



PROCESSING LARGE AMOUNT OF TEXT DATA: A PRACTICAL EXAMPLE FROM A META-EVALUATION USING COMPUTATIONAL TEXT ANALYSIS

Oliver Serfling, Taya Romanov

Agenda

1

Introduction: The Research Project

2

Data & Pipeline

3

Beneficiary Section Extraction

4

Level Identification via CTA

5

Sample Construction Funnel

6

Key Findings

7

Reflection, Limitations & Takeaways

Q

Q&A

Relevance

Evaluation systems generate ‘mountains’ of textual data

project appraisal documents, implementation and completion reports, evaluation reports

This data is rarely analyzed systematically

Conventional evaluation and meta-evaluations struggle with large-n, text-heavy sources

Computational text analysis offers new opportunities to systematically analyze large document collections.

The webinar presents a practical, real-world example rather than a purely technical method demonstration.

Part 1

Introduction: The Research Project

Capacity Building for Sustainable Development

Identification of a common definition for CD

Theoretical foundations of CD

Quantitative Significance in Development Practice

(Meta-)Evaluation of CD projects

**WE
SHOW:**

- that the popularity of CD remains on a consistently high level
- that income level of the recipient country predicts the amount of CD funding
- CD projects have worse evaluation outcomes

CD → worse evaluation outcomes

Mean evaluation outcome

0.32 for CD

VS

0.43 for non-CD

Wilcoxon-Mann-Whitney test
p-value < .001

WB IEG evaluation outcomes

3 dimensions:

- ✓ Relevance
- ✓ Efficacy
- ✓ Effectiveness

1.5	Highly satisfactory
1	Satisfactory
0.5	Moderately satisfactory
- 0.5	Moderately unsatisfactory
- 1	Unsatisfactory
- 1.5	Highly unsatisfactory

(World Bank 2023, p.3)

The complex nature of CD

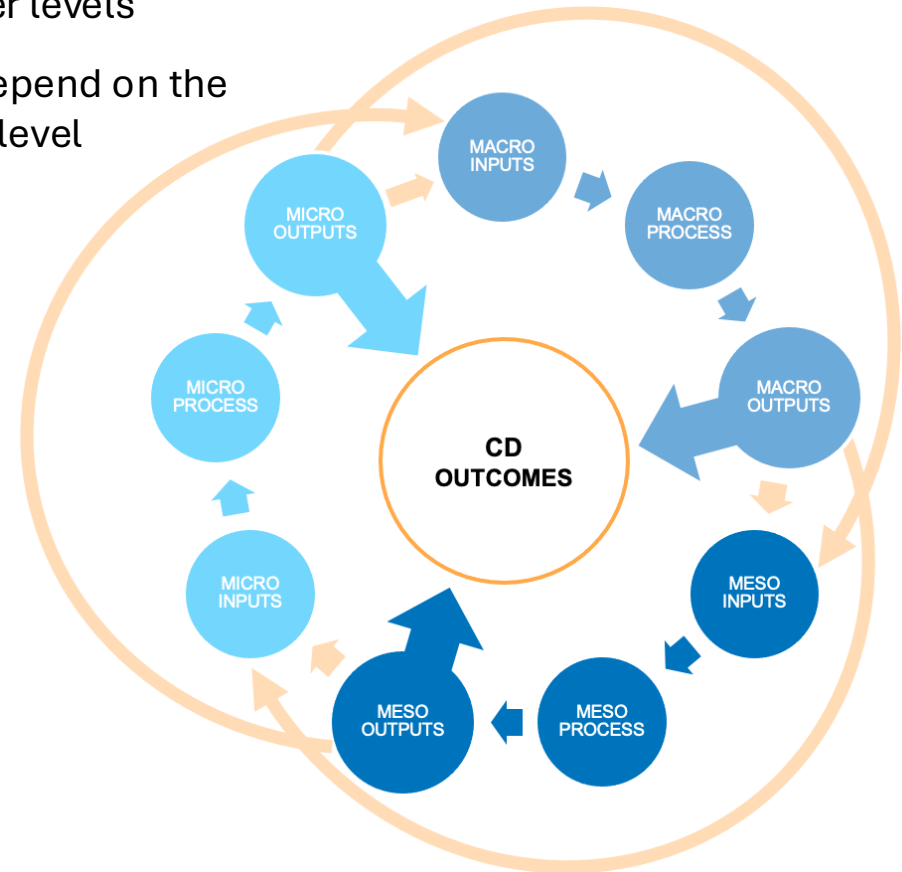


Own representation, based on FAO, 2010

A holistic CD Intervention exists on three levels that are interconnected and highly complementary to each other.

