IDB TECHNICAL NOTE IDB-TN-3399

What Do Development Banks Learn from their Operations?

A Textual Analysis of Project Completion Reports from the IDB

Magdalena Rossetti

Inter-American Development Bank
Knowledge and Learning Division

September, 2026

What Do Development Banks Learn from their Operations?

A Textual Analysis of Project Completion Reports from the IDB

Magdalena Rossetti

Inter-American Development Bank
Knowledge and Learning Division

September, 2026

INDEX

Introduction

Multilateral development banks finance operations across many different organizational sectors, countries, and institutional contexts. Because conditions vary and no two projects unfold the same way, the ability to learn systematically from experience is central to improving development effectiveness (Koporcic et al. 2025; Sillito and Pope 2024; Mahfouz et al. 2019; Lis 2012). The impact of even a well-designed project can be diminished by weak execution, governance bottlenecks, and inadequate monitoring. How experience gets documented, interpreted, and transmitted across the organization matters, then, as much as individual project outcomes. This centrality of learning is reflected in how development institutions increasingly position themselves as knowledge banks, whose comparative advantage lies in translating operational experience into better interventions.

To institutionalize learning, multilateral development banks embed formal documentation requirements into their project cycles. At the Inter-American Development Bank (IDB), operational teams are required to include a dedicated "lessons learned" section in every Project Completion Report, a standardized instrument produced when an operation closes. These lessons are intended to gather the most important insights from implementation and make them available to inform future work.

In principle, lessons learned inform the preparation of new operations by signaling recurrent implementation risks and highlighting the institutional conditions associated with better results. At an institutional level, they also contribute to the scaling of knowledge across sectors and countries, affecting the Bank's ability to translate dispersed operational experience into systematic improvements in effectiveness.

The IDB Knowledge Model structures institutional learning around a knowledge loop that connects analytical and operational cycles across four stages of a project: programming, design, execution, and closure and evaluation. In the operational cycle, the closure and evaluation stage is explicitly defined as the moment when insights and lessons feed into internal policy improvement and inform future projects. (see Inter-American Development Bank, IDB Invest, and IDB Lab 2026). It is the point at which operational experience is supposed to reenter the system as structured learning, feeding forward into the next cycle of programming and design. Understanding what kinds of learning are actually codified in formal documentation is, therefore, central not only to knowledge management but to core operational decision making. When this mechanism does not function as intended, it breaks the knowledge loop, with consequences that extend well beyond compliance with reporting requirements.

Yet despite their centrality in operational reporting, little is known about the collective architecture of operational lessons. Although individual lessons are typically read at the project or sector level, they have not been analyzed systematically as a corpus. As a result, basic questions remain unanswered about how institutional learning is codified in formal documentation. Does recorded

learning reflect sector-specific operational experience, or does it emphasize more managerial concerns? Does it evolve over time in response to changing operational contexts, or does it remain structurally stable? And how is learning articulated discursively—that is, how prescriptive are the lessons, and how explicitly do they assign responsibility to implementing actors? Computational analysis of World Bank annual reports suggests that, over time, institutional language tends to become more abstract and less connected to concrete actors (Moretti and Pestre 2015), but whether this pattern characterizes operational learning documents specifically remains an open question.

This study constructs and analyzes a corpus of 6,620 lesson statements extracted from Project Completion Reports of IDB loan operations published between 2014 and 2024. A computational text analysis pipeline is applied, combining multilingual sentence embeddings, density-based clustering, and topic modeling to recover the latent thematic structure of the corpus. Thematic assignments are then linked to sectoral metadata, temporal patterns, and discursive features, which are examined through logistic regression models with year and sector fixed effects to assess how learning varies across institutional units and over time. The analysis produces three main findings. First, the structure of IDB lessons is highly concentrated, with over 93 percent of all documented lessons falling within a single dominant thematic area, centered on project design and implementation management. Second, this concentration is stable across sectors and over time, though its internal composition shifts gradually. Third, the discursive structure of lessons varies systematically by thematic area, in ways that speak to how responsibility for action is assigned within the codified record.

Lessons Learned in the IDB Operational Cycle

The IDB defines lessons learned as knowledge gained from reflecting on a completed project's design and implementation (Roca, Zepeda, and García Ferro 2025). The purpose of this process is to gain an understanding of which factors, decisions, and contextual conditions shaped the project's results and how they did so. Lessons come in two parts. A *finding* describes the evidence connecting project outcomes to specific actions, institutional arrangements, or external conditions that helped or hindered performance. A *recommendation* builds on that evidence to propose practical steps that can be applied in similar situations to reduce risks, solve problems, or replicate what worked.

Rather than being confined to project closure, lessons learned are embedded throughout the operational cycle. During the design stage, for example, teams are required to document relevant lessons from prior operations in their loan proposals and explain how those lessons will inform the new intervention. During execution, they are encouraged to record implementation challenges and emerging insights in periodic monitoring reports. At closure, the lessons accumulated across the cycle are consolidated and formalized in the Project Completion Report, the Bank's main self-evaluation instrument.

The Project Completion Report is a standardized document prepared by the unit responsible for the operation at the end of its execution phase. It covers implementation performance, outputs, outcomes, risk management, and fiduciary compliance. The lessons learned section is intended to distill the most important insights from the project. Teams are explicitly guided to address what worked well, what could have been done differently, which decisions or contextual factors shaped results, and whether particular approaches should be replicated or avoided in future operations.

Beyond documenting project experience, lessons learned are intended to inform evidence-based operational decision making and improve development results and impacts. Institutional guidance emphasizes their role in improving the preparation of new operations, strengthening supervisory practices, and supporting the design of monitoring and evaluability frameworks. By making operational knowledge visible and transferable across sectors and countries, lessons are also expected to contribute to institutional learning agendas, inform portfolio-level strategy and performance, and support the scaling of effective approaches. In this sense, the lessons learned system is not only a reporting requirement but a continuous learning mechanism embedded across the project cycle, through which dispersed project experience can shape future interventions, portfolio management, and development outcomes.

Why Codified Learning May Diverge from Operational Experience

The Competing Logic of the Formalization of Knowledge

Formal lessons learned systems convert project experience into reusable organizational knowledge. In theories of organizational learning, experience becomes institutional learning when it is "encoded" into routines, rules, or shared understandings that guide future action (Levitt and March 1988; Huber 1991). This is particularly important in project-based organizations where teams are temporary, which presents a risk that operational knowledge will be lost once projects close (Brady and Davies 2004; Sense 2011). Codification reduces dependence on individual memory and facilitates the transfer of experience across sectors, countries, and project cycles.

Three somewhat competing rationales shape what gets documented in formal lessons learned systems. The first is *accountability*. In public and development organizations, formal reporting demonstrates rationality, oversight, and responsible use of resources (Power 1999; OECD 2021). Performance systems can simultaneously support learning and reinforce compliance-oriented behavior, depending on how incentives are structured (Moynihan 2008; Moynihan and Landuyt 2009). When accountability concerns are foremost, lessons are more likely to emphasize measurable and comparable managerial thematic areas, such as fiduciary controls, monitoring systems, and risk management.

The second rationale is *knowledge retention*. In institutions with high staff mobility and geographically dispersed portfolios, formal lessons are especially important to preserve experiential knowledge beyond the life of individual project teams (Huber 1991; Argote and Miron-Spektor 2011). To remain transferable among contexts, such knowledge tends to be expressed at a level of generality that allows it to travel across sectors and countries, favoring portable managerial insights over context-specific technical learning.

The third rationale is *operational feedback*. Lessons learned systems are also intended to connect project closure to future design decisions by identifying implementation constraints and assigning responsibility for corrective action (Schindler and Eppler 2003). When this rationale is strong, lessons are more likely to specify decision points, identify institutional actors, and frame recommendations in actionable terms.

