlling for sector, country, decade, and document era.

Models G1 through G4 estimate the theory-by-tangibility interactions, restricting the sample to projects classified as either Tangible or Intangible (N = 804). Model G1 used the results framework approach (RBM) scores with the binary tangible indicator, Models G2 and G3 interact with problem-driven approach (PDIA) and specificity respectively and lastly, G4 enters all three interactions simultaneously. Getting a positive and significant RBM×Tangible interaction would effectively show that RBM's predictive power is concentrated in infrastructure and service delivery sectors, consistent with the expectation that measurable targets are more meaningful where outputs are physically countable. A positive PDIA×Intangible interaction would indicate that adaptive design pays off specifically in institutionally complex contexts.

### **6.5. Identification and Limitations**

Five identification concerns are relevant to interpreting the regression results.

First, while reverse causation in the strict sense is not plausible, since the plan precedes the evaluation by construction, endogeneity remains a concern through selection on unobservables. More experienced or capable implementing agencies might produce both better plans and also better outcomes, which might generate an upward bias in the theory-score coefficients. The sector and country fixed effects partly address this but the within-sector and within-country variation in institutional capacity is not fully controlled for. This is a standard limitation of observational designs in the aid effectiveness literature (Clemens et al., 2012; Deaton, 2010).

Second, the project outcomes as rated by IEG cannot be attributed solely to the World Bank's planning quality. The development outcomes seen in/via the projects are shaped by a full

ecosystem of actors operating there: bilateral donors, NGOs, other multilateral institutions, and domestic government programmes; whose activities may coincide with/complement/undermine a World Bank project over its implementation period. The IEG rating does not take into account the control for co-interventions, meaning that if a project rated Satisfactory may have benefited from parallel investments by other actors, while a project rated Unsatisfactory might have faced a more contradictory aid environment. This has an impact on the extent to which variation in outcome can be attributed to World Bank project designs specifically (Roodman, 2007; Winters, 2010).

Third, the text scores generated are indirect measures of the theoretical constructs they are meant to measure. Keyword counting identifies the presence of relevant language but cannot directly assess whether the language reflects genuine analytical thinking. An experienced implementing agency could produce a high RBM score by populating the planning document with a standard results framework template without internalising the measurement logic. This limitation has been curbed by the inclusion of the specificity score: if high-scoring projects are genuinely analytically engaged with context, they should also use more local-specific language.

Fourth, the tangibility classification is made at the sector level and therefore might misclassify individual projects that cross sectoral boundaries. A water project which aims to reform regulatory processes is coded as tangible because it is in the Water sector even though its actual objects are intangible. This measurement error attenuates the tangibility interaction towards zero, making the non-significant results a conservative test of the tangibility hypothesis.

Fifth, attrition is also a possible concern. Projects for which no planning document could be retrieved may differ systematically from those that make up the analytical sample. As mentioned in Chapter 4, the document retrieval rate is 84% of the full IEG-rated IPF portfolio, and those not retrieved are more likely to be the older (SAR-era) and from countries with historically weaker document management systems. If older, less well-documented projects also tend to perform worse, the analytical sample might be skewed towards better-performing projects, reducing the range of outcome variation and making it harder to detect statistically significant effects.

## 7. Results and Findings

### 7.1. Descriptive Overview

In the analytical sample, the mean IEG outcome score is 4.11 on a one-to-six scale (SD = 1.20), which roughly corresponds to a ‘Moderately Satisfactory’ rating. All three theory scores are standardized to a mean of zero and a standard deviation of one. This standardization puts all three scores on the same scale so that their effects can be compared directly. Among projects classified as either tangible or intangible (N = 804, excluding the 306 mixed projects), approximately 72 per cent are tangible and 28 per cent are intangible. Across the full analytical sample of 1,110 projects, tangible projects account for 52.1 per cent, intangible for 20.4 per cent, and mixed for 27.6 per cent.

**Table 2. Descriptive Statistics**

| Variable                           | Mean  | SD    | Min    | Max   |
|------------------------------------|-------|-------|--------|-------|
| IEG outcome score                  | 4.107 | 1.202 | 1.000  | 6.000 |
| RBM score (std.)                   | 0.000 | 1.000 | -2.482 | 5.136 |
| PDIA score (std.)                  | 0.000 | 1.000 | -1.280 | 9.221 |
| Specificity score (std.)           | 0.000 | 1.000 | -1.705 | 4.456 |
| Share tangible (among non-mixed)   | 0.719 |       |        |       |
| Share intangible (among non-mixed) | 0.281 |       |        |       |

*Note: N = 1,110. Theory scores standardised (mean = 0, SD = 1). Share tangible and share intangible are computed over the binary subsample (N = 804) used in Models G1–G4, excluding mixed projects.*

The distributions of the theory scores show sufficient variations across the sample to identify meaningful effects in regression. RBM score ranges from -2.48 to 5.14 standard deviations, which reflects substantial variation in the quality of results infrastructure across projects over time. PDIA scores are similarly distributed, -1.28 to 9.22, with the right tail reflecting a small number of projects that score very highly on the adaptive design language. Specificity and generic density scores have a narrower range but again show enough variation to identify effects in regression.

### 7.2. Isolation Models: Testing Each Theory Separately (Models A1–A3)

The first set of models, the isolation models, tested each theory score independently against the full set of control, including sector, country and decade fixed effects and document-era

dummy distinguishing between PADs and SARs. The isolation models directly address the first research question: does the quality of project design documents, measured at appraisal, predict the development outcomes those projects eventually achieve?

The first hypothesis (H1) states that the projects with stronger results-based management infrastructure at appraisal, more quantified targets, baseline data, and formal monitoring and evaluation frameworks, will achieve better outcomes. Model A1 provides clear support for this hypothesis. The isolation models serve as the first test for whether planning quality leaves any detectable footprint on outcomes at all. Each theory is given its best chance to shine before being put into competition with the others. The RBM score is positive and has a statistically significant coefficient ( $\beta = 0.095$ ,  $SE = 0.040$ ,  $p < 0.05$ ), which in practical terms means that if a projects scores one standard deviation higher on the results infrastructure at appraisal, then there is roughly a 0.10-point improvement on the six-point outcomes scale. This result is modest and will not save project operation in a fragile state but it is consistent and also survives a demanding set of controls.

**Table 3. OLS Regressions: Each Theory Tested Separately (A1-A3)**

|                          | IEG Outcome Score (1-6) |                     |                     |
|--------------------------|-------------------------|---------------------|---------------------|
|                          | RBM (A1)                | PDIA (A2)           | Specificity (A3)    |
| RBM score (std.)         | 0.095**<br>(0.040)      |                     |                     |
| PDIA score (std.)        |                         | 0.015<br>(0.033)    |                     |
| Specificity score (std.) |                         |                     | 0.012<br>(0.042)    |
| Doc. era: PAD (ref. SAR) | -0.412**<br>(0.168)     | -0.351**<br>(0.167) | -0.343**<br>(0.167) |
| Sector FE                | Yes                     | Yes                 | Yes                 |
| Country FE               | Yes                     | Yes                 | Yes                 |
| Decade FE                | Yes                     | Yes                 | Yes                 |
| Observations             | 1,110                   | 1,110               | 1,110               |
| R <sup>2</sup>           | 0.089                   | 0.085               | 0.084               |
| Adjusted R <sup>2</sup>  | 0.064                   | 0.059               | 0.059               |

*Note: Robust standard errors (HCl) in parentheses. Theory scores standardised (mean = 0, SD = 1). Reference categories: Education (sector), India (country), SAR (document era), 1950s (decade). \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .*

The second hypothesis (H2) predicts that projects with greater adaptive design language, consistent with PDIA principles, will perform better in complex institutional environments. Model A2 provides support for this hypothesis in its unconditional form. The PDIA score has a small and insignificant coefficient ( $\beta = 0.015$ ,  $SE = 0.033$ ,  $p > 0.1$ ), indicating that adaptive design language in isolation does not reliably predict outcomes across the full sample. This finding is not entirely surprising because if the PDIA-style planning is most crucial for institutionally complex settings then its effect would be expected to appear in conditional models that interact the score with project type rather than in isolation.

Hypothesis three (H3) states that higher generic boilerplate density is negatively associated with outcomes, reflecting the mimicry problems described in the theoretical framework. Model A3 also fails to support this hypothesis. The specific score shows a negligible and insignificant coefficient ( $\beta = 0.012$ ,  $SE = 0.042$ ,  $p > 0.1$ ), and this null result persists across all subsequent specifications.

Across all three isolation models, the document-era dummy is large and consistently significant. Projects designed in the old project planning reports (PAD) era score systematically lower on evaluation outcomes than those in the newer planning documents (SAR) ( $\beta \approx -0.35$  to  $-0.41$ ,  $p < 0.05$ ), net of decade fixed effects. This likely reflects compositional shifts in the portfolio, the PAD era corresponds to an expansion into more institutionally demanding sectors, and proves to be a useful reminder that project design did not improve linearly with the introduction of more formal planning templates. The adjusted  $R^2$  values across the isolation models range from 0.059 to 0.064, which is modest as expected given the difficulty of explaining outcome variance from document-level variables alone.

### ***7.3. The Horse Race: Competing Theories Simultaneously (Models B and C)***

The horserace model (Model B) enters all three theories at the same time, alongside the full controls. This is aimed at testing whether the RBM findings from Model A1 reflects a genuine relationship or is instead driven by correlation between the theory scores. It also partially answers the second research question: which theory of effective planning best predicts outcomes when they compete against one another?