CD Results Chain

- Outputs of each level also serve as inputs to other levels
- CD outcomes depend on the outputs at each level together
- Non-linear



Part 1:

Research Idea and Hypotheses

Research Idea

Hypothesis

Holistic CD interventions, i.e. those addressing the complementarity amongst all levels, achieve better evaluation outcomes

Data

1. The World Bank Project Appraisal Documents (PADs)
2. IEG Project Evaluation Outcomes Database

Methodology

Step 1. Classification of projects based on the number of addressed levels using a computational text analysis methods (CTA)
Step 2: Regression analysis with number of addressed levels as an independent variable

Empirical Strategy: Three Interconnected Hypotheses

H1

CD projects receive better evaluation results than non-CD projects

H2

CD projects can be identified by more level-balanced approach

H3

A level-balanced approach increases the evaluation success in addition to the capacity-building nature of the project

Part 2:

Data & Pipeline

Three Interlocking Data Sources

1 World Bank Project Database

- 27,534 projects
- Downloaded from [the Website](#)
- Key variables: Project title, description, financing, timelines, regions.
- Basis for CD identification.

27,534

WB projects total

2 Project Appraisal Documents (PADs)

- Investment Project Financing (IPF) documents produced at project approval.
- Retrieved via [WB Documents & Report Open API](#) as .txt files.
- Contains beneficiary information.

5,845

PADs downloaded

3 IEG Evaluation Ratings

- Independent Evaluation Group (IEG) outcome ratings for completed projects
- Downloaded from [the Website](#)
- Six-point scale from Highly Unsatisfactory to Highly Satisfactory.

13,542

projects rated

How We Identify Capacity Development Projects

CD projects = Projects that explicitly mention CD keywords in their title and/or description

Computational Text Analysis

proximity-based n-gram search on project titles and descriptions, identifying words signalling “capacity” and words signalling “building” or “development” (“action” verbs: build, develop, strengthen, enhance, improve, expand, increase, or extend) when they occurred within a maximum window of two words from each other

Example — captured patterns:

"capacity building activities" ✓
"build organizational management capacity" ✓
"strengthening institutional capacity" ✓

Proximity-based n-gram search on project titles and descriptions

Action words:

build, develop, strengthen, enhance, improve, expand, increase, extend

Capacity words:

capacity, capacities, cpcty (abbreviated forms)

Rule:

Verb and capacity words within a window of max 2 words apart — in either order

2,516 CD projects in total — 9% of overall portfolio

PADs & IEG data - Our Key Sources

What is a PAD?

- 1 Produced at the approval stage of World Bank Investment Project Financing (IPF) operations
- 2 Standardised document structure including project description, objectives, financing, and — crucially — Beneficiaries section
- 3 Should be available for ~17,781 approved IPF projects, with only ~6,900 in the WB documents repository and ~5500 disclosed

What is in IEG data?

Project evaluation outcomes based either:

- 1) Implementation Completion Report Review (ICRRs), or
- 2) Project Performance Assessment Report (PPAR)

IEG Evaluation outcome rating:

“The extent to which a project efficiently achieved <..> its relevant objectives. The outcome rating integrates three underlying dimensions: relevance, efficacy, and efficiency” (IEG, n.d.-b, p. 2).

World Bank Project Lifecycle



We use PAD texts and IEG outcomes

Beneficiary Section Extraction: Pipeline & Outcomes

Line-by-line extraction pipeline from raw .txt files



Target heading patterns (case-insensitive):

B. Project Beneficiaries
C. Beneficiaries
3. Direct Project Beneficiaries
IV. Beneficiaries and Target Population
B. Target Population

Required format: numbered prefix (B. / 3. / IV.) + heading keyword

Our core assumption:

Whenever a project targets micro-, meso-, or macro-level actors, it will name level-specific keywords in its Beneficiaries section.