In multilateral development banks, the three rationales coexist and may pull documentation practices in different directions. Evaluation systems, for example, are shaped not only by learning objectives but by reputational and accountability pressures (Buntaine, Parks, and Buch 2017), and strong oversight incentives may encourage compliance-oriented reporting rather than adaptive learning (Honig 2018). The thematic and discursive structure of lessons learned can, therefore, provide indirect evidence on which rationale is most influential in practice.

Structural Limitations on Codifying Learning

Investing in lessons learned systems does not guarantee the knowledge they produce will reflect the full range of operational experience. Even well-designed systems operate within four structural constraints that shape what gets recorded and how. This pushes codified learning in predictable directions, with implications for its policy relevance and institutional utility.

The first limitation is *fragmentation*. Projects generate rich experiential knowledge, but closure often coincides with staff turnover and the dissolution of teams, leaving knowledge localized unless it is deliberately institutionalized (Bakker et al. 2016). Although learning occurs simultaneously at project and organizational levels, mechanisms that support one do not automatically support the other (Sydow, Lindkvist, and DeFillippi 2020; Brady and Davies 2004; Duffield and Whitty 2015; 2016). Evidence from World Bank staff surveys suggests that learning remains heavily dependent on informal knowledge held by individuals, with inadequate handover arrangements contributing to its loss (IEG 2015). In fragmented environments, lesson systems risk preserving episodic insights rather than building cumulative institutional memory.

The second obstacle is *standardization pressure*. Multilateral development banks operate under strong external demands for evaluation, benchmarking, and performance measurement that favor standardized reporting categories comparable across contexts (Broome, Homolar, and Kranke 2018). Cross-cutting managerial thematic areas, such as risk management, monitoring systems, and fiduciary controls, are more easily standardized than sector-specific technical insights, making them more likely to appear in formal documentation. Organizations in structured institutional fields often end up using similar reporting formats over time (DiMaggio and Powell 1983), and prevailing conventions can shape how operational reality is framed and recorded (Cornelissen et al. 2015; Christensen and Lægreid 2017). As a result, codified learning may emphasize what is measurable and comparable rather than what is substantively distinctive.

The third limitation is *transfer friction*. Codifying experience oversimplifies causal relationships, removes contextual detail, and resolves ambiguities whose inclusion may be analytically important (Nonaka 1994). Transferring knowledge within organizations is inherently difficult, as recipients may lack the capacity to absorb it, the conditions that made a lesson applicable in its original setting may not be present elsewhere, and meanings can shift during transmission (Szulanski 1996; Cohen and Levinthal 1990). In institutions that operate across different sectors and country environments, lessons are, therefore, often written at a level of generality broad enough to travel across operations. This favors portable managerial principles over context-specific insights that may be more directly actionable in similar implementation settings.

The fourth obstacle is *accountability pressure*. In organizations subject to external scrutiny, formal documents function not only as records of experience but as signals of rationality, control, and responsible management (Meyer and Rowan 1977; Power 1999). Strong accountability demands can create incentives to acknowledge problems while avoiding explicit attribution of responsibility (Overman and Schillemans 2022; Schillemans and Busuioc 2015). This can result in lessons

framed in prescriptive terms but with limited clarity about who should act on them, reducing their practical usefulness for future decision making.

Methodology

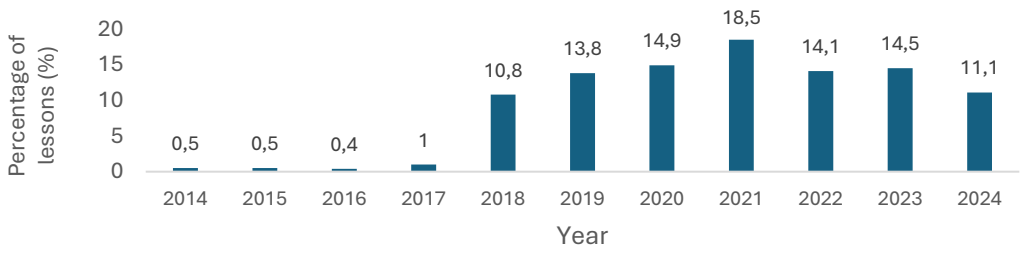
Data

The analysis draws on a corpus of 6,620 lesson statements extracted from Project Completion Reports (PCRs) published between 2014 and 2024. Each observation corresponds to a single lesson statement, as recorded in the PCR's lessons learned section. The unit of analysis is the individual lesson rather than the operation. Since each operation can report more than one lesson, this choice captures within-project variation in what is emphasized and how learning is articulated, rather than collapsing the experience of an entire project into a single summary measure.

Each lesson is linked to operation-level metadata from its corresponding PCR, including operation number and name, country, sector, division, subsector, approval year, and document number. These metadata allow us to examine how documented learning varies across institutional units and over time. Spanish-language lessons are retained in their original form where available; otherwise, the original English text is used. Lesson statements average approximately 110 words in length, with a median of around 93 words.

The corpus is unevenly distributed across IDB divisions and countries. Three divisions—Water and Sanitation (WSA), Rural Development and Natural Disasters (RND), and Transport (TSP)—together account for about 40 percent of all lessons, and Brazil alone contributes 16.9 percent of the total (see Appendix, figures A1 and A2).

Figure 1. Distribution of Lessons by Approval Year



Source: Author's calculations, based on Project Completion Reports published by the Inter-American Development Bank between 2014 and 2024.

Figure 1 shows the distribution of lessons by approval year. Annual volume remains below 1 percent of the corpus between 2014 and 2017, followed by a sharp increase beginning in 2018,

peaking at 18.5 percent in 2021 and stabilizing at high levels through 2023 before declining to 11.1 percent in 2024. Given the low volumes before 2018, the main analysis focuses on the 2018–24 period, where coverage is sufficient for reliable inference.

Analytical Strategy

The objective of the empirical analysis is not prediction but, rather, measurement, to recover the latent thematic structure of lessons learned, quantify how lessons are distributed across thematic areas, and evaluate the patterns recovered against the interpretive frameworks developed above. The raw lesson text is transformed into structured representations, and a sequence of analytical steps is applied to identify thematic groupings.[1]

The first step in the analysis is to *preprocess* the text. We apply a consistent preprocessing procedure to all lessons in the corpus. This procedure corrects character-encoding problems that can cause words or accented characters to be displayed incorrectly, removes formatting artifacts such as URLs, numerical strings, and nonalphabetic symbols, and applies lemmatization—that is, it reduces inflected words to their base, or dictionary, forms. Common stop words in Spanish and English, such as articles and prepositions, are removed, except for modal and normative terms such as "should" and "must," which signal obligation or recommendation and carry substantive analytical content. Highly generic institutional terms, such as "project" and "program," are also excluded, as they are mechanically frequent but analytically uninformative.

The second step is *semantic representation*. Each lesson is represented as a numerical summary of its meaning, generated by a multilingual language model (Reimers and Gurevych 2019), which locates texts expressing similar ideas near one another, even if they use different vocabulary, in a continuous semantic space. This is particularly important for a bilingual corpus, where the same operational concept may be expressed differently in the two languages. Similarity among statements is measured using cosine distance.

The third step is *dimensionality reduction*. We compress the numerical summaries of lessons into a lower-dimensional representation using Uniform Manifold Approximation and Projection (UMAP; McInnes, Healy, and Melville 2018), which preserves relative distances among semantically similar statements while improving statistical stability for the clustering stage.