The horse race is the most demanding regression, where we find out whether any of the planning quality signals survive when forced to compete against each other simultaneously. For RBM, the answer is yes. The results based management score retains almost exactly the same score in the horse race as it did in the isolation model ( $\beta = 0.094$ ,  $SE = 0.041$ ,  $p < 0.05$ ), which highlights an important finding. The RBM signal is not a statistical artefact of correlation with the other dimensions of design quality but rather it also stands on its own. PDIA and specificity both enter with near-zero and wholly insignificant coefficients ( $\beta = 0.002$  and  $0.002$  respectively), consistent with their null results in isolation. The horse race is not close. Results-based management infrastructure is the only project design dimension that independently predicts outcomes in this sample.

**Table 4. Horse Race OLS (B) and Ordinal Logistic Regression (C)**

|                          | <i>IEG Outcome Score</i> |                      |
|--------------------------|--------------------------|----------------------|
|                          | OLS horse race (B)       | Ordinal logistic (C) |
| RBM score (std.)         | 0.094**<br>(0.041)       | 0.171**<br>(0.068)   |
| PDIA score (std.)        | 0.002<br>(0.033)         | -0.014<br>(0.060)    |
| Specificity score (std.) | 0.002<br>(0.042)         | -0.021<br>(0.069)    |
| Doc. era: PAD (ref. SAR) | -0.411**<br>(0.168)      | -0.881***<br>(0.280) |
| Sector FE                | Yes                      | Yes                  |
| Country FE               | Yes                      | Yes                  |
| Decade FE                | Yes                      | Yes                  |
| Observations             | 1,110                    | 1,110                |
| R <sup>2</sup>           | 0.089                    |                      |
| Adjusted R <sup>2</sup>  | 0.062                    |                      |

*Note: Robust standard errors (HC1) in parentheses for Model B. Model C: ordinal logistic (proportional odds); standard errors from model Hessian; threshold parameters suppressed. Theory scores standardised (mean = 0, SD = 1). Reference: Education (sector), India (country), SAR (document era), 1950s (decade). \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .*

Model C replicates the horse race specification using an ordered logistic regression, treating the IEG outcome score as an ordinal rather than cardinal variable. The RBM coefficient in Model C is 0.171 ( $SE = 0.068$ ,  $p < 0.05$ ), consistent with the OLS result. PDIA and specify once again produce negligible and insignificant results. The convergence of the OLS and ordinal logistic results strengthens confidence that the RBM finding is not an artefact of the assumption that IEG outcome categories are equally spaced.

#### **7.4. *Sector Interactions (Models D1-D3)***

Having established that results-based planning predicts outcomes on average, the next question then becomes whether this relationship is uniform across sectors or whether certain contexts are particularly responsive to structured planning. The third set of models interacts with each theory with the sector variable, allowing the relationship between design quality and outcomes to vary across the twelve sectors categories in the sample. Full results are reported in Table 11 (Appendix G). This addresses the question whether certain sectors are particularly responsive to RBM-style planning or to the other frameworks.

For RBM, Model D1, the main effect in the base category (Education) turns negative and insignificant once the interactions are introduced, as expected since the reference category absorbs baseline variation. The more interesting finding is that two sectors are statistically significant and positively interact with RBM: Energy and Extractives ( $\beta = 0.390$ ,  $SE = 0.157$ ,  $p < 0.05$ ) and Finance and Competitiveness ( $\beta = 0.297$ ,  $SE = 0.137$ ,  $p < 0.05$ ). From both cases, it can be deduced that projects with stronger RBM infrastructure are associated with better outcomes.

For the PDIA (Model D2), there is only one significant interaction and that is with the Finance and Competitiveness sector which returns a negative and marginally significant coefficient ( $\beta = -0.444$ ,  $SE = 0.267$ ,  $p < 0.1$ ). This suggests that adaptive design language is associated with worse outcomes in this sector, possibly showing that it is difficult to translate flexibility into results in an environment where institutional reform depends on sustained political commitments rather than project-level adaptations. All other PDIA sector interactions are close to zero and statistically indistinguishable from noise.

The specificity results tell a more differentiated story, and one finding stands out clearly enough to warrant emphasis before turning to the full picture. In Model D3, the Governance sector interaction is large and highly significant ( $\beta = 0.557$ ,  $SE = 0.168$ ,  $p < 0.01$ ), the strongest sector-level finding in the entire analysis. Its interpretation is relatively straightforward: in governance reform, the document's theory of change either engages with the actual institutional landscape or it doesn't, and evaluators can apparently tell the difference. Beyond this, the overall  $R^2$  increases only modestly compared to the main-effect

models, suggesting that sector heterogeneity explains a limited share of outcome variance outside of this specific finding.

### **7.5. Temporal Trends (Models E1-E3)**

The decade interaction models test whether the predictive power of each theory has changed over time. Full results are reported in Table 12 (Appendix H). This addresses hypothesis (H6) that RBM effects strengthened after the formal institutionalisation of Results Frameworks in the early 2000s, and that the specificity penalty deepened as boilerplate language spread through the standardised PAD template. The temporal models are the most technically constrained of the sixteen specifications, and the collinearity problem that runs through Models E1 and E2 means the story they can tell is limited. The clearest signal comes from Model E3.

Model E1 (RBM  $\times$  decade) produces large standard errors on all of the interaction terms. This is a consequence of collinearity between the decade dummies and the RBM score, since the score itself has evolved substantially over time as the Bank's M&E requirements were progressively formalised. Although, the main effect of RBM in the 2020s reference decade is positive and large ( $\beta = 0.306$ ), the interaction terms for earlier decades are all negative and of similar magnitude, which leaves the net effect across decades imprecisely estimated. Overall, the data does not provide strong evidence that RBM became more predictive over time, however, the collinearity problem denotes that this hypothesis cannot be definitively ruled out.

The PDIA decade interaction in Model E2 shows a parallel pattern of multicollinearity-driven imprecision, with all the interaction terms being large in absolute value and none statistically significant. In other words, the temporal story for the adaptive design argument can not be told from the appraisal documents alone.

In this context, the specificity decade interactions (Model E3), has the most interesting findings. In the 2020s reference decade, the specificity score is positive and significant ( $\beta = 0.417$ ,  $SE = 0.187$ ,  $p < 0.05$ ), suggesting context-specific content predicts better outcomes in recent projects. But this effect was weaker in the early decades. The interactions for the 1960s ( $\beta = -0.534$ ,  $p < 0.1$ ), 1980s ( $\beta = -0.525$ ,  $p < 0.05$ ), 1990s ( $\beta = -0.591$ ,  $p < 0.01$ ), and 2000s ( $\beta$

= -0.388,  $p < 0.05$ ) are all significantly negative. A possible explanation could be that context-specific language mattered less, or had no effect, in earlier eras when project appraisal was less standardised. Its predictive power has increased since the introduction of the PAD template and specificity has become a more meaningful signal which distinguishes between genuinely trailers documents from the boilerplate ones.

### **7.6. *Tangibility as a Moderating Factor (Models F and G1-G4)***

The tangibility models can be deemed as the most theoretically central to the analysis in the thesis. The said models test hypothesis H4, whether RBM effects are stronger for tangible projects, and hypothesis H5, whether PDIA effects are stronger for intangible ones. When taken together these hypotheses predict that the right kind of planning discipline varies systematically with the nature of the problem that the project is trying to solve. Figure 8 illustrates this pattern descriptively. Intangible projects record the lowest mean IEG outcome score (3.92), compared to tangible (4.15) and mixed projects (4.17), with confidence intervals suggesting the gap between intangible and the other two categories is meaningful. This descriptive pattern is consistent with the theoretical expectation that institutionally complex projects face structural disadvantages that are difficult to compensate through design quality alone. Figure 9 adds a further descriptive observation that complicates the picture. Intangible projects score higher on RBM than tangible ones despite achieving worse outcomes. This divergence between planning quality and outcomes in the intangible subsample is consistent with the isomorphic mimicry hypothesis: results framework language in institutionally complex projects may reflect compliance with donor requirements rather than genuine analytical investment in monitoring and evaluation design (Pritchett, Woolcock and Andrews, 2013).

### Mean IEG outcome score by project tangibility

Tangible = infrastructure/service delivery; Intangible = institutional reform

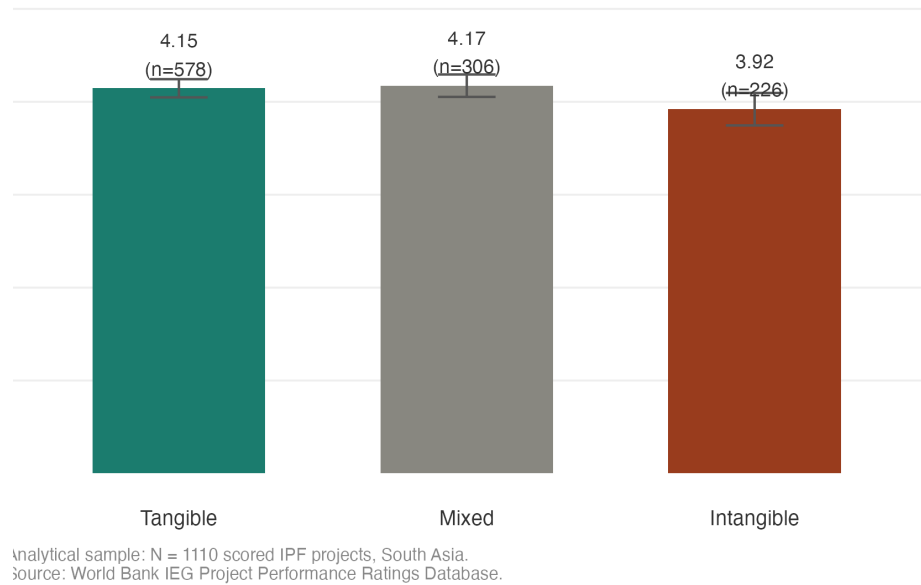


Figure 8. Mean IEG outcome score by project tangibility with 95% CI.

Tangible = infrastructure and service delivery; Intangible = institutional reform and governance.