Extraction outcomes

3,555

Successfully extracted

~1,800

No Beneficiaries section

~190

Section < 200 chars (re-run)

~40

Heading with no body text

14

Inaccessible documents/ Undisclosed

Part 3:

Level Identification via CTA

Dictionary Design: High- and Low-Confidence Level Terms

An iterative process: literature → identification → manual inspection → dictionary update (×3)

MICRO

Individual Level

Human development, knowledge and skills training, individual capacity enhancement

High confidence ×3

Examples: individuals, households, families

Low confidence ×1

Examples: communities, training, users

MESO

Organisational Level

Structural and procedural change, organizational development

High confidence ×3

Examples: agencies, municipalities, organizations

Low confidence ×1

Examples: communities, staff, local administration

MACRO

Systemic / Enabling Level

Legal and regulatory transformation, governance reform, systemic enabling environment

High confidence ×3

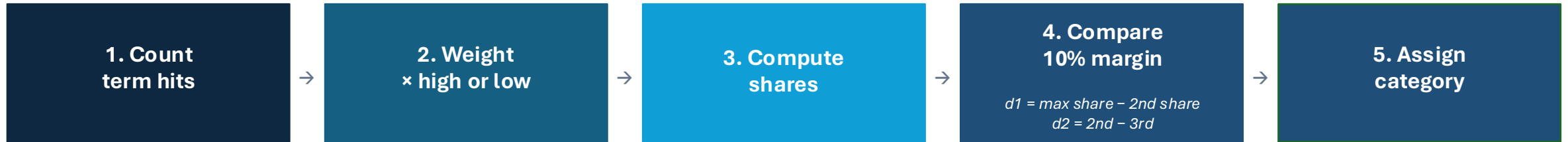
Examples: national, policy reform, regulator, ministry/ministries

Low confidence ×1

Examples: system, government, legal framework

Note: Ambiguous terms (e.g. 'community', 'government') appear across levels assigned with low weight

From Words to Scores to Classification



The resulting classification (10% threshold):

Level dominance	Number of projects
Micro-dominant	1367
Meso-dominant	984
Macro-dominant	336
Micro–Meso joint emphasis	288
Meso–Macro joint emphasis	226
Micro–Macro joint emphasis	79
Tri-level balanced	229
Low evidence	46

Alternative: Balance Indices

Herfindahl-Hirschman Balance Index

Originally from market concentration analysis.

$$B_{HHI} = 1 - (s_{micro}^2 + s_{meso}^2 + s_{macro}^2)$$

- Value = 0: complete concentration on one level
- Value → 0.67: perfect tri-level balance

Atkinson Balance Index

From welfare economics (inequality measurement).

$$B_A = 3(s_{micro} + s_{meso} + s_{macro})^{1/3}$$

- Sensitive to small levels: penalises zero-emphasis on any level more strongly
- $\epsilon = 1$: equal aversion to all types of imbalance

Both indices are inverted for intuitive interpretation: higher values = more balance

Questions? Comments?

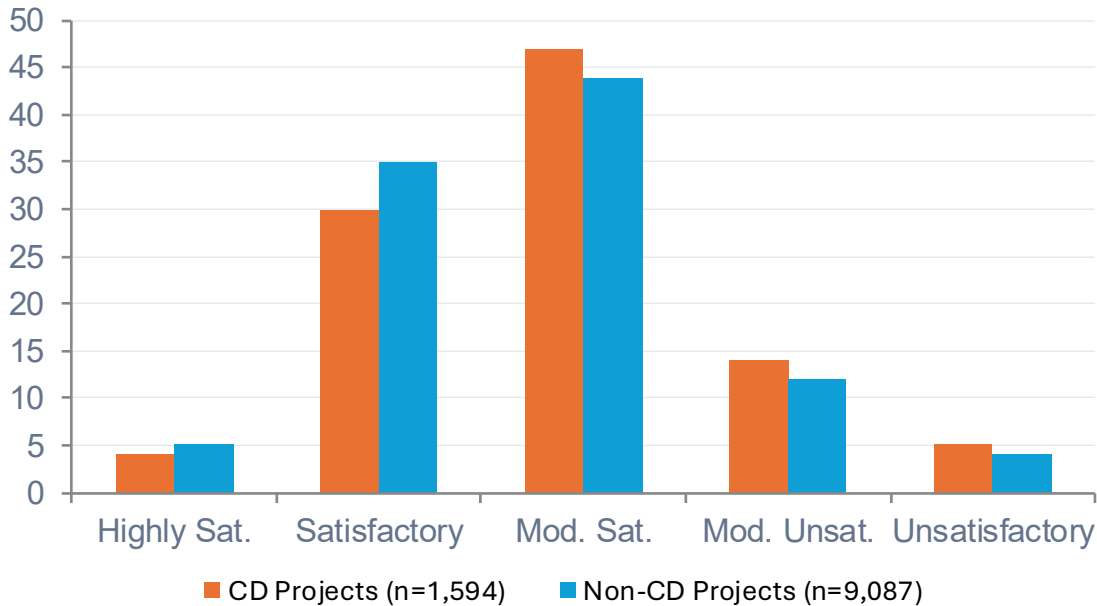
Part 5:

Key Findings

H1 rejected: CD Projects Systematically Underperform

CD-labelled projects receive worse evaluation outcomes on average

Descriptive: IEG Rating Distribution



Regression: CD Penalty holds controlling for...

CD coefficient, all 7 specs — N = 10,239

M1: Project volume + duration	-0.268***
M2: M1 + GDP per capita + population	-0.271***
M3: M2 + 7 WB regions	-0.265***
M4: M3 + 12 sector categories	-0.261***
M5: M4 + 17 lending instruments	-0.278***
M6: M5 + 4 WB practice groups	-0.272***
M7: M6 + Bank performance rating	-0.319***

The 'CD Penalty'

CD projects are significantly less likely to receive satisfactory evaluation outcomes — even after controlling for project size, country context, sector, instrument, and practice group. The penalty is robust across all specifications.

H2 confirmed: CD Projects Are More Level-Balanced

If CD is multi-level in design in practice, CD projects should exhibit broader level coverage — and they do

Logit models: dependent variable = CD indicator (Cap3), **Ref.level category** = “Tri-level balanced”, final sample (N = 2,514)

Micro-dominant	-0.904***	Most pronounced negative effect — micro-only projects are significantly less likely to be CD
Micro–Meso joint	-0.509**	Joint emphasis also negative vs. tri-level balance
Meso-dominant	-0.279	Not significant
Macro-dominant	0.063	Not significant
Atkinson Balance	3.371***	Higher balance strongly predicts CD status

What this tells us:

Micro-dominant projects are significantly LESS likely than Tri-level balanced projects to be classified as CD (-0.904***).

This aligns with theory — single-level 'training-only' projects should not claim CD status.

Higher balance (Atkinson index) is strongly associated with CD status (+3.371***).

CD projects are designed with broader beneficiary scopes — this is visible in the text.

H3: Does Level Balance Predict Better Outcomes?

Ordered logit models on final analysis sample (N = 2,514)

Model	CD Indicator (Cap3)	Level Variable Added	Effect on Outcome	McFadden R ²
M1: Baseline	-0.248**	— (no level variable)	—	0.126
M2: M1 + Categories	-0.234**	Level categories (7 dummies)	Micro–Macro: -0.54*; Meso–Macro: +0.31	0.128
M3: M2 + HHI balance	-0.230**	HHI balance index	-0.497**	0.125
M4: M3 + Atkinson balance	-0.221**	Atkinson balance index	-0.892***	0.125

“CD Penalty” persists

After controlling for level balance, the CD indicator remains negative and significant — the penalty cannot be explained by level composition alone

Balance matters — direction unexpected

Higher balance indices show significant negative effects — balanced projects perform worse on average. This contradicts H3 but opens further discussion.

Part 6: Reflection & Takeaways

Takeaways for Evaluation Practice

1 Text project data carries evaluation-relevant information

Large-scale project documentation routinely produced by development organisations contains systematic information that can be operationalized by evaluation practitioners

2 CTA enables meta-evaluation at scale

Dictionary-based approaches allow analysts to work across thousands of documents, extracting structured variables from unstructured text. The method is transparent, replicable, and low-cost.

Discussion

What can explain the negative performance of CD projects?

What can explain the negative performance of multi-level projects?

Are current evaluation frameworks suitable for capturing CD effectiveness?

If project-based development programming suitable for CD interventions at all?