The fourth step is *density-based clustering*. We apply Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN; Campello, Moulavi, and Sander 2013), which locates regions of the semantic space where many statements are concentrated and treats them as clusters. The algorithm determines both the number and size of clusters endogenously and allows a small share of observations to remain unassigned. Under the baseline specification, the procedure yields one large cluster containing the majority of observations, several smaller clusters, and a limited share of unclassified statements. Robustness checks confirming that the dominant cluster is not an algorithmic artifact are reported in the Appendix.

[1] Technical parameters and robustness checks are documented in the Appendix.

The fifth step is *subtopic modeling* within the dominant cluster. Nonnegative matrix factorization (NMF) is applied to extract recurring word patterns and assign each statement to the subtheme that best characterizes its vocabulary (Lee and Seung 1999).[2] This step works directly with lexical content—that is, the words in the text, as opposed to the grammar, sentence structure, or meaning—to improve interpretability. This comes at the cost of some sensitivity to exact wording, but it complements the clustering based on meaning by associating the vocabulary with each previously identified thematic area. The use of this methodology produces eight distinct subthemes (for details, see the Appendix).

The sixth (and final) step is the *analysis of discursive features*. Two analytically distinct dimensions are highlighted: normative language, measuring whether a lesson frames its recommendations as suggestions or as mandates, and responsibility attribution, tracking whether a lesson names specific actors who are expected to act on those recommendations. We construct a separate binary indicator for each dimension: one records whether normative language is present, and the other records whether one or more institutional actors are identified. We then construct a third, more restrictive indicator that takes the value of one only when both conditions are met: the lesson identifies one or more implementing actors and frames the recommendation in normative terms. This joint indicator is treated as the strictest proxy for explicit responsibility attribution. Logistic regression models are estimated for each outcome with year and sector fixed effects, and average marginal effects are reported in percentage-point terms. Standard errors are clustered at the operation level.[3]

[2] Unlike the embedding-based approach used in the clustering stage, NMF does not determine cluster boundaries but provides a descriptive decomposition of the preidentified dominant thematic area. The sequential design combines semantic coherence at the clustering stage with lexical transparency at the interpretation stage.

[3] Year fixed effects absorb time-varying shocks common to all lessons in a given year, including changes in reporting guidelines or organizational priorities. Sector fixed effects absorb systematic differences in discursive style across operational areas, ensuring that estimated differences across thematic areas are not confounded by sectoral composition.

What Does the IDB Learn from Its Operations?

A Highly Concentrated Architecture of Institutional Learning

The clustering procedure reveals a highly concentrated thematic structure in the operational lessons in the sample. A single cluster, referred to here as Project Design and Implementation Management, contains 6,172 observations and accounts for 93.2 percent of all documented lessons. Lessons in this cluster address the organizational, contractual, and governance conditions that determine whether projects translate their designs into actual results. The conditions include procurement readiness, institutional capacity, sustainability arrangements, and results measurement. The remaining substantive clusters are small and include Integrated Tourism Development and Destination Governance (1.1 percent), Integrity and Oversight in Procurement (0.7 percent), Technology Systems and Data Capacity (0.6 percent), and Integrating Gender into the Design of Operations (0.5 percent).

Table 1. Distribution of Lessons by Cluster

Cluster number	Cluster Name	Percentage of lessons
1	Project Design and Implementation Management	93.2
2	Integrated Tourism Development and Destination Governance	1.1
3	Integrity and Oversight in Procurement	0.7
4	Technology Systems and Data Capacity	0.6
5	Integrating Gender into the Design of Operations	0.5
-	Unclassified/Residual	3.9

Source: Author's calculations, based on Project Completion Reports published by the Inter-American Development Bank between 2014 and 2024.

The four minority clusters account for less than 3 percent of all documented lessons. They are, however, substantive and internally coherent. Cluster 2 includes lessons on aligning infrastructure investment, stakeholder coordination, and long-term sustainability within a shared territorial vision, while Cluster 3 covers fiduciary safeguards, due diligence, and the importance of independent oversight in contracting. Cluster 4 documents how investments in digital platforms and information systems require strong technical specifications and sustained institutional capacity to deliver results, and Cluster 5 emphasizes the importance of incorporating gender objectives and appropriate indicators from the design stage onward.

A question that arises is whether the concentration of lessons in a dominant cluster simply reflects how the portfolio is structured. If most operations belong to a few sectors, most lessons might naturally cluster together for compositional reasons alone. The data do not support this interpretation. The dominant cluster accounts for between 81 and 99 percent of lessons within every individual sector (see Table A6 in the Appendix), meaning the pattern holds regardless of which part of the portfolio is analyzed. What gets formally recorded as a lesson, across all sectors and departments, tends to be framed in general managerial terms rather than in terms specific to

any sector or policy area. This may reflect how reporting templates are structured, how project teams draw lessons from their experiences at project closure, or broader institutional conventions about what counts as a lesson worth documenting. The sections that follow examine whether the pattern holds across the internal structure of the dominant cluster, its sectoral and temporal composition, and the way lessons are linguistically framed.

The Dominant Cluster

Within the dominant cluster, through subtopic modeling, eight distinct subthemes were identified. Table 2 reports their distribution. Three subtopics together account for roughly two-thirds of lessons within the cluster, while the remaining five cover progressively smaller shares. None of the eight corresponds exactly with the technical content of a particular sector or policy area. Each instead isolates a different aspect of the same underlying problem: what determines whether a project translates its design into actual results, and where in the implementation cycle things tend to go wrong.

The eight subtopics can be divided into three groups. Topics 1–3, as listed in Table 2, address the structural conditions, organizational arrangements, physical infrastructure, and governance architecture that projects must inherit or create before implementation can begin. Topics 4–6 capture process conditions—that is, how projects are managed, monitored, and adapted once execution is underway. The final two subtopics concern strategic conditions, which characterize the political and institutional environment that determines whether investments and reforms produce durable results after the project closes.

Table 2. Subtopic Distribution within the Dominant Cluster

Subtopic	Share within cluster (%)
Operational Sustainability and Service Viability	25.7
Institutional Architecture and Fiduciary Capacity	20.5
Procurement and Execution	19.4
Governance and Political Sustainability of Reforms	15.7
Results, Attribution, and Evaluability	12.1
Monitoring and Information Systems	4.1
Risk Management	1.6
Operational Rules and Execution Routines	1.0

Source: Author's calculations, based on Project Completion Reports published by the Inter-American Development Bank between 2014 and 2024.
Notes: Subtopics are derived from a nonnegative matrix factorization (NMF) model applied to the 6,172 observations in the dominant cluster.

The following describes the subtopics in greater detail.

Operational Sustainability and Service Viability (25.7 percent) is the largest subtopic and the one most directly concerned with what happens after a project closes. Lessons here reveal that long-term effectiveness depends on whether services remain financially viable, institutionally governed, and socially embedded once Bank support ends. A recurring finding is that sustainability

risks do not typically stem from technical design failures. They stem, rather, from the absence of secured maintenance budgets, realistic arrangements for cost recovery, clear transfer of operational responsibility, and genuine community ownership.

Institutional Architecture and Fiduciary Capacity (20.5 percent) addresses the organizational conditions under which projects are executed. Lessons repeatedly find that implementation bottlenecks are less often attributable to what was designed than to who was responsible for implementing it. Among the bottlenecks that arise are staffing instability, fragmented mandates, and weak financial and procurement management within executing units. Projects performed better when executing units had clear authority, dedicated personnel, and robust fiduciary systems in place from the start.

Procurement and Execution (19.4 percent) concerns the technical and contractual readiness of specific investment operations, from preparation through contract management and works supervision. Lessons record that delays and cost overruns typically originate not in external shocks but in insufficient project maturity at approval, characterized by incomplete studies, poor cost estimates, unresolved permits, and misaligned designs that generate amendments and retendering once execution is underway. Procurement strategy is itself a binding constraint, with poorly designed bidding packages, limited market analysis, and inadequate risk allocation in contract terms reported to have contributed to inefficiencies independent of executing unit capacity. While the previous subtopic covers the organizational conditions of the unit responsible for execution, this one addresses the technical and contractual readiness of the operation itself.