Dashed line = overall mean (N = 1,110).

### Mean theory scores by project tangibility

Intangible projects score higher on RBM despite worse outcomes — consistent with mimicry hypothesis

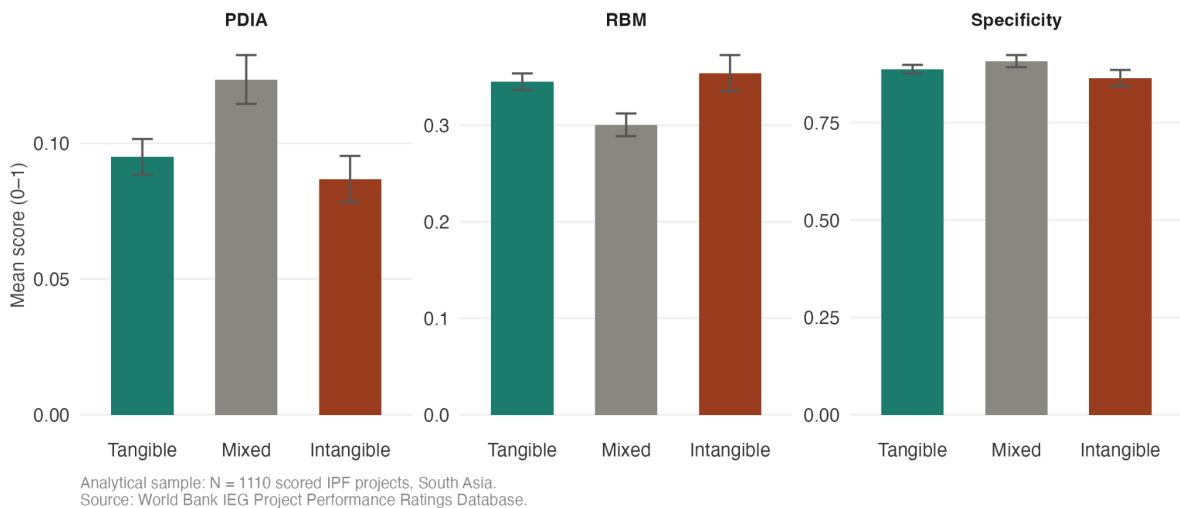


Figure 9. Mean theory scores by project tangibility with 95% CI.

Scores presented on the unstandardised 0-1 scale. Intangible projects score higher on RBM despite worse average outcomes, consistent with the mimicry hypothesis. Analytical sample: N = 1,110 (N = 1,110).

**Table 5. Tangibility Models: Main Effect (F) and Interactions (G1-G4)**

|                                         | IEG Outcome Score (1-6) |                     |                     |                     |                     |
|-----------------------------------------|-------------------------|---------------------|---------------------|---------------------|---------------------|
|                                         | Main effect<br>(F)      | RBM<br>(G1)         | PDIA<br>(G2)        | Spec.<br>(G3)       | Joint<br>(G4)       |
| RBM score (std.)                        | 0.089**<br>(0.038)      | 0.120*<br>(0.065)   | 0.102**<br>(0.043)  | 0.099**<br>(0.044)  | 0.102<br>(0.069)    |
| PDIA score (std.)                       | -0.003<br>(0.032)       | 0.005<br>(0.040)    | -0.051<br>(0.091)   | 0.005<br>(0.040)    | -0.043<br>(0.091)   |
| Specificity score (std.)                | 0.011<br>(0.040)        | -0.020<br>(0.047)   | -0.019<br>(0.047)   | 0.041<br>(0.078)    | 0.037<br>(0.082)    |
| Tangibility: Mixed (ref. Tangible)      | 0.082<br>(0.088)        |                     |                     |                     |                     |
| Tangibility: Intangible (ref. Tangible) | -0.220**<br>(0.099)     |                     |                     |                     |                     |
| Tangible (ref. Intangible)              |                         | 0.245**<br>(0.100)  | 0.252**<br>(0.098)  | 0.232**<br>(0.100)  | 0.243**<br>(0.100)  |
| RBM × Tangible                          |                         | -0.030<br>(0.080)   |                     |                     | -0.007<br>(0.085)   |
| PDIA × Tangible                         |                         |                     | 0.069<br>(0.097)    |                     | 0.059<br>(0.097)    |
| Specificity × Tangible                  |                         |                     |                     | -0.091<br>(0.089)   | -0.085<br>(0.094)   |
| Doc. era: PAD (ref. SAR)                | -0.421**<br>(0.168)     | -0.497**<br>(0.240) | -0.501**<br>(0.241) | -0.507**<br>(0.240) | -0.498**<br>(0.242) |
| Country FE                              | Yes                     | Yes                 | Yes                 | Yes                 | Yes                 |
| Decade FE                               | Yes                     | Yes                 | Yes                 | Yes                 | Yes                 |
| Observations                            | 1,110                   | 804                 | 804                 | 804                 | 804                 |
| R <sup>2</sup>                          | 0.074                   | 0.088               | 0.088               | 0.089               | 0.089               |
| Adjusted R <sup>2</sup>                 | 0.057                   | 0.064               | 0.065               | 0.065               | 0.063               |

*Note: Model F uses the full sample (tangible, mixed, intangible) with tangibility as a main-effect control replacing sector FE. Models G1-G4 use the binary subsample (tangible and intangible only; mixed excluded; N = 804). Reference category: Intangible. Country and decade FE included in all models. Robust standard errors (HCl) in parentheses. Theory scores standardised (mean = 0, SD = 1). Reference: India (country), SAR (document era), 1950s (decade). \*p<0.1; \*\*p<0.05; \*\*\*p<0.01.*

The theoretical stakes are the highest in the tangibility models. If the theory stands true, then the data should show that results infrastructure matters more for road-building than for governance reform, and that adaptive design pays off specifically where institutional complexity is greatest. However, the results from the regression are surprising.

Model F incorporates tangibility as a three-category main effect (tangible, mixed and intangible) along with the full theory scores and controls. The most notable result from this regression is the large and significant coefficient on the intangible category ( $\beta = -0.220$ , SE =

0.099,  $p < 0.05$ ). This indicates that the intangible projects perform systematically worse than tangible ones, *ceteris paribus*. This is in line with the theoretical expectation that the projects in institutionally complex sectors face structural disadvantages, which are difficult to compensate through design quality alone. The mixed category is not significantly different from tangible projects.

Models G1 through G4 incorporate just the binary subsample of clearly tangible and clearly intangible projects ( $N = 804$ , excluding mixed projects) and introduces interaction terms between each theory score and tangible status. In this scenario, the tangible dummy enters with consistently positive and significant coefficients across all specifications, ( $\beta \approx 0.23-0.25$ ,  $p < 0.05$ ), reinforcing the main-effect result.

Looking at the interaction terms, the results are instructive in their null. The  $RBM \times Tangible$  interaction in Model G1 is  $-0.030$  ( $SE = 0.080$ ), statistically indistinguishable from zero. Hence, H4 is not supported. Similarly,  $PDIA \times Tangible$  interaction in Model G2 is  $0.069$  ( $SE = 0.097$ ), again insignificant. Therefore, H5 is not supported either. The joint model G4, which includes all three interaction terms at the same time, confirms this picture: none of the interactions approach significance, and the Adjusted  $R^2$  of  $0.063$  is barely above the main-effect model.

When taken together, the tangibility models tell a clear story. Project type shakes the baseline outcome, where intangible projects are genuinely harder, but it does not appear to moderate the relationship between design quality and outcomes in the way the theory predicted. RBM predicts the outcomes in the full sample regardless of the tangibility category and neither RMB nor PDIA are definitely effective depending on what kind of project is being designed. This is not a failure of the analysis, it is a substantively important result. It suggests that the accountability structures created by a well-designed results framework carry value regardless of whether a project is about laying pipes or reforming a civil service.

### **7.7. Summary of Findings Against Hypotheses**

Across sixteen regression models, the analysis produces a consistent and interpretable pattern of results. Table 6 below summarises the status of each hypothesis.

**Table 6. Summary of Hypotheses and Findings**

| H         | Hypothesis                                                                                                                    | Predicted Direction     | Verdict         | Key Evidence                                                                                                                              |
|-----------|-------------------------------------------------------------------------------------------------------------------------------|-------------------------|-----------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| <b>H1</b> | Projects with stronger RBM infrastructure at appraisal achieve better outcomes                                                | +                       | Supported       | $\beta = 0.095^{**}$ in A1; holds in horse race (B) and all tangibility models                                                            |
| <b>H2</b> | Projects with greater PDIA-style adaptive design language perform better in complex settings                                  | +                       | Not supported   | $\beta = 0.015$ in A2; near zero in B and all interaction models                                                                          |
| <b>H3</b> | Higher generic boilerplate density is negatively associated with outcomes; context-specific language is positively associated | - / +                   | Partial support | Null in aggregate (A3, B); Specificity $\times$ Governance positive and significant (D3: $\beta = 0.557^{***}$ )                          |
| <b>H4</b> | RBM effects are stronger for tangible projects than intangible ones                                                           | Larger + for tangible   | Not supported   | RBM $\times$ Tangible = -0.030 (G1), ns; RBM significant regardless of project type                                                       |
| <b>H5</b> | PDIA effects are stronger for intangible projects than tangible ones                                                          | Larger + for intangible | Not supported   | PDIA $\times$ Tangible = 0.069 (G2), ns; PDIA insignificant across all models                                                             |
| <b>H6</b> | RBM predictive power increased post-2000; specificity penalty deepened as boilerplate spread                                  | Increasing over time    | Partial support | RBM temporal trend unidentified due to collinearity (E1); specificity more predictive in recent decades (E3: 2020s $\beta = 0.417^{**}$ ) |

Note:  $** p < 0.05$ ;  $*** p < 0.01$ . Model labels refer to regression specifications described in Chapter 5. Verdict shading: green = supported, red = not supported, yellow = partial support.