Questions & Discussion



Oliver Serfling

Professor of Economic Policy & Development Economics
Rhine-Waal University of Applied Sciences
oliver.serfling@hochschule-rhein-waal.de

Taya Romanov

Junior Researcher, TransRegINT Project
Rhine-Waal University of Applied Sciences
taysiya.romanov@hochschule-rhein-waal.de

Limitations: What This Study Cannot Claim

Honest transparency is part of the methodology — not a weakness

IEG ratings as CD performance proxies

The IEG Evaluation Outcome rating captures whether a project achieved its stated development objectives — it is not specifically designed to assess capacity development. CD outcomes are long-term, emergent, and non-linear. Short project evaluation cycles (4–6 years) may systematically underestimate CD impact.

Omitted variable bias

Key theoretically relevant variables — project manager quality, institutional quality, political stability — are not available in public datasets. Research shows these can have substantial explanatory power. Our estimates are associational and should not be interpreted causally.

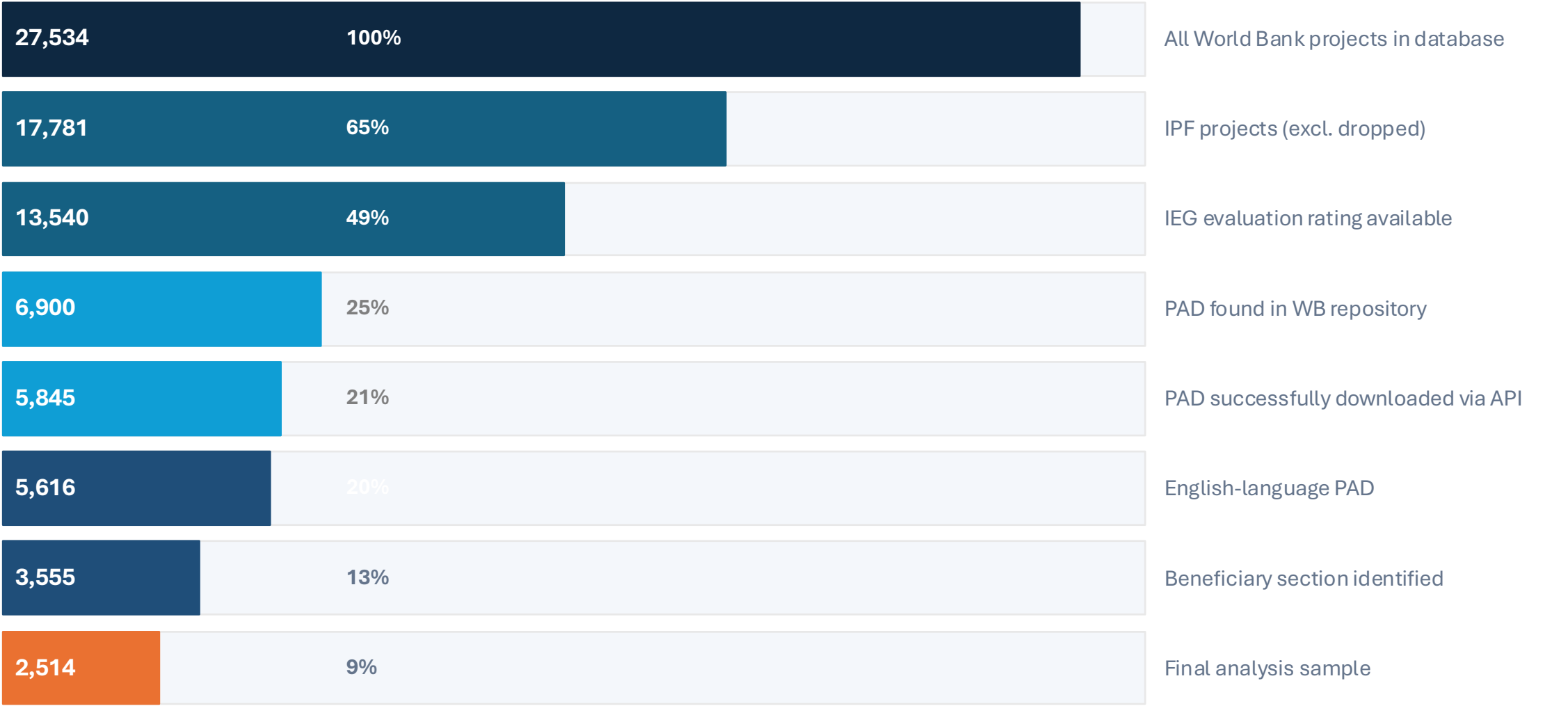
Section extraction

Extraction success depends on document formatting consistency. PADs vary widely in structure; differences in documentation quality across time and sectors may partly drive variation.