Governance and Political Sustainability of Reforms (15.7 percent) brings together lessons from policy-based and programmatic operations, where the binding constraint is not organizational capacity but political durability. Lessons consistently find the effectiveness of reform depends less on formal legislative approval than on the governance arrangements that sustain measures over time, including sequencing, coordination among institutions, and political anchoring. Electoral cycles, changes in administration, and high turnover among senior officials are identified as recurrently disruptive. This subtopic is distinct from the sustainability subtopic in that it concerns the political conditions needed for durable reforms rather than the operational conditions needed for service continuity.

Results, Attribution, and Evaluability (12.1 percent) identifies recurring flaws in the design of results measurement frameworks. The core problem flagged across operations is not a lack of monitoring capacity but, rather, a conceptual issue. It includes indicators too aggregated to be attributable to the intervention, targets that are unrealistic relative to the scope of the project, and baselines that are fragile or abandoned midcycle. In short, this subtopic is about what gets measured and whether it can plausibly be linked to what the project did—in other words, problems that originated at the design stage.

Monitoring and Information Systems (4.1 percent) is the operational counterpart to Results, Attribution, and Evaluability. Where the latter subtopic concerns the logic of what to measure, this one concerns whether the institutional machinery to do the measuring exists and functions.

Lessons indicate that management information systems were frequently absent, fragmented, or implemented too late; that monitoring plans included in project documents were undermined by insufficient staffing and unclear responsibilities; and that monitoring efforts were oriented toward compliance reporting rather than informing decisions. The distinction is between knowing what to measure and having the systems and capacity to measure it.

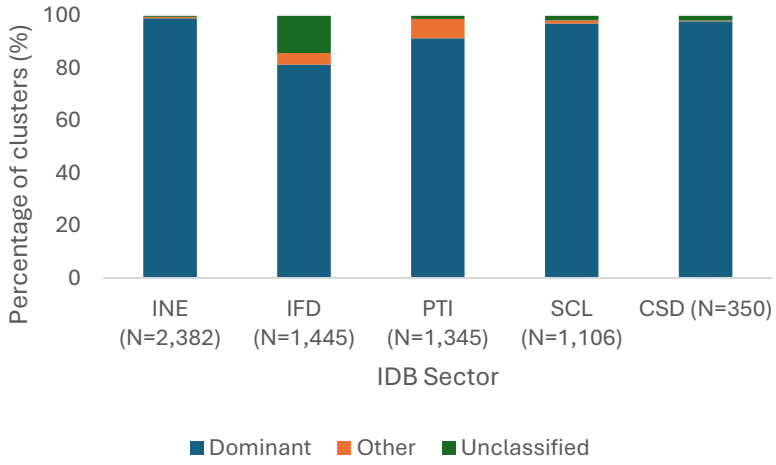
Risk Management (1.6 percent) focuses on a specific and recurring institutional failure in the form of risk assessment treated as a compliance exercise rather than a management tool. Lessons show across operations that risk matrices were prepared before approval but not updated during execution, not integrated into supervisory routines, and not linked to concrete mitigation measures with allocated budgets.

Operational Rules and Execution Routines (1.0 percent), the smallest subtopic, addresses a distinct failure mode at the level of day-to-day management. Execution falters here not because of unclear objectives or insufficient capacity but because of lax operational discipline resulting in overly optimistic timelines, delayed preparation of key documents, and insufficiently formalized coordination and procurement routines. Where Institutional Architecture concerns stable organizational structures and fiduciary systems, this subtopic concerns the procedural considerations that ensure those structures can produce consistent practice.

Sectoral Composition

The concentration of lessons documented at the beginning of this section is not specific to any organizational sector of the IDB. It characterizes the learning record across all operational departments. Figure 2 shows the share of lessons assigned to the dominant and other clusters, respectively, and the unclassified residual, by sector. The dominant cluster accounts for 99.0 percent of lessons in the Infrastructure and Energy (INE) Sector, 97.9 percent in Climate Change and Sustainable Development (CSD), 97.0 percent in the Social Sector (SCL), and 91.4 percent in Productivity, Trade and Innovation (PTI). The lowest share is in Institutions for Development (IFD), at 81.4 percent, where 14.2 percent of lessons remain unclassified. Even there, more than four in five lessons fall within the dominant managerial thematic area.

Figure 2. Distribution of HDBSCAN Clusters by Sector

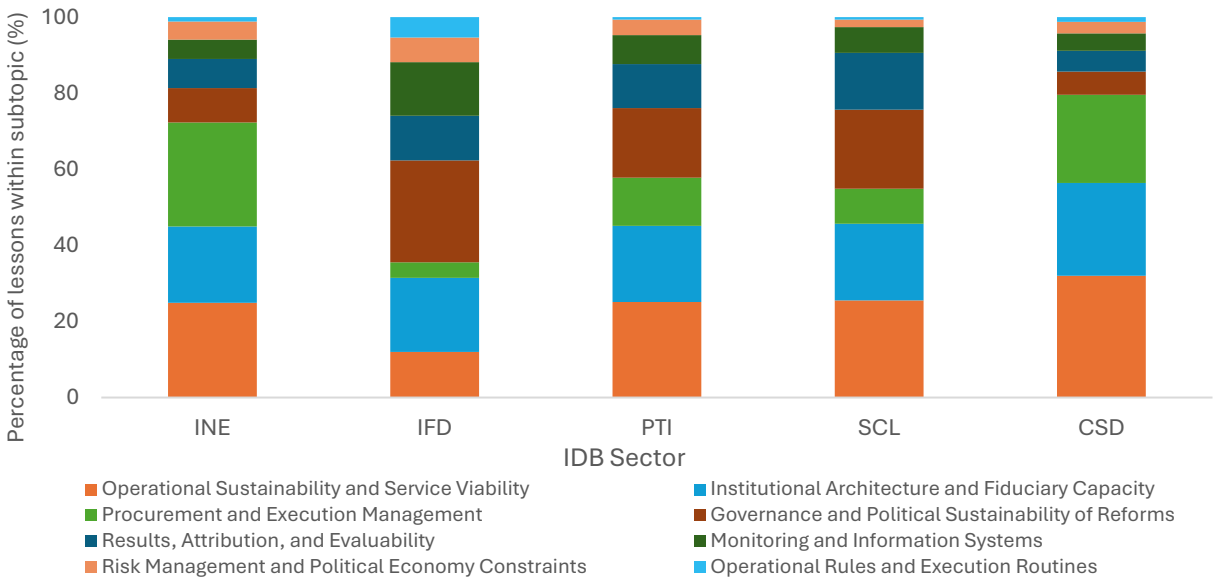


Source: Author's calculations, based on Project Completion Reports published by the Inter-American Development Bank between 2014 and 2024.
Note: INE = Infrastructure and Energy; IFD = Institutions for Development; PTI = Productivity, Trade and Innovation; SCL = Social Sector; and CSD = Climate Change and Sustainable Development.

The higher unclassified share in IFD is worth a brief note. The clustering procedure operates on the full corpus without separating observations by sector. Cluster boundaries are, therefore, defined by the global vocabulary of the data, which is anchored primarily by the language of investment operations. The higher unclassified share in IFD, a department whose portfolio centers on financial markets, fiscal management, and institutional reform, most plausibly reflects genuine semantic distance from that corpus-wide vocabulary rather than a data problem.