H1 is supported: results-based management infrastructure at appraisal is a positive and significant predictor of IEG outcomes across all main specifications. H2 is not supported: PDIA-style adaptive design language shows no significant relationship with outcomes in any model. H3 is not supported in its aggregate form, though the Governance sector interaction in Model D3 provides conditional support for the role of context-specific language in complex institutional settings. H4 and H5 are not supported: tangibility moderates baseline outcome levels but not the design-quality relationship. H6 finds partial support for the specificity dimension, with context-specific language becoming more predictive in recent decades, but the RBM temporal trend cannot be identified precisely due to collinearity.

## 8. Discussion

### ***8.1. RBM Works - But Not for the Reasons the Reform Assumed***

This section is aimed at asking what the RBM finding actually means, and where a closer reading of the evidence complicates the standard pro-reform narrative. The finding from the regression analysis that stronger results infrastructure at appraisal predicts better IEG outcomes appears to support the World Bank's shift to mandatory Results Frameworks. Research shows that embedding baseline data, milestones, and formal M&E systems into project designs gives managers necessary tools which allow them to detect problems early and intervene before they compound (Kusek and Rist, 2004). The data are consistent with this logic.

However, the document-era result does not necessarily sit comfortably next to this. The PAD-era projects, those produced under the very regime that made results frameworks mandatory, score systematically lower on IEG outcomes than their SAR-era predecessors, net of decade and sector controls. Holzapfel (2016) in her comparative review of donor results-management systems has identified that standardized indicators tend to crowd out medium-term outcomes and skew the portfolio toward what is easily counted. What this means is that the RBM signal captured in this thesis reflects genuine results infrastructure, while the document-era penalty reflects portfolio-wide dilution as compliance spreads. These, however, are not contradictory since some projects invest in baseline measurement and monitoring design whereas others procure the same linguistic surface without actually having the underlying analytical work (Natsios, 2010).

The practical implication of this finding is straightforward. Appraisal quality review should weight indicator quality over template completeness. It can be argued that a results framework with three credible, variable indicators is more informative than one with fifteen indicators that cannot be measured. The OECD's (2019) synthesis of donor RBM evaluations reaches a similar conclusion: the agencies that see the strongest performance gains from RBM adoption are those that invest in indicator quality rather than indicator volume.

### ***8.2. PDIA as a Management Practice, Not a Design Feature***

The null PDIA finding across all sixteen models does not refute the adaptive management theory. Instead, it suggests where PDIA operates and what kind of test can detect it. Honig

(2018) highlights that managerial autonomy during implementation predicts better outcomes in complex projects. But this is based on the assumption how a project is managed, not how it was written. Given this, it is believed that a task team that genuinely intends to iterate adaptively has little incentive to signal this in the PAD, which is reviewed for format compliance rather than adaptive intent. Hence, it can be argued that the null result is in line with the PDIA literature which states that adaptive management may leave no detectable trace at the design stage.

There is wider evidence on thinking and working politically which reinforces this interpretation. Laws and Marquette (2018), reviewed a decade of thinking and working politically integration in development outcomes and concluded that the strongest confirmation for adaptive management emerges during implementation rather than design. Hadley and Tilley (2023), in their empirical study of PDIA in public financial management reform across six African countries, find that PDIA delivers short-term gains which are tied to authoring environments. This finding once again hints to the notion that it is about the implementation conditions not the appraisal language. The design document, in this reading, is the wrong place to look for the PDIA signal.

The direct practical implications of this would be that if the Bank wants to promote adaptive management, then they should leverage their attention to the supervision and restructuring architecture, not in the PAD. Allowing the teams to revise targets mid-term without penalty, and treating restructuring as evidence of learning rather than failure, would operationalise PDIA into practice in a way that would mandate adaptive language in appraisal documents.

### ***8.3. Boilerplate as a Governance-Specific Problem***

Similarly, for the specificity score, the aggregate null is not a straightforward negative result. It is a finding which pinpoints where generic document language is and is not costly. In infrastructure sectors, physical delivery is sufficiently standardized that template-driven planning rarely comprises implementation. For example, a road can be built to specification regardless of whether the appraisal document used context-specific analysis or generic donor vocabulary. The causal change, in these scenarios, is short enough for the document to be read more as a procurement specification rather than a theory of change.

However in the case of the Governance sector, the picture is different. The Governance × Specificity interaction is large and highly significant, indicating that projects in this sector that use more context-specific language perform substantially better. This can be deemed theoretically coherent. Governance interventions, such as, public financial management reform, decentralisation, civil service restructuring, rely entirely on accurate institutional diagnosis. A planning document which includes agencies, specific political economy constraints, and empirically grounded baseline conditions shows that the task team was aware of the environment they were intervening in. appraisal that rely on generic language such ‘institutional strengthening’ and ‘good governance’ signals the opposite (Andrews, 2013). A recent text analysis by Goldemberg, Jordan and Kenyon (2025), using machine learning embeddings of World Bank project documents, finds that the strongest predictor of projects' contributions to development outcomes is their degree of adaptation to country context rather than donor self-evaluation ratings, providing independent empirical support for the specificity finding in this thesis.

Practical implication of this finding is that boilerplate is not uniformly costly, it is especially costly in sectors where the causal chain runs through political and institutional dynamics that cannot be anticipated generically. Planning document review processes should demand for better specific standards to governance and institutional reform operations than to infrastructure ones, and this differentiation should be made more explicit in review criteria.

#### ***8.4. The Tangibility Null - Rethinking What RBM Measures***

The tangibility conditioning included in the thesis is intuitive and theoretically grounded, but the data rejects it. The tangibility factor of a project shapes baseline difficulty, and intangible projects are structurally harder but this does not moderate the relationship between planning quality and outcomes. RBM predicts better outcomes regardless of whether a project is building roads or reforming institutions.

This null finding invites a reversion to the standard account of how RBM adds value. The conventional argument puts forth the idea that the results infrastructure matters most where the outputs are physically countable, since that is where the targets are meaningful and monitoring systems can detect underperformance. But the data shows something different here. It shows that RBMs value may lie primarily in the accountability structures it creates

rather in the measurability of specific outputs. What baseline data does is that it forces teams to confront the gap between stated intent and initial conditions. Additionally, time-bound milestones create pressure for early problem identification. Procedures like formal verification makes it harder to claim success on activities alone. These mechanisms are valuable across sector types. Bulman, Kolkman and Kraay (2017), extending Denizer, Kaufmann and Kraay (2013) to the Asian Development Bank, find that within-country variation explains most of the outcome variance and that project-level management quality is a significant moderator. This is consistent with the view that accountability structures matter independently of the sector.

The practical implication of these findings goes against a common pattern in Bank portfolios, where infrastructure teams invest heavily in monitoring architecture while governance teams invest primarily in technical assistance and policy dialogue. If accountability structures drive outcomes, then the governance portfolio is where investment in genuine results infrastructure would have the largest marginal return.

### ***8.5. Design as a Necessary but Insufficient Condition***

The overall pattern of results positions project design quality as a modest, consistent, and robust contributor to development outcomes. But, it operates within a structural context it cannot override. The adjusted  $R^2$  values across the main models sit around 0.06, meaning that the full set of fixed effects and theory scores together explain roughly six per cent of outcome variance. The remainder can be attributed to the implementation environment, such as the country's institutional capacity, political economy, supervision quality, and the contingent factors that determine whether a project's theory of change holds in practice (Denizer, Kaufmann and Kraay, 2013; Eilers et al., 2026).

This is not an argument against design reform. Kilby (2015) shows that the project preparation on outcomes is conditional on economic vulnerability, meaning that counties under stress benefit more from careful preparation. Heinzl and Liese (2021) show that task team leader quality and country knowledge also shapes the recipient performance, independently of the country-level institutions, suggesting that the Bank-side inputs to planning genuinely matter. Together these findings indicate that the design quality is a necessary condition for good outcomes even if they are not sufficient, they consistently raise

the probability of success within a given institutional context, even when it cannot overcome severe structural disadvantages.

An implication for the Bank could be how to communicate the value of design reform internally and externally. Better project design will not transform the outcomes in the most difficult operating environments, and promising the opposite sets up the reform efforts to be dismissed when the results are disappointing. What it can do and what this thesis shows is the shift of the distribution of outcomes upward across the portfolio in a consistent and replicable manner.

## **9. Conclusion**

The focal point of this thesis was to assess whether the effort that donors put in project planning actually matters for development outcomes. Across 1,110 World Bank operations in South Asia spanning across six decades, it can be concluded that the answer is a qualified yes. Results-based management infrastructure at appraisal is a consistent and significant predictor of better IEG outcomes, which holds true across sectors, project types, and estimation strategies. It is the only one of the three theoretical frameworks that survives the horse race design, and it can be argued that its predictive power is not confined to physically tractable projects. This suggests that the accountability structures created by results infrastructure carry value across the portfolio independent of output measurability (Bulman, Kolkman and Kraay, 2017). This interpretation also implies that the value of the results framework lies less in their technical function and more in the organizational discipline that they impose. It forces teams to confront the gap between the stated intent and the actual conditions before the implementation begins (Mayne, 2007). Specifying the context only matters for projects in the Governance sectors where local institutional diagnosis is the intervention itself (Andrews, 2013). This finding raises the possibility that as governance operations become an increasingly large share of the global development portfolio, the cost of generic boilerplate planning may also grow considerably, which is a dynamic worth monitoring as the Bank continues to expand its institutional reform lending (World Bank IEG, 2020). Adaptive design language, as captured in appraisal documents, does not predict anything. This is not because the PDIA argument is wrong, it is because adaptive management is a practice of implementation, not a feature of documentation (Honig, 2018; Laws and Marquette, 2018). If further research can confirm this finding using

implementation-stage documents, it would have significant implications for how donors structure supervision and restructuring processes. This would suggest that flexibility should be built into implementation architecture rather than being signalled at appraisal (Andrews, Pritchett and Woolcock, 2017).