Selection bias

Projects with accessible PADs are not randomly distributed. The ~49% gap between expected and available documents introduces potential selection. Non-disclosed projects may differ systematically.

Sample Construction: From 27,534 to 2,514 Projects



Main attrition: PAD availability

Backup: Full Regression Results — H1 (Full Sample, N = 10,239)

	M1	M2	M3	M4	M5	M6	M7
log(Project Volume)	0.129***	0.070***	0.077***	0.069***	0.086***	0.082***	0.099***
Project Duration	-0.006	-0.009	-0.012**	-0.020***	-0.043***	-0.046***	-0.063***
CD Indicator (Cap3)	-0.268***	-0.271***	-0.265***	-0.261***	-0.278***	-0.272***	-0.319***
log(GDP per capita)	—	0.167***	0.079***	0.080***	0.092***	0.092***	0.073***
log(Population)	—	-0.105***	-0.029	-0.026	-0.049*	-0.050*	-0.052*
Controls: Region (7)	—	—	✓	✓	✓	✓	✓
Controls: Sector (12)	—	—	—	✓	✓	✓	✓
Controls: Instrument (17)	—	—	—	—	✓	✓	✓
Controls: Practice Group (4)	—	—	—	—	—	✓	✓
Controls: Bank Performance	—	—	—	—	—	—	✓
McFadden R ²	0.005	0.009	0.015	0.018	0.022	0.023	0.092

Note: *p<0.1. **p<0.05. ***p<0.01. Standard errors in parentheses. Ordered logit models.

Backup: Full Regression Results — H3 (Analysis Sample, N = 2,514)

	M1	M2	M3	M4	M5
log1p(Project Volume)	0.051	0.053	0.039	0.041	0.050
Project Duration	-0.161***	-0.159***	-0.163***	-0.162***	-0.165***
CD Indicator (Cap3)	-0.248**	-0.234**	-0.230**	-0.221**	-0.237**
log(GDP per capita)	0.165***	0.161***	0.148**	0.143**	0.150**
log(Population)	-0.108	-0.111*	-0.089	-0.085	-0.100
Macro-dominant	—	-0.116	—	—	-0.136
Meso-dominant	—	0.024	—	—	0.029
Meso–Macro joint	—	0.311	—	—	0.315
Micro-dominant	—	0.172	—	—	0.137
Micro–Macro joint	—	-0.540*	—	—	-0.569*
Micro–Meso joint	—	0.333*	—	—	0.315
HHI Balance Index	—	—	-0.497**	—	—
Atkinson Balance Index	—	—	—	-0.892***	—
Controls: Region/Sector/Practice	✓	✓	✓	✓	✓
Controls: Bank Performance	✓	✓	✓	✓	✓

Note: *p<0.1. **p<0.05. ***p<0.01. Reference category for level: category: Tri-level balanced.

Backup: Analytical Workflow — 10 Scripts in Sequence

1	<code>WB_CDIdentification.R</code>	Data cleaning, sector imputation, CD identification (Cap1/2/3)	<i>WB.RData</i>
2	<code>WB_LendingInstrumentType.R</code>	Lending instrument correspondence table	<i>Stats</i>
3	<code>WB_PADs_get.R</code>	WB API loop — download PADs as .txt files	<i>5,846 .txt files</i>
4	<code>PADs_dataframeCreation.R</code>	Pull .txt files into R dataframe	<i>PADs.RData</i>
5	<code>PAD_Beneficiaries.R</code>	Language detection, line-by-line section extraction, QA	<i>PADs_BenefiSection.RData</i>
6	<code>Levels_dictionaries.R</code>	High/low confidence dictionaries for micro/meso/macro	<i>R objects</i>
7	<code>LevelCoverageVariableConstruction.R</code>	Term counting, weighting, shares, categories, balance indices	<i>PADs_Levels_m10.RData</i>
8	<code>FinalSampleConstruction.R</code>	Merge project + IEG + levels data; descriptive statistics	<i>EO_Paper_AnalysisSample10.RData</i>
9	<code>WB_IEG_EO_H1.R</code>	Exploratory models on full sample (H1 formulation)	<i>Models, tables, plots</i>
10	<code>WB_IEG_EO_H2H3.R</code>	Final models on analysis sample; all tables and figures	<i>Models, tables, plots</i>