Within the dominant cluster, Figure 3 shows that the internal composition of lessons is broadly similar across sectors, with most departments sharing the same basic profile. The two largest subtopics in virtually every sector are Operational Sustainability and Service Viability and Institutional Architecture and Fiduciary Capacity, while the ranking of remaining subtopics does not vary dramatically across INE, PTI, SCL, and CSD. The main exception is IFD, which shows the highest share in Governance and Political Sustainability of Reforms (26.8 percent) and the second highest in Monitoring and Information Systems (14.1 percent), while Procurement and Execution is almost absent (4.1 percent).

Figure 3. Subtopic Distribution within the Dominant Cluster, by Sector



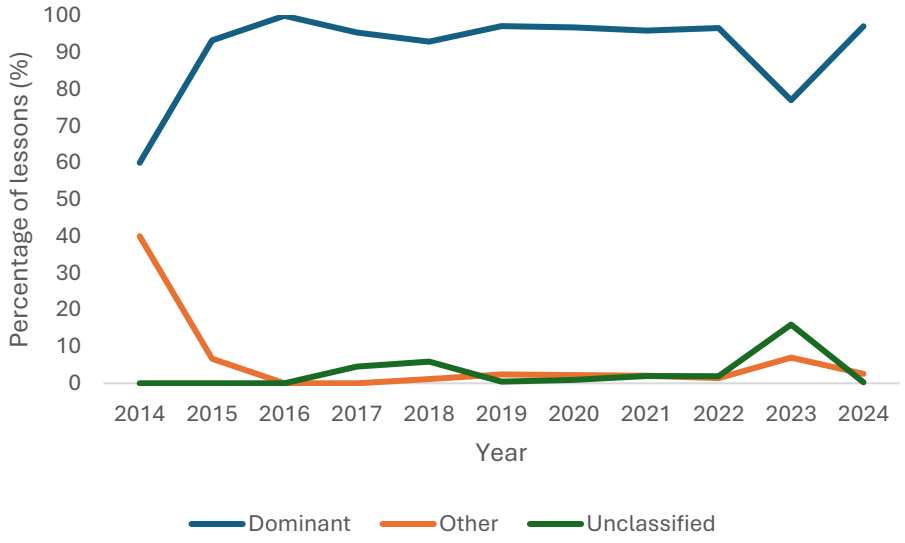
Source: Author's calculations, based on Project Completion Reports published by the Inter-American Development Bank between 2014 and 2024.

The uniformity across the remaining four sectors is itself part of the finding. It reinforces the central argument: that the architecture of codified learning is not primarily shaped by operational content but by something more structural and consistent across the IDB.

Temporal Patterns

The share of lessons in the dominant cluster is broadly stable over the period covered by the analysis (Figure 4). From 2018 onward, the dominant cluster accounts for between 92.9 and 97.2 percent of lessons in every year except 2023, where its share drops to 77.0 percent and the unclassified residual rises sharply. This is likely a compositional artifact reflecting the mixture of operations completing that year rather than a genuine shift in what project teams were learning. The years before 2018 show instability, but the volumes are too low for reliable inference. The overall picture is one of stability throughout the period of adequate coverage.

Figure 4. Cluster-Group Shares over Time

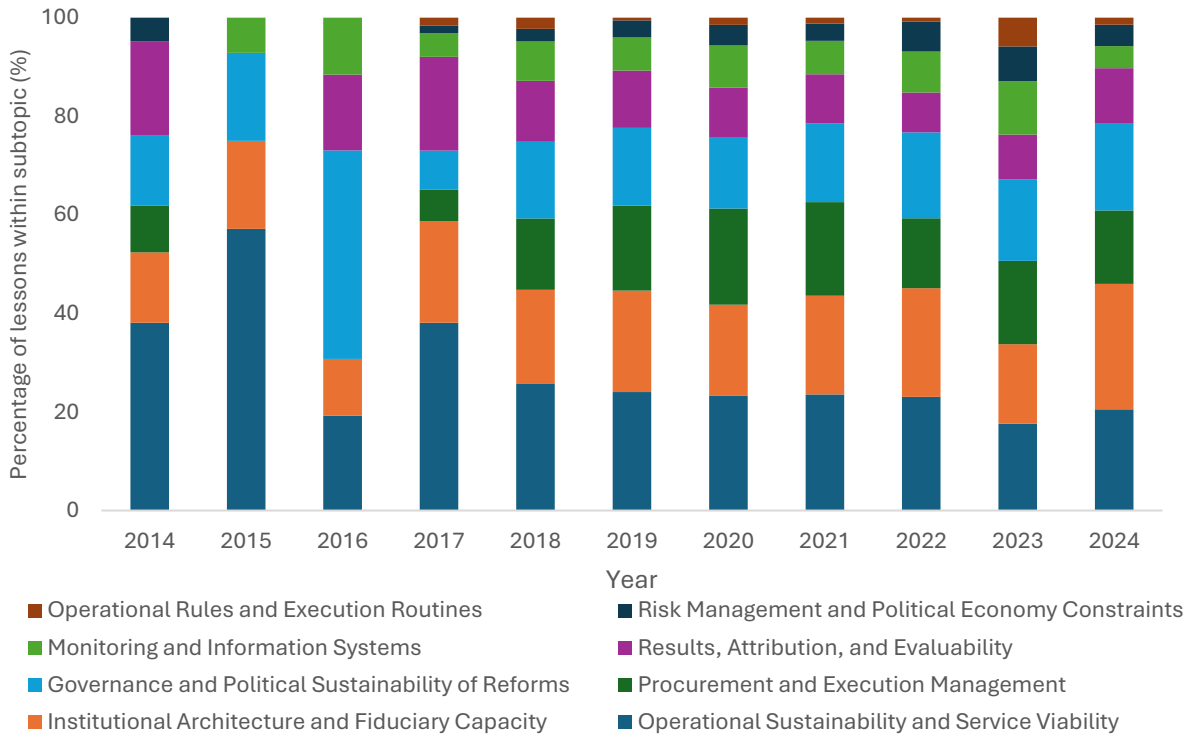


Source: Author's calculations, based on Project Completion Reports published by the Inter-American Development Bank between 2014 and 2024.

Figure 5 shows a gradual shift in the internal composition of the dominant cluster over time. Operational Sustainability and Service Viability declines from 25.8 percent in 2018 to 20.5 percent in 2024, while the two monitoring and evaluation subtopics grow. Monitoring and Information Systems rises from 8.0 percent in 2018 to a peak of 10.8 percent in 2023 before easing to 4.5 percent in 2024, and Results, Attribution, and Evaluability fluctuates between 8.1 and 12.4 percent across the period.

What drives this shift is harder to establish. It may point to genuine changes in operational priorities, or it may reflect changes in how teams were expected to write lessons, rather than what they were actually learning. The data do not distinguish between these possibilities, and without direct evidence on changes in reporting guidelines the temporal results are best read as documentation of a shift in the codified record rather than an indication of its cause.

Figure 5. Subtopic Composition within the Dominant Cluster over Time



Source: Author's calculations, based on Project Completion Reports published by the Inter-American Development Bank between 2014 and 2024.

Discursive Structure of Lessons

Beyond their thematic content, lessons vary systematically in how they are written. Two dimensions are of particular interest: whether a lesson uses prescriptive language to make recommendations and whether it names specific implementing individuals or departments as responsible for acting on the recommendations. These dimensions are analytically distinct. A lesson can be highly prescriptive without naming anyone responsible, and it can name an actor without framing the recommendation in obligatory terms.

Across the 2018–24 sample, prescriptive language is the norm: About 62.8 percent of lessons contain at least one normative cue—that is, language indicating a suggestion or recommendation. Actor attribution is considerably less common, appearing in approximately 47.2 percent of lessons. The strictest measure, combining the two, is present in roughly 31.5 percent of lessons. Most lessons contain prescriptive language, but fewer than one in three both name a specific actor and frame the recommendation in obligatory terms.