Given that the adjusted  $R^2$  values across all models sit at around six per cent, locates this finding precisely within the broader aid effectiveness literature, where, country institutions, political economy, and implementation conditions remain the dominant determinants of project success (Denizer, Kaufmann and Kraay, 2013; Easterly, 2006). The quality of the design raises the probability of success within a given context and it cannot substitute for context itself (Kilby, 2015). The persistence of the PAD-era outcome penalty is a reminder that mandatory compliance and genuine planning quality are not the same thing, and that the spread of results framework has not automatically produced the analytical rigour that those templates were designed to generate (Pritchett, Woolcock and Andrews, 2013; Holzapfel, 2016).

At a broader level, this thesis addresses a fundamental question in development economics: whether direct public investment can systematically accelerate economic development, and under what conditions the institutional apparatus surrounding that investment adds or destroys value (Sachs, 2005; Collier, 2007). South Asia's development trajectory over the study period includes rapid growth accompanied by persistent structural poverty, divergent institutional capacities across eight countries, which makes it an especially revealing setting to ask that question (World Bank, 2022; UNDP, 2016). The inference from the thesis that planning quality matters, modestly but robustly, is overall an argument for treating appraisal documents as a first-order of concern in development policy rather than an administrative formality. More speculatively, if the observed relationship between planning quality and outcomes holds across other regional portfolios and lending instruments, it would suggest that the hundreds and millions that donors invest annually in project preparations are not merely a transaction cost of aid delivery, but genuine attempts at improving development effectiveness; one that has been systematically undervalued in the macro aid effectiveness debate (Savedoff, Levine and Birdsall, 2005).

### ***9.1. Future Research***

Three extensions would most productively build on this thesis. First, validating the keyword scores against expert-coded quality rating from a subsample of the planning documents would strengthen the construct validity, building on recent text-analysis approaches in the literature (Goldemberg, Jordan and Kenyon, 2025). Second, applying the same methodology to another type of planning document, such as the Implementation Status Reports or mid-term review documents rather than appraisal documents alone, would allow a direct test of whether PDIA operates during implementation rather than the design stage, like the null finding here suffers to and is implied in Honig (2018). Third, an application of the analysis across the Sub-Saharan Africa portfolio, where governance constraints tend to be more severe and sector compositions differ substantially, would help in establishing whether the RBM and specificity findings can be generalized beyond South Asia or whether they are features of a particular regional lending context.

This thesis began with a tension: donors invest heavily in project planning on the premise that better-designed projects produce better outcomes, yet roughly one in three World Bank investment projects in South Asia has been rated Moderately Unsatisfactory or worse, persistently across decades of design reform. The empirical record presented here offers a qualified answer to that tension. Planning quality does matter, modestly and consistently, but it operates within structural constraints it cannot override. Results infrastructure raises the probability of success within a given context; it does not determine it. The democratic accountability argument raised at the outset, that public money spent on planning processes should demonstrably improve outcomes, finds cautious support here, but with an important caveat. What matters is not the presence of a results framework but the quality of the analytical work behind it. Mandatory compliance and genuine planning quality are not the same thing, and the spread of standardised templates has not automatically produced the rigour those templates were designed to generate. If there is a single practical message from this thesis, it is to treat appraisal documents as a genuine planning instrument rather than an administrative formality, to invest in indicator quality over indicator volume, and to recognise that for governance operations in particular, the cost of generic boilerplate is not merely bureaucratic. It predicts worse outcomes.

# References

- Andrews, M. (2013) *The Limits of Institutional Reform in Development: Changing Rules for Realistic Solutions*. Cambridge: Cambridge University Press.
- Andrews, M., Pritchett, L. and Woolcock, M. (2017) *Building State Capability: Evidence, Analysis, Action*. Oxford: Oxford University Press.
- Bevan, G. and Hood, C. (2006) 'What's measured is what matters: Targets and gaming in the English public health care system', *Public Administration*, 84(3), pp. 517-538.
- Birdsall, N. and Savedoff, W. (2011) *Cash on Delivery: A New Approach to Foreign Aid*. Washington, DC: Center for Global Development.
- Briggs, R.C. (2017) 'Does foreign aid target the poorest?', *International Organization*, 71(1), pp. 187–206.
- Bulman, D., Kolkman, W. and Kraay, A. (2017) 'Good countries or good projects? Comparing macro and micro correlates of World Bank and Asian Development Bank project performance', *Review of International Organizations*, 12(3), pp. 335-363.
- Burnside, C. and Dollar, D. (2000) 'Aid, policies, and growth', *American Economic Review*, 90(4), pp. 847-868.
- Carlsson, J., Köhlin, G. and Ekbom, A. (2016) *The Political Economy of Evaluation: International Aid Agencies and the Effectiveness of Aid*. Springer.
- Chauvet, L. and Collier, P. (2008) 'What are the preconditions for turnarounds in failing states?', *Conflict Management and Peace Science*, 25(4), pp. 332-348.
- Clemens, M.A., Radelet, S., Bhavnani, R.R. and Bazzi, S. (2012) 'Counting chickens when they hatch: Timing and the effects of aid on growth', *Economic Journal*, 122(561), pp. 590-617.
- Collier, P. (2007) *The Bottom Billion: Why the Poorest Countries Are Failing and What Can Be Done About It*. Oxford: Oxford University Press.
- Cracknell, B.E. (2000) *Evaluating Development Aid: Issues, Problems and Solutions*. London: Sage Publications.
- Deaton, A. (2010) 'Instruments, randomization, and learning about development', *Journal of Economic Literature*, 48(2), pp. 424-455.
- Denizer, C., Kaufmann, D. and Kraay, A. (2013) 'Good countries or good projects? Macro and micro correlates of World Bank project performance', *Journal of Development Economics*, 105, pp. 288-302.

DiMaggio, P. and Powell, W.W. (1983) 'The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields', *American Sociological Review*, 48(2), pp. 147-160.

Easterly, W. (2006) *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*. New York: Penguin Press.

Easterly, W. (2014) *The Tyranny of Experts: Economists, Dictators, and the Forgotten Rights of the Poor*. New York: Basic Books.

Easterly, W., Levine, R. and Roodman, D. (2004) 'Aid, policies, and growth: Comment', *American Economic Review*, 94(3), pp. 774-780.

Eilers, Y., Kluve, J., Langbein, J. and Reiners, L. (2026) 'Volume, risk, complexity: what makes development finance projects succeed or fail?', *World Bank Economic Review*, 40(1), pp. 172-195.

Geli, P., Kraay, A. and Nobakht, H. (2014) *Predicting World Bank Project Outcome Ratings*. World Bank Policy Research Working Paper 7001. Washington, DC: World Bank.

Gentzkow, M., Kelly, B. and Taddy, M. (2019) 'Text as data', *Journal of Economic Literature*, 57(3), pp. 535-574.

Goldemberg, D., Jordan, L. and Kenyon, T. (2025) 'Minding the gap: Aid effectiveness, project ratings and contextualization', *The World Bank Economic Review*, advance article. doi: 10.1093/wber/lhaf005.

Grimmer, J. and Stewart, B.M. (2013) 'Text as data: The promise and pitfalls of automatic content analysis methods for political texts', *Political Analysis*, 21(3), pp. 267-297.

Grimmer, J., Roberts, M.E. and Stewart, B.M. (2022) *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton: Princeton University Press.

Hadley, S. and Tilley, H. (2023) 'Is the problem driven iterative adaptation approach (PDIA) a panacea for public financial management reform? Evidence from six African countries', *World Development Perspectives*, 31, p. 100525.

Heinzel, M. and Liese, A. (2021) 'Managing performance and winning trust: how World Bank staff shape recipient performance', *Review of International Organizations*, 16(3), pp. 625-653.

Holzapfel, S. (2016) 'Boosting or hindering aid effectiveness? An assessment of systems for measuring donor agency results', *Public Administration and Development*, 36(1), pp. 3-19.

Honig, D. (2018) *Navigation by Judgment: Why and When Top-Down Management of Foreign Aid Does Not Work*. Oxford: Oxford University Press.

Hood, C. (1991) 'A public management for all seasons?', *Public Administration*, 69(1), pp. 3-19.

World Bank Independent Evaluation Group (IEG) (2008) *Public Sector Reform: What Works and Why? An IEG Evaluation of World Bank Support*. Washington, DC: World Bank.

World Bank Independent Evaluation Group (2012) *Designing a Results Framework for Achieving Results: A How-to Guide*. Washington, DC: World Bank.

World Bank Independent Evaluation Group (2020) *Results and Performance of the World Bank Group 2020*. Washington, DC: World Bank. Available at: <https://ieg.worldbankgroup.org/evaluations/rap2020> (Accessed 15 March 2026).

World Bank Independent Evaluation Group (2023) *Results and Performance of the World Bank Group 2023: An Independent Evaluation*. Washington, DC: World Bank. Available at: <https://ieg.worldbankgroup.org/evaluations/results-and-performance-world-bank-group-2023> (Accessed 17 March 2026).

World Bank Independent Evaluation Group (2025) *Results and Performance of the World Bank Group 2025: Concept Note*. Washington, DC: World Bank. Available at: [https://ieg.worldbankgroup.org/sites/default/files/Data/reports/report-Concept\\_Note-RAP\\_2025.pdf](https://ieg.worldbankgroup.org/sites/default/files/Data/reports/report-Concept_Note-RAP_2025.pdf) (Accessed March 21 2026).

Isham, J. and Kaufmann, D. (1999) 'The forgotten rationale for policy reform: The productivity of investment projects', *Quarterly Journal of Economics*, 114(1), pp. 149-184.