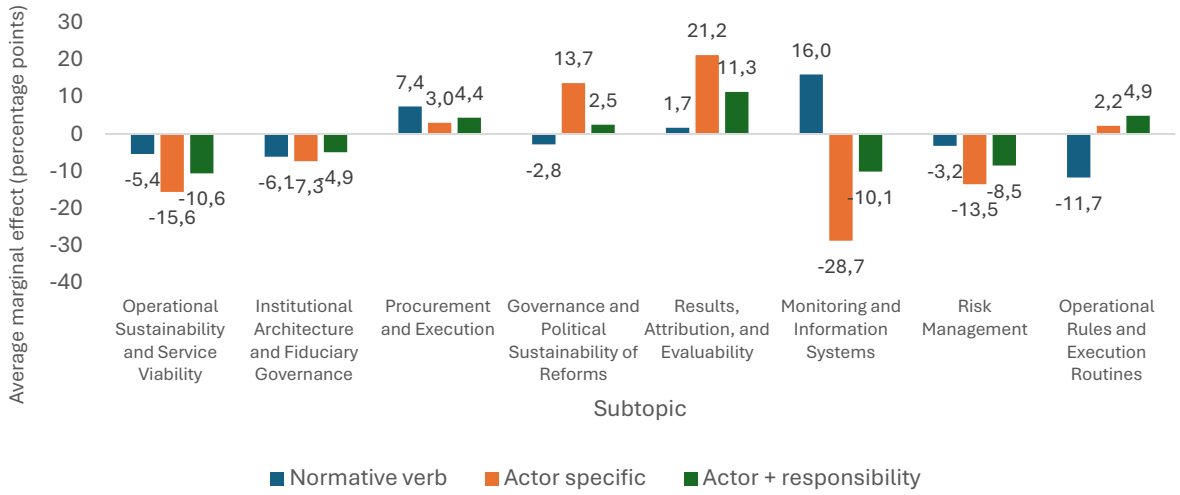
Table 3. Marginal Effects on Discursive Features of Lessons: Dominant Cluster Effect, 2018–24

Variables	(1) Normative verb	(2) Actor (specific)	(3) Actor + responsibility
Dominant cluster	–0.005	0.013	0.069**
	(0.038)	(0.031)	(0.030)
Year FE	Yes	Yes	Yes
Sector FE	Yes	Yes	Yes
Observations	6,456	6,456	6,456

Source: Author's calculations, based on Project Completion Reports published by the Inter-American Development Bank between 2014 and 2024.
Note: Standard errors in parentheses. Clustered SE at operation level. Entries are average marginal effects (percentage-point changes).
* p < 0.10, ** p < 0.05, *** p < 0.01.

Table 3 examines whether belonging to the dominant cluster is associated with different discursive patterns. Dominant cluster membership is not associated with higher prescriptiveness or higher actor attribution overall. Dominant cluster lessons are, however, 6.9 percentage points more likely to combine actor attribution with a normative cue.

Figure 6. Marginal Effects on Discursive Features, by Subtopic



Source: Author's calculations, based on Project Completion Reports published by the Inter-American Development Bank between 2014 and 2024.

Note: Each bar reports the average marginal effect of belonging to the indicated subtopic relative to all other lessons in the dominant cluster, estimated from a logistic regression model with year and sector fixed effects. The sample is restricted to lessons in the dominant cluster with non-missing subtopic assignment, 2018–24 (N = 6,028). Standard errors are clustered at the operation level. Entries are in percentage-point terms. Full results, including standard errors and confidence intervals, are reported in the Appendix.

Three subtopics show distinctive discursive patterns. Operational Sustainability and Service Viability is associated with substantially lower normative language (−5.4 percentage points), lower actor attribution (−15.6 percentage points), and a lower probability of combining actor attribution with normative language (−10.6 percentage points). This pattern is consistent with the subtopic's focus on post-closure outcomes, for which responsibility may be distributed across multiple actors rather than assigned to a single identifiable entity. Risk Management is also associated with significantly lower actor attribution (−13.5 percentage points) and a lower probability of combining actor attribution with normative language (−8.5 percentage points), while its marginal effect on normative language is close to zero. Operational Rules and Execution Routines, by contrast, shows the largest negative marginal effect on normative language of any subtopic (−11.7 percentage points), while its marginal effect on actor attribution is not statistically distinguishable from zero. The estimates for Risk Management and Operational Rules and Execution Routines are based on smaller within-sample sizes and should therefore be interpreted with caution.

The most striking profile belongs to Monitoring and Information Systems, which combines the highest normative language of any subtopic (+16.0 percentage points) with the most negative actor attribution (−28.7 percentage points) and significantly below-average actor + responsibility (−10.1 percentage points). Lessons here are highly prescriptive regarding what monitoring frameworks should look like but do not name who is responsible for building them. Results, Attribution, and Evaluability shows the opposite pattern: no significant normative effect (+1.7 percentage points) but the strongest positive actor attribution of any subtopic (+21.2 percentage points) and above-average actor + responsibility (+11.3 percentage points). Procurement and Execution and Governance and Political Sustainability of Reforms also show positive actor attribution (+3.0 and +13.7 percentage points, respectively), with Governance the stronger and more precisely estimated effect. Institutional Architecture and Fiduciary Governance shows a consistently negative profile across all three outcomes (−6.1, −7.3, and −4.9 percentage points, respectively), similar in structure to Operational Sustainability.

This pattern connects to the temporal shift documented in Figure 5. As Monitoring and Information Systems grows within the dominant cluster, the corpus increasingly produces lessons that specify what good measurement should look like without naming who must deliver it. The contrast with Results and Evaluability is instructive. Both subtopics address measurement problems, but with opposite accountability structures: Monitoring prescribes actions without assigning them, while Results names actors without framing obligations. Whether this reflects genuine differences in how accountability is understood across these problem thematic areas or differences in reporting conventions is a question discussed below.

Lessons Learned on Lessons Learned

This paper set out to examine what has been formally recorded as institutional learning across a decade of IDB Project Completion Reports, and what that record reveals about how the organization codifies experience. The findings point to three conclusions with practical implications.

The first is that codified learning is more concentrated than the portfolio would lead one to expect. More than nine in ten lessons across 26 countries, 16 divisions, and 5 sector departments fall within a single cross-cutting managerial thematic area centered on project design and implementation management. Within that thematic area, the lessons are substantive and coherent; they address operational problems concerning procurement readiness, institutional capacity, sustainability arrangements, governance of reforms, and results measurement. But they do so in terms that are general enough to apply across sectors, which means they also lack the specificity that makes lessons directly actionable in any particular context.

The second is that lessons tend to be prescriptive without consistently assigning responsibility. Most lessons indicate the need for some form of action, but fewer than one in three both name a specific actor and frame the recommendation in obligatory terms. This gap is widest in Monitoring and Information Systems, which is the most prescriptive of any subtopic while also being the least likely to identify who is responsible for acting. A lesson that calls for stronger monitoring systems, clearer reporting arrangements, or better data infrastructure without identifying which unit, team, or counterpart must deliver them may state a broadly desirable course of action without creating a clear basis for assigning responsibility or tracking implementation.

The third is that the overall architecture is stable. The IDB's dominant learning domain has absorbed the overwhelming majority of documented lessons throughout the period of adequate coverage, and the internal composition of that domain has shifted only gradually. While this stability may point to the genuine universality of implementation challenges across the portfolio, it may also stem from the pull of reporting templates, drafting conventions, and institutional expectations that shape what gets written at project closure, independently of what was actually learned during execution.

Viewed from the perspective of the knowledge loop of the IDB Knowledge Model, these findings suggest the main challenge may not be the absence of lessons at the closure and evaluation stage of Bank projects but, rather, the form in which those lessons reenter the institutional learning system. Lessons are being produced, but their ability to inform programming and design depends on whether they preserve sufficient information about the context in which a problem arose, the operational decision that shaped the outcome, the actor with authority to respond, and the point in the project cycle at which action was required. Strengthening these elements would make lessons easier to interpret, retrieve, and apply to future operations.

The findings also point to concrete ways in which lessons could better support operational design. The recurrent themes identified in the corpus could be translated into structured questions for teams preparing new operations. Depending on the nature of the intervention, these questions could examine whether procurement inputs are sufficiently mature, whether executing units have adequate authority and fiduciary capacity, whether maintenance and financing arrangements are viable beyond project closure, whether monitoring systems will be operational early enough to inform implementation, and whether reform measures have the institutional and political support required to endure. The value of the lessons system would, therefore, lie not only in preserving experience but in helping teams identify recurring implementation conditions before they become constraints.

These implications extend beyond the drafting of individual lessons. A stronger feedback mechanism would connect lessons to comparable operations, make relevant evidence visible during programming and design, and allow teams to document how experience gained from other projects influenced operational choices. Clearer stewardship would also be needed to organize accumulated lessons, identify recurring patterns, and determine when a lesson is sufficiently reliable and transferable to inform future decisions. Such mechanisms would help convert a collection of project-level statements into cumulative institutional learning.

The thematic concentration documented in this paper should not be interpreted as evidence that sector-specific knowledge is absent from the institution. More specialized learning may be documented in technical documents, thematic publications, and other knowledge products outside the PCR lessons learned sections analyzed here. The institutional question is, therefore, also one of connection: whether knowledge produced through different instruments is integrated and made accessible at the stages of the operational cycle where teams can use it.

What emerges from this analysis is that the current lessons learned system appears to capture a substantive but relatively concentrated dimension of institutional experience. It documents recurring implementation challenges with considerable consistency and coherence. Its contribution to the knowledge loop could, however, be strengthened by preserving greater contextual specificity, making responsibility more explicit, connecting lessons with complementary sources of sector knowledge with one another, and creating clearer mechanisms through which accumulated experience informs the design and execution of future operations.

References

Argote, L., and E. Miron-Spektor. 2011. "Organizational Learning: From Experience to Knowledge." *Organization Science* 22 (5): 1123–37. https://doi.org/10.1287/orsc.1100.0621.

Bakker, R. M., R. J. DeFillippi, A. Schwab, and J. Sydow. 2016. "Temporary Organizing: Promises, Processes, Problems." *Organization Studies* 37 (12): 1703–19. https://doi.org/10.1177/0170840616655982.

Brady, T., and A. Davies. 2004. "Building Project Capabilities: From Exploratory to Exploitative Learning." *Organization Studies* 25 (9): 1601–21. https://doi.org/10.1177/0170840604048001.

Broome, A., A. Homolar, and M. Kranke. 2018. "Bad Science: International Organizations and the Indirect Power of Global Benchmarking." *European Journal of International Relations* 24 (3): 514–39. https://doi.org/10.1177/1354066117719320.

Buntaine, M. T., B. C. Parks, and B. P. Buch. 2017. "Aiming at Effectiveness: The Politics of Performance in Multilateral Development Banking." *International Organization* 71 (4): 791–823. https://doi.org/10.1017/S0020818317000242.

Campello, R. J. G. B., D. Moulavi, and J. Sander. 2013. "Density-Based Clustering Based on Hierarchical Density Estimates." In *Advances in Knowledge Discovery and Data Mining*, edited by J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu. *PAKDD 2013. Lecture Notes in Computer Science* 7819: 160–72. Springer.

Christensen, T., and P. Lægreid. 2017. "Performance and Accountability—A Theoretical Discussion and Empirical Assessment." *Public Administration* 95 (4): 857–72. https://doi.org/10.1111/padm.12319.

Cohen, W. M., and D. A. Levinthal. 1990. "Absorptive Capacity: A New Perspective on Learning and Innovation." *Administrative Science Quarterly* 35 (1): 128–52. https://doi.org/10.2307/2393553.

Cornelissen, J., R. Durand, P. C. Fiss, J. C. Lammers, and E. Vaara. 2015. "Putting Communication Front and Center in Institutional Theory." *Academy of Management Review* 40 (1): 10–27. https://doi.org/10.5465/amr.2014.0271.

DiMaggio, P. J., and W. W. Powell. 1983. "The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields." *American Sociological Review* 48 (2): 147–60. https://doi.org/10.2307/2095101.

Duffield, S., and S. J. Whitty. 2015. "Developing a Systemic Lessons Learned Knowledge Model for Organisational Learning through Projects." *International Journal of Project Management* 33 (2): 311–24. https://doi.org/10.1016/j.ijproman.2014.07.004.

Duffield, S., and S. J. Whitty. 2016. "Application of the Systemic Lessons Learned Knowledge Model for Organisational Learning through Projects." *International Journal of Project Management* 34 (7): 1280–93. https://doi.org/10.1016/j.ijproman.2016.07.001.

Honig, D. 2018. *Navigation by Judgment: Why and When Top-Down Management of Foreign Aid Doesn't Work*. Oxford University Press.

Huber, G. P. 1991. "Organizational Learning: The Contributing Processes and the Literatures." *Organization Science* 2 (1): 88–115. https://doi.org/10.1287/orsc.2.1.88.

IEG (Independent Evaluation Group). 2015. *Learning and Results in World Bank Operations: Toward a New Learning Strategy, Evaluation 2*. Washington, DC: World Bank.

Inter-American Development Bank, IDB Invest, and IDB Lab. 2026. "IDB Group Knowledge Annual Report 2025: Turning Knowledge into Impact." https://doi.org/10.18235/0013948.

Koporcic, N., D. Sjödin, M. Kohtamäki, and V. Parida. 2025. "Embracing the 'Fail Fast and Learn Fast' Mindset: Conceptualizing Learning from Failure in Knowledge-Intensive SMEs." *Small Business Economics* 64: 181–202. https://doi.org/10.1007/s11187-024-00897-0.

Lee, D. D., and H. S. Seung. 1999. "Learning the Parts of Objects by Non-Negative Matrix Factorization." *Nature* 401 (6755): 788–91.

Levitt, B., and J. G. March. 1988. "Organizational Learning." *Annual Review of Sociology* 14:319–40. https://doi.org/10.1146/annurev.so.14.080188.001535.

Lis, A. 2012. "Positive Organizational Behaviors as the Key Success Factors for Lessons Learned Systems: The Case of Military Organizations." *Journal of Positive Management* 3 (1): 82–93. https://doi.org/10.12775/JPM.2012.006.

Mahfouz, A., A. Labib, S. Hadleigh-Dunn, and M. Gentile. 2019. "Operationalizing Learning from Rare Events: Framework for Middle Humanitarian Operations Managers." *Production and Operations Management* 28 (9): 2323–37. https://doi.org/10.1111/poms.13054.

McInnes, L., J. Healy, and J. Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." *arXiv preprint arXiv:1802.03426*.

Meyer, J. W., and B. Rowan. 1977. "Institutionalized Organizations: Formal Structure as Myth and Ceremony." *American Journal of Sociology* 83 (2): 340–63. https://doi.org/10.1086/226550.

Moretti, F., and D. Pestre. 2015. "Bankspeak: The Language of World Bank Reports, 1946–2012." Literary Lab Pamphlet 9. Stanford Literary Lab. https://litlab.stanford.edu/LiteraryLabPamphlet9.pdf.