Kilby, C. (2015) 'Assessing the impact of World Bank preparation on project outcomes', *Journal of Development Economics*, 115, pp. 111-123.

Krippendorff, K. (2018) *Content Analysis: An Introduction to Its Methodology*. 2nd edn. Thousand Oaks: Sage Publications.

Kusek, J.Z. and Rist, R.C. (2004) *Ten Steps to a Results-Based Monitoring and Evaluation System: A Handbook for Development Practitioners*. Washington, DC: World Bank.

Laws, E. and Marquette, H. (2018) *Thinking and Working Politically: Reviewing the Evidence on the Integration of Politics into Development Practice over the Past Decade*. Birmingham: University of Birmingham.

Lombard, M., Snyder-Duch, J. and Bracken, C.C. (2022) 'Content analysis in mass communication: Assessment and reporting of intercoder reliability', *Human Communication Research*, 28(4), pp. 587-604.

Mayne, J. (2007) 'Challenges and lessons in implementing results-based management', *Evaluation*, 13(1), pp. 87-109.

Michaelowa, K. and Borrmann, A. (2006) 'Evaluation bias and incentive structures in bi- and multilateral aid agencies', *Review of Development Economics*, 10(2), pp. 313-329.

Mintzberg, H. (1994) *The Rise and Fall of Strategic Planning*. New York: Free Press.

Natsios, A. (2010) *The Clash of the Counter-Bureaucracy and Development*. Washington, DC: Center for Global Development.

OECD DAC (2010) *Quality Standards for Development Evaluation*, DAC Guidelines and Reference Series. Paris: OECD Publishing. doi: 10.1787/9789264083905-en

OECD (2019) *Managing for Sustainable Development Results: Guiding Principles*. Paris: OECD Publishing. doi: 10.1787/44a288bc-en

Pollitt, C. (1993) *Managerialism and the Public Services: Cuts or Cultural Change in the 1990s?* 2nd edn. Oxford: Blackwell.

Power, M. (1997) *The Audit Society: Rituals of Verification*. Oxford: Oxford University Press.

Pressman, J.L. and Wildavsky, A. (1984) *Implementation: How Great Expectations in Washington Are Dashed in Oakland*. Berkeley: University of California Press.

Pritchett, L. and Sandefur, J. (2015) 'Learning from experiments when context matters', *American Economic Review: Papers and Proceedings*, 105(5), pp. 471-475.

Pritchett, L. and Woolcock, M. (2004) 'Solutions when the solution is the problem: Arraying the disarray in development', *World Development*, 32(2), pp. 191-212.

Pritchett, L., Woolcock, M. and Andrews, M. (2013) 'Looking like a state: Techniques of persistent failure in state capability for implementation', *Journal of Development Studies*, 49(1), pp. 1-18.

Rajan, R.G. and Subramanian, A. (2008) 'Aid and growth: What does the cross-country evidence really show?', *Review of Economics and Statistics*, 90(4), pp. 643-665.

Ramalingam, B. (2013) *Aid on the Edge of Chaos: Rethinking International Cooperation in a Complex World*. Oxford: Oxford University Press.

Rogers, P. (2008) 'Using programme theory to evaluate complicated and complex aspects of interventions', *Evaluation*, 14(1), pp. 29-48.

Roodman, D. (2007) 'The anarchy of numbers: Aid, development, and cross-country empirics', *World Bank Economic Review*, 21(2), pp. 255-277.

Sachs, J. (2005) *The End of Poverty: Economic Possibilities for Our Time*. New York: Penguin Press.

Savedoff, W., Levine, R. and Birdsall, N. (2005) *When Will We Ever Learn? Improving Lives through Impact Evaluation*. Washington, DC: Center for Global Development.

Snowden, D.J. and Boone, M.E. (2007) 'A leader's framework for decision making', *Harvard Business Review*, 85(11), pp. 68-76.

UNDP (2016) *Human Development Report 2016: Human Development for Everyone*. New York: United Nations Development Programme. Available at: <https://hdr.undp.org/content/human-development-report-2016> (Accessed 14 April 2026).

Valters, C. and Whitty, B. (2017) *The Politics of the Results Agenda in DFID: 1997-2017*. London: Overseas Development Institute. Available at: <https://odi.org/en/publications/the-politics-of-the-results-agenda-in-dfid-1997-2017/>

Weiss, C.H. (1997) 'Theory-based evaluation: Past, present, and future', *New Directions for Evaluation*, 76, pp. 41-55.

Winters, M.S. (2010) 'Accountability, participation and foreign aid effectiveness', *International Studies Review*, 12(2), pp. 218-243.

World Bank (1992) *Effective Implementation: Key to Development Impact. Report of the Portfolio Management Task Force*. Washington, DC: World Bank.

World Bank (2022) *Reshaping Norms: A New Way Forward*. South Asia Economic Focus, Spring 2022. Washington, DC: World Bank. doi: 10.1596/978-1-4648-1857-8

World Bank (2026a) *IEG Project Performance Ratings Database*. Washington, DC: World Bank. Available at: <https://financesone.worldbank.org/ieg-world-bank-project-performance-ratings/DS00053> [Accessed 27 March 2026].

World Bank (2026b) *Documents and Reports: Implementation Completion and Results Reports, South Asia*. Washington, DC: World Bank. Available at: [https://documents.worldbank.org/en/publication/documents-reports/documentlist?countcode=AF%5EPK%5EIN%5ENP%5EBT%5EBD%5ELK%5EMV&srt=docdt&order=asc&strdate=01-01-2000&enddate=03-03-2026&docty\\_exact=Implementation+Completion+and+Results+Report](https://documents.worldbank.org/en/publication/documents-reports/documentlist?countcode=AF%5EPK%5EIN%5ENP%5EBT%5EBD%5ELK%5EMV&srt=docdt&order=asc&strdate=01-01-2000&enddate=03-03-2026&docty_exact=Implementation+Completion+and+Results+Report) [Accessed 27 March 2026].

# Appendix A

## *AI Contribution Statement*

### **1. Tools**

Claude (Anthropic) was the only AI-based tool used in the completion of this thesis.

### **2. Degree of Use**

Research and idea generation: Claude was used to discuss and refine the theoretical framing of the three competing frameworks, Results-Based Management, Problem-Driven Iterative Adaptation, and the isomorphic mimicry hypothesis, and to identify relevant literature and sources across all chapters. All analytical decisions, theoretical arguments, and interpretations of findings are my own.

Coding and data analysis: Claude assisted with writing and debugging the R code used for PDF text extraction, keyword scoring, sigmoid normalisation, and regression analysis. The research design, scoring dictionaries, variable definitions, and interpretation of all regression outputs were developed and conducted independently by me.

Language and editing: Claude was used to proofread and improve the clarity and flow of written sections throughout the thesis. In all cases, the substance, structure, and arguments of the text are my own, and all writing was reviewed and revised by me before inclusion in the thesis.

Formal aspects: Claude was used to assist with formatting references in the bibliography and checking citation consistency throughout the thesis.

All information generated or suggested by Claude was treated with critical caution and independently verified against primary sources before being used in the thesis. The core intellectual contributions of this thesis, the research design, the theoretical horse-race framework, the construction of the scoring dictionaries, the regression analysis, and the interpretation of findings, are entirely my own.

## Appendix B

**Table 7. Scoring Dictionary For Dimension 1: Results-Based Management (RBM)**

| Category                             | Weight        | Terms / Phrases Used                                                                                                                       | What This Captures                                                                                                            |
|--------------------------------------|---------------|--------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|
| Quantified targets and outputs       | 1.5 per match | Numbers combined with units: %, percent, million, thousand, household, beneficiary-, student, patient, farmer, km, ha, mw, unit, number of | Presence of quantified objectives or outputs                                                                                  |
| Baseline language                    | 2 per match   | baseline                                                                                                                                   | Explicit reference to starting conditions                                                                                     |
| Strong baseline phrasing             | 3 per match   | baseline of, baseline data, baseline value, baseline survey, baseline level, from a baseline, against a baseline                           | Stronger evidence that baseline measurement is built into design                                                              |
| Target language                      | 1 per match   | target                                                                                                                                     | General reference to intended results                                                                                         |
| Strong target phrasing               | 2 per match   | target of [number], target: [number], target value, endline target                                                                         | More concrete specification of target values                                                                                  |
| Time-bound commitments               | 1 per match   | by [year], by year [x], by end of project, by project completion, annually, by mid-term                                                    | Whether targets are linked to a time horizon                                                                                  |
| Monitoring and evaluation systems    | 4 (binary)    | monitoring and evaluation, M&E framework, results framework, monitoring framework                                                          | Formal mechanisms for tracking implementation and results; highest weight reflecting diagnostic value of genuine M&E adoption |
| Verification and measurement sources | 1.5 per match | means of verification, method of measurement, data source, data collection, verified by, survey data, administrative record                | Whether indicators are linked to evidence sources or measurement methods                                                      |
| Results-chain terminology            | 3 (binary)    | PDO-level, outcome indicator, output indicator, intermediate result, results chain, results framework                                      | Structured results logic connecting activities to outputs and outcomes                                                        |

## Appendix C

**Table 8. Scoring Dictionary For Dimension 2: Problem-Driven Iterative Adaptation (PDIA)**

| Category                                 | Weight        | Terms / Phrases Used                                                                                                                                                                 | What This Captures                                                                                                                                                       |
|------------------------------------------|---------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Uncertainty and complexity               | 0.5 per match | uncertain, unpredictable-, complex, challeng-, difficult to predict, context-depend-, contingent on                                                                                  | Recognition that implementation conditions are uncertain or complex; low weight because such terms appear in routine risk sections regardless of genuine adaptive intent |
| Adaptation and flexibility               | 2 per match   | adaptive management, adaptation, flexibility, flexible implementation, adjustment, adjusting the design, revise, course correction, learning loop, iterative                         | Whether the project design allows revision and adjustment over time                                                                                                      |
| Piloting and phased implementation       | 2 per match   | phase 1/2/3, phase one/two/three, phase I/II/III, pilot phase, pilot program, pilot project, pilot test, phased approach, phased implementation, pilot and scale, test and learn     | Use of experimentation, sequencing, or staged rollout                                                                                                                    |
| Contingency planning                     | 1.5 per match | contingency, contingent, if the project, in the event, should the project, alternative approach, fallback, exit strategy                                                             | Whether alternatives are built in if implementation conditions change                                                                                                    |
| Learning and feedback                    | 1.5 per match | lesson learned, lessons learned, learning from, learning agenda, learning review, mid-term review, course correction, feedback loop, feedback mechanism, review and adjust           | Mechanisms for reflection, review, and adaptive learning                                                                                                                 |
| Political economy constraints            | 2 per match   | political economy, political will, political commitment, political constraint, political risk, vested interest, stakeholder power, resistance, opposition, political feasibility     | Recognition that reform is shaped by politics and power; most distinctively PDIA-aligned element                                                                         |
| Local ownership and contextual grounding | 1 per match   | local context, local knowledge, local condition, local capacity, local need, local partner, local solution, country-driven, demand-driven, locally led, locally owned, community-led | Sensitivity to local context and emphasis on local agency                                                                                                                |

## Appendix D

**Table 9. Scoring Dictionary For Dimension 3: Context Specificity**

| Signal Type                              | Direction    | Terms / Phrases Used                                                                                                                                                                                                                                                                                                                    | What This Captures                                                                                                 |
|------------------------------------------|--------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------|
| Specific institutions and agencies       | Specific (+) | ministry of, department of, directorate of, bureau of, authority of, board of, national commission, national agency, national authority, national institute, national council, national bank, national fund, provincial, district office, district government, district authority, municipality, panchayat, upazila, tehsil, gram sabha | Concrete identification of implementing institutions or governance structures                                      |
| Specific places and administrative areas | Specific (+) | district(s) of, province of, region of, state of, division of, upazila, thana, taluk, mandal, ward, village of, town of                                                                                                                                                                                                                 | Geographic specificity                                                                                             |
| Empirical and statistical references     | Specific (+) | according to the [year], [year] census, [year] survey, [year] data, [year] report, [year] assessment, national survey, national data, national statistic-, household survey, [number] percent of the population / households / farmers / women / children                                                                               | Use of data, surveys, and population-based evidence                                                                |
| Specific beneficiary groups              | Specific (+) | smallholder farmer, marginal farmer, landless, ultra-poor, female-headed household, tribal, scheduled caste, scheduled tribe, indigenous people, indigenous community, pastoralist, artisan, fisherfolk                                                                                                                                 | Clearly identified target groups rather than broad categories                                                      |
| Generic development language             | Generic (–)  | best practice, international standard, global experience, lessons from, in line with, consistent with, aligned with, in accordance with, world class, state of the art, capacity building, institutional strengthening, good governance, gender mainstreaming, inclusive growth, holistic approach, integrated approach                 | Boilerplate or broad policy language with limited contextual specificity; negative signal reducing the ratio score |

## Appendix E

**Table 10. Sector Classification by Tangibility**

| Category                           | Sectors                                                                                                 | Rationale                                                                                                       |
|------------------------------------|---------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| Tangible (1)<br><i>n</i> = 578     | Energy & Extractives, Transport, Water,<br>Agriculture and Food, Digital<br>Development                 | Physical infrastructure or directly<br>measurable service delivery; short causal<br>chain from input to outcome |
| Intangible (0)<br><i>n</i> = 226   | Governance, Finance & Competitiveness,<br>Macroeconomics & Trade, Social Policy                         | Institutional change, policy reform, or<br>behaviour change; long causal chain<br>dependent on political will   |
| Mixed (baseline)<br><i>n</i> = 306 | Education, Health, Urban & Resilience,<br>Environment & NRM, Social<br>Sustainability, Poverty & Equity | Combines physical delivery and<br>institutional elements; retained as separate<br>category in regression        |

# Appendix F

## Econometric Model Specifications

### Models A1-A3: Isolation Models

$$Y_i = \alpha + \beta_1 \text{RBM}_i + \gamma \text{Country}_i + \delta \text{Sector}_i + \theta \text{Decade}_i + \lambda \text{PAD}_i + \varepsilon_i \quad (\text{A1})$$

$$Y_i = \alpha + \beta_1 \text{PDIA}_i + \gamma \text{Country}_i + \delta \text{Sector}_i + \theta \text{Decade}_i + \lambda \text{PAD}_i + \varepsilon_i \quad (\text{A2})$$

$$Y_i = \alpha + \beta_1 \text{Spec}_i + \gamma \text{Country}_i + \delta \text{Sector}_i + \theta \text{Decade}_i + \lambda \text{PAD}_i + \varepsilon_i \quad (\text{A3})$$

### Model B: OLS Horse Race

$$Y_i = \alpha + \beta_1 \text{RBM}_i + \beta_2 \text{PDIA}_i + \beta_3 \text{Spec}_i + \gamma \text{Country}_i + \delta \text{Sector}_i + \theta \text{Decade}_i + \lambda \text{PAD}_i + \varepsilon_i \quad (\text{B})$$

### Model C: Ordinal Logistic Horse Race

$$\log P(Y_i \leq k) / P(Y_i > k) = \alpha_k - (\beta_1 \text{RBM}_i + \beta_2 \text{PDIA}_i + \beta_3 \text{Spec}_i + \gamma \text{Country}_i + \delta \text{Sector}_i + \theta \text{Decade}_i + \lambda \text{PAD}_i) \quad (\text{C})$$

### Models D1-D3: Theory × Sector Interactions

$$Y_i = \alpha + \beta_1 \text{RBM}_i + \sum \beta_2 (\text{RBM}_i \times \text{Sector}_i) + \gamma \text{Country}_i + \delta \text{Sector}_i + \theta \text{Decade}_i + \lambda \text{PAD}_i + \varepsilon_i \quad (\text{D1})$$

$$Y_i = \alpha + \beta_1 \text{PDIA}_i + \sum \beta_2 (\text{PDIA}_i \times \text{Sector}_i) + \gamma \text{Country}_i + \delta \text{Sector}_i + \theta \text{Decade}_i + \lambda \text{PAD}_i + \varepsilon_i \quad (\text{D2})$$

$$Y_i = \alpha + \beta_1 \text{Spec}_i + \sum \beta_2 (\text{Spec}_i \times \text{Sector}_i) + \gamma \text{Country}_i + \delta \text{Sector}_i + \theta \text{Decade}_i + \lambda \text{PAD}_i + \varepsilon_i \quad (\text{D3})$$

### Models E1-E3: Theory × Decade Interactions

$$Y_i = \alpha + \beta_1 \text{RBM}_i + \sum \beta_2 (\text{RBM}_i \times \text{Decade}_i) + \gamma \text{Country}_i + \delta \text{Sector}_i + \theta \text{Decade}_i + \lambda \text{PAD}_i + \varepsilon_i \quad (\text{E1})$$

$$Y_i = \alpha + \beta_1 \text{PDIA}_i + \sum \beta_2 (\text{PDIA}_i \times \text{Decade}_i) + \gamma \text{Country}_i + \delta \text{Sector}_i + \theta \text{Decade}_i + \lambda \text{PAD}_i + \varepsilon_i \quad (\text{E2})$$

$$Y_i = \alpha + \beta_1 \text{Spec}_i + \sum \beta_2 (\text{Spec}_i \times \text{Decade}_i) + \gamma \text{Country}_i + \delta \text{Sector}_i + \theta \text{Decade}_i + \lambda \text{PAD}_i + \varepsilon_i \quad (\text{E3})$$

### Model F: Tangibility Main Effect

$$Y_i = \alpha + \beta_1 \text{RBM}_i + \beta_2 \text{PDIA}_i + \beta_3 \text{Spec}_i + \beta_4 \text{Tangible}_i + \beta_5 \text{Intangible}_i + \gamma \text{Country}_i + \theta \text{Decade}_i + \lambda \text{PAD}_i + \varepsilon_i \quad (\text{F})$$

### Models G1-G4: Theory × Tangibility Interactions

$$Y_i = \alpha + \beta_1 \text{RBM}_i + \beta_2 \text{Tangible}_i + \beta_3 (\text{RBM}_i \times \text{Tangible}_i) + \gamma \text{Country}_i + \theta \text{Decade}_i + \lambda \text{PAD}_i + \varepsilon_i \quad (\text{G1})$$

$$Y_i = \alpha + \beta_1 \text{PDIA}_i + \beta_2 \text{Tangible}_i + \beta_3 (\text{PDIA}_i \times \text{Tangible}_i) + \gamma \text{Country}_i + \theta \text{Decade}_i + \lambda \text{PAD}_i + \varepsilon_i \quad (\text{G2})$$

$$Y_i = \alpha + \beta_1 \text{Spec}_i + \beta_2 \text{Tangible}_i + \beta_3 (\text{Spec}_i \times \text{Tangible}_i) + \gamma \text{Country}_i + \theta \text{Decade}_i + \lambda \text{PAD}_i + \varepsilon_i \quad (\text{G3})$$

$$Y_i = \alpha + \beta_1 \text{RBM}_i + \beta_2 \text{PDIA}_i + \beta_3 \text{Spec}_i + \beta_4 \text{Tangible}_i + \beta_5 (\text{RBM}_i \times \text{Tangible}_i) + \beta_6 (\text{PDIA}_i \times \text{Tangible}_i) + \beta_7 (\text{Spec}_i \times \text{Tangible}_i) + \gamma \text{Country}_i + \theta \text{Decade}_i + \lambda \text{PAD}_i + \varepsilon_i \quad (\text{G4})$$

Where:

$Y_i$  = IEG outcome rating for project  $i$  (1-6)

$\text{RBM}_i$  = standardised RBM score for project  $i$

$\text{PDIA}_i$  = standardised PDIA score for project  $i$

$\text{Spec}_i$  = standardised specificity score for project  $i$

$\text{Tangible}_i$  = binary indicator equal to 1 if project  $i$  is classified as tangible

$\text{Country}_i$  = vector of country fixed effects

$\text{Sector}_i$  = vector of sector fixed effects

$\text{Decade}_i$  = vector of approval decade fixed effects

$\text{PAD}_i$  = binary document-era indicator equal to 1 if project  $i$  is a post-1995 appraisal document

$\varepsilon_i$  = error term

$\alpha_k$  = threshold parameters for ordinal logistic model (Model C only)

# Appendix G

**Table 11. Theory × Sector Interactions, OLS (D1-D3)**

|                                 | IEG Outcome Score (1-6) |                       |                        |
|---------------------------------|-------------------------|-----------------------|------------------------|
|                                 | RBM × Sector<br>(D1)    | PDIA × Sector<br>(D2) | Spec. × Sector<br>(D3) |
| RBM score (std.)                | −0.079<br>(0.105)       |                       |                        |
| PDIA score (std.)               |                         | −0.011<br>(0.126)     |                        |
| Specificity score (std.)        |                         |                       | −0.034<br>(0.107)      |
| Doc. era: PAD (ref. SAR)        | −0.407**<br>(0.172)     | −0.336**<br>(0.169)   | −0.356**<br>(0.168)    |
| <i>RBM interactions</i>         |                         |                       |                        |
| RBM × Agriculture & Food        | 0.064<br>(0.142)        |                       |                        |
| RBM × Energy & Extractives      | 0.390**<br>(0.157)      |                       |                        |
| RBM × Environment & NRM         | 0.403<br>(0.268)        |                       |                        |
| RBM × Finance & Competitiveness | 0.297**<br>(0.137)      |                       |                        |
| RBM × Governance                | 0.423<br>(0.285)        |                       |                        |
| RBM × Health, Nutrition & Pop.  | 0.220<br>(0.185)        |                       |                        |
| RBM × Macroeconomics & Trade    | 0.096<br>(0.167)        |                       |                        |
| RBM × Social Policy             | −0.099<br>(0.259)       |                       |                        |
| RBM × Transport                 | 0.172<br>(0.154)        |                       |                        |
| RBM × Urban, Resilience & Land  | 0.051<br>(0.183)        |                       |                        |
| RBM × Water                     | 0.081<br>(0.165)        |                       |                        |
| RBM × Digital Development       | 0.304<br>(0.323)        |                       |                        |
| <i>PDIA interactions</i>        |                         |                       |                        |
| PDIA × Agriculture & Food       |                         | −0.034                |                        |

|                                         |          |  |
|-----------------------------------------|----------|--|
|                                         | (0.139)  |  |
| PDIA × Energy & Extractives             | 0.002    |  |
|                                         | (0.204)  |  |
| PDIA × Environment & NRM                | −0.032   |  |
|                                         | (0.154)  |  |
| PDIA × Finance & Competitiveness        | −0.444*  |  |
|                                         | (0.267)  |  |
| PDIA × Governance                       | 0.143    |  |
|                                         | (0.240)  |  |
| PDIA × Health, Nutrition & Pop.         | 0.029    |  |
|                                         | (0.166)  |  |
| PDIA × Macroeconomics & Trade           | −0.025   |  |
|                                         | (0.290)  |  |
| PDIA × Social Policy                    | 0.109    |  |
|                                         | (0.168)  |  |
| PDIA × Transport                        | 0.009    |  |
|                                         | (0.147)  |  |
| PDIA × Urban, Resilience & Land         | 0.143    |  |
|                                         | (0.152)  |  |
| PDIA × Water                            | 0.086    |  |
|                                         | (0.155)  |  |
| PDIA × Digital Development              | −0.187   |  |
|                                         | (0.412)  |  |
| <i>Specificity interactions</i>         |          |  |
| Specificity × Agriculture & Food        | −0.182   |  |
|                                         | (0.143)  |  |
| Specificity × Energy & Extractives      | −0.047   |  |
|                                         | (0.163)  |  |
| Specificity × Environment & NRM         | 0.022    |  |
|                                         | (0.187)  |  |
| Specificity × Finance & Competitiveness | 0.106    |  |
|                                         | (0.149)  |  |
| Specificity × Governance                | 0.557*** |  |
|                                         | (0.168)  |  |
| Specificity × Health, Nutrition & Pop.  | 0.274    |  |
|                                         | (0.347)  |  |
| Specificity × Macroeconomics & Trade    | −0.231   |  |

|                                           |       |       |         |
|-------------------------------------------|-------|-------|---------|
|                                           |       |       | (0.179) |
| Specificity × Social Policy               |       |       | 0.067   |
|                                           |       |       | (0.207) |
| Specificity × Transport                   |       |       | 0.177   |
|                                           |       |       | (0.136) |
| Specificity × Urban, Resilience<br>& Land |       |       | 0.149   |
|                                           |       |       | (0.162) |
| Specificity × Water                       |       |       | −0.015  |
|                                           |       |       | (0.151) |
| Specificity × Digital<br>Development      |       |       | 0.386   |
|                                           |       |       | (0.274) |
| Sector FE                                 | Yes   | Yes   | Yes     |
| Country FE                                | Yes   | Yes   | Yes     |
| Decade FE                                 | Yes   | Yes   | Yes     |
| Observations                              | 1,110 | 1,110 | 1,110   |
| R <sup>2</sup>                            | 0.100 | 0.091 | 0.100   |
| Adjusted R <sup>2</sup>                   | 0.064 | 0.055 | 0.063   |

*Note: Main effects of sector, country, and decade FE included but suppressed. Robust standard errors (HC1) in parentheses. Theory scores standardised (mean = 0, SD = 1). Reference: Education (sector), India (country), SAR (document era), 1950s (decade). \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .*

# Appendix H

**Table 12. Theory × Decade Interactions, OLS (E1-E3)**

|                                          | <i>IEG Outcome Score (1-6)</i> |                       |                        |
|------------------------------------------|--------------------------------|-----------------------|------------------------|
|                                          | RBM × Decade<br>(E1)           | PDIA × Decade<br>(E2) | Spec. × Decade<br>(E3) |
| RBM score (std.)                         | 0.306<br>(0.973)               | 0.095**<br>(0.041)    | 0.089**<br>(0.041)     |
| PDIA score (std.)                        | 0.003<br>(0.034)               | 1.159<br>(1.071)      | −0.006<br>(0.033)      |
| Specificity score (std.)                 | 0.002<br>(0.043)               | 0.007<br>(0.043)      | 0.417**<br>(0.187)     |
| Doc. era: PAD (ref. SAR)                 | −0.413**<br>(0.169)            | −0.443***<br>(0.171)  | −0.441***<br>(0.169)   |
| <i>RBM × Decade interactions</i>         |                                |                       |                        |
| RBM × 1960s                              | −0.096<br>(0.988)              |                       |                        |
| RBM × 1970s                              | −0.095<br>(0.977)              |                       |                        |
| RBM × 1980s                              | −0.270<br>(0.977)              |                       |                        |
| RBM × 1990s                              | −0.206<br>(0.976)              |                       |                        |
| RBM × 2000s                              | −0.183<br>(0.979)              |                       |                        |
| RBM × 2010s                              | −0.299<br>(0.977)              |                       |                        |
| RBM × 2020s                              | (ref.)                         |                       |                        |
| <i>PDIA × Decade interactions</i>        |                                |                       |                        |
| PDIA × 1960s                             |                                | −0.991<br>(1.075)     |                        |
| PDIA × 1970s                             |                                | −1.133<br>(1.084)     |                        |
| PDIA × 1980s                             |                                | −1.192<br>(1.078)     |                        |
| PDIA × 1990s                             |                                | −1.053<br>(1.075)     |                        |
| PDIA × 2000s                             |                                | −1.192<br>(1.072)     |                        |
| PDIA × 2010s                             |                                | −1.181<br>(1.071)     |                        |
| PDIA × 2020s                             |                                | (ref.)                |                        |
| <i>Specificity × Decade interactions</i> |                                |                       |                        |

|                         |       |       |           |
|-------------------------|-------|-------|-----------|
| Specificity × 1960s     |       |       | −0.534*   |
|                         |       |       | (0.300)   |
| Specificity × 1970s     |       |       | −0.326    |
|                         |       |       | (0.257)   |
| Specificity × 1980s     |       |       | −0.525**  |
|                         |       |       | (0.215)   |
| Specificity × 1990s     |       |       | −0.591*** |
|                         |       |       | (0.212)   |
| Specificity × 2000s     |       |       | −0.388**  |
|                         |       |       | (0.193)   |
| Specificity × 2010s     |       |       | −0.296    |
|                         |       |       | (0.199)   |
| Specificity × 2020s     |       |       | (ref.)    |
| Sector FE               | Yes   | Yes   | Yes       |
| Country FE              | Yes   | Yes   | Yes       |
| Decade FE               | Yes   | Yes   | Yes       |
| Observations            | 1,110 | 1,110 | 1,110     |
| R <sup>2</sup>          | 0.092 | 0.091 | 0.096     |
| Adjusted R <sup>2</sup> | 0.060 | 0.059 | 0.064     |

---

*Note: Decade main effects and other FE controls included but suppressed. 2020s is the omitted decade category (reference). Robust standard errors (HCl) in parentheses. Theory scores standardised (mean = 0, SD = 1). Reference: Education (sector), India (country), SAR (document era). \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$ .*