Moynihan, D. P. 2008. *The Dynamics of Performance Management: Constructing Information and Reform*. Georgetown University Press.

Moynihan, D. P., and N. Landuyt. 2009. "How Do Public Organizations Learn? Bridging Cultural and Structural Perspectives." *Journal of Public Administration Research and Theory* 19 (4): 819–40. https://doi.org/10.1093/jopart/mun025.

Nonaka, I. 1994. "A Dynamic Theory of Organizational Knowledge Creation." *Organization Science* 5 (1): 14–37. https://doi.org/10.1287/orsc.5.1.14.

Organisation for Economic Co-operation and Development (OECD). 2021. *Applying Evaluation Criteria Thoughtfully*. OECD Publishing. https://doi.org/10.1787/543e84ed-en.

Overman, S., and T. Schillemans. 2022. "Toward a Public Administration Theory of Accountability Overload." *Public Administration Review* 82 (6): 1072–84. https://doi.org/10.1111/puar.13507.

Power, M. 1999. *The Audit Society: Rituals of Verification*. Oxford University Press.

Reimers, N., and I. Gurevych. 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–92. Association for Computational Linguistics.

Roca, M. E., D. Zepeda, and L. Á. García Ferro. 2025. *Guide to Leveraging Lessons Learned: Enhancing Operational Effectiveness at the Inter-American Development Bank*. Washington, DC: Inter-American Development Bank. https://doi.org/10.18235/0013773.

Sense, A. J. 2011. "The Project Workplace for Organizational Learning Development." *International Journal of Project Management* 29 (8): 986–93. https://doi.org/10.1016/j.ijproman.2010.05.003.

Schillemans, T., and M. Busuioc. 2015. "Predicting Public Sector Accountability: From Agency Drift to Forum Drift." *Journal of Public Administration Research and Theory* 25 (1): 191–215. https://doi.org/10.1093/jopart/muu024.

Schindler, M., and M. J. Eppler. 2003. "Harvesting Project Knowledge: A Review of Project Learning Methods and Success Factors." *International Journal of Project Management* 21 (3): 219–28. https://doi.org/10.1016/S0263-7863(02)00096-0.

Sillito, J., and M. Pope. 2024. "Learning from Lessons Learned: Preliminary Findings from a Study of Learning from Failure." *Proceedings of the 2024 IEEE/ACM 17th International Conference on Cooperative and Human Aspects of Software Engineering (CHASE)*, 97–102. https://doi.org/10.1145/3641822.3641867.

Sydow, J., L. Lindkvist, and R. DeFillippi. 2020. "Project-Based Organizations and the Duality of Learning." *Research in the Sociology of Organizations* 69:45–68. https://doi.org/10.1108/S0733-558X20200000069004.

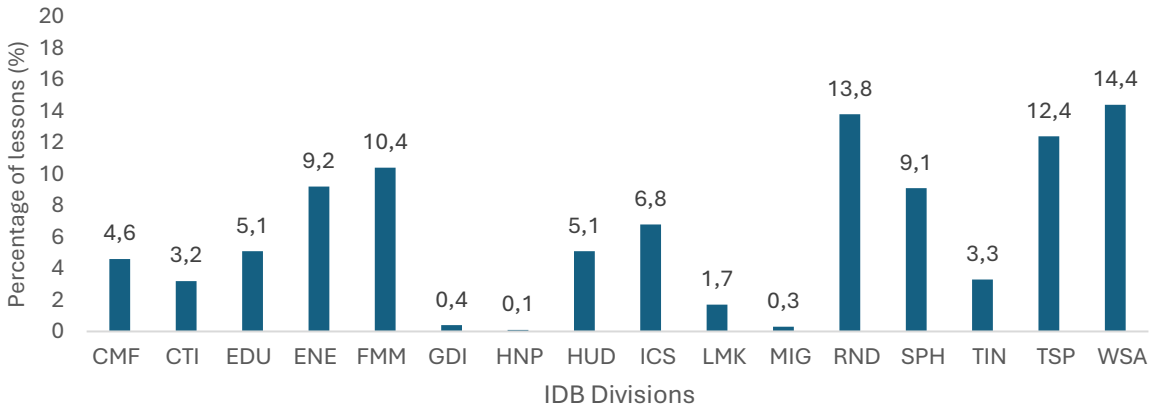
Szulanski, G. 1996. "Exploring Internal Stickiness: Impediments to the Transfer of Best Practice within the Firm." *Strategic Management Journal* 17 (Winter Special Issue): 27–43. https://doi.org/10.1002/smj.4250171105.

Appendix: Methodological Details and Robustness Checks

Corpus Description

Figure A1 reports the distribution of lessons across the IDB's sectoral divisions. The corpus is institutionally concentrated, with three divisions together accounting for approximately 40 percent of all documented lessons. WSA (Water and Sanitation) contributes the largest share at 14.4 percent, followed by RND (Rural Development and Natural Disasters) at 13.8 percent and TSP (Transport) at 12.4 percent. The remaining divisions each contribute between 1 and 8 percent of the total. This concentration partly reflects differences in portfolio size and average lessons per operation across divisions, rather than systematic differences in reporting behavior.

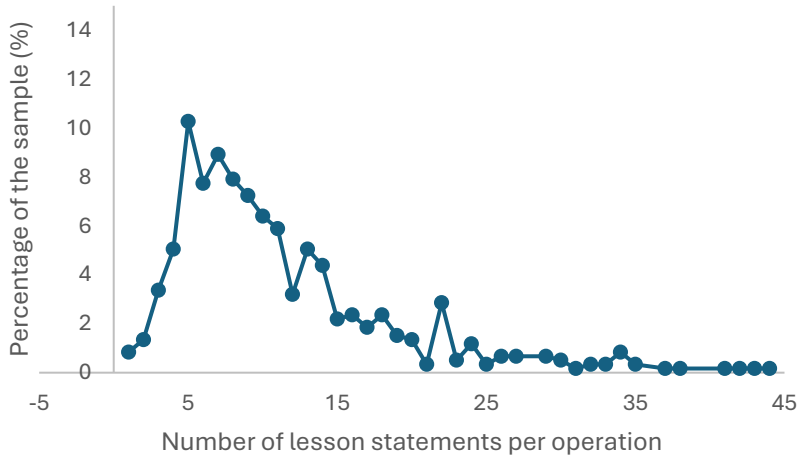
Figure A1. Share of Lessons by IDB Division



Source: Author's calculations, based on Project Completion Reports published by the Inter-American Development Bank between 2014 and 2024.
Note: Division acronyms are as follows: CMF = Connectivity, Markets and Finance; CTI = Competitiveness, Technology and Innovation; EDU = Education; ENE = Energy; FMM = Fiscal Management; GDI = Gender and Diversity; HNP = Health, Nutrition and Population; HUD = Housing and Urban Development; ICS = Institutional Capacity of the State; LMK = Labor Markets; MIG = Migration; RND = Rural Development, Environment and Disaster Risk Management; SPH = Social Protection and Health; TIN = Trade and Investment; TSP = Transport; and WSA = Water and Sanitation.

Figure A2 shows the distribution of lessons by number of lessons per operation. The pattern is highly skewed. A few projects account for a relatively large share of the total corpus, while many operations contribute only one or two lessons. Operations reporting five to ten lessons represent the modal range and together account for a substantial portion of the sample. Operations with five lessons represent 10.3 percent of the corpus, and those with seven to nine lessons each contribute between about 7 and 9 percent. By contrast, operations with only one or two lessons represent less than 2.5 percent combined. At the upper tail, a limited number of operations report more than fifteen lessons, but each individual category accounts for less than 3 percent of the sample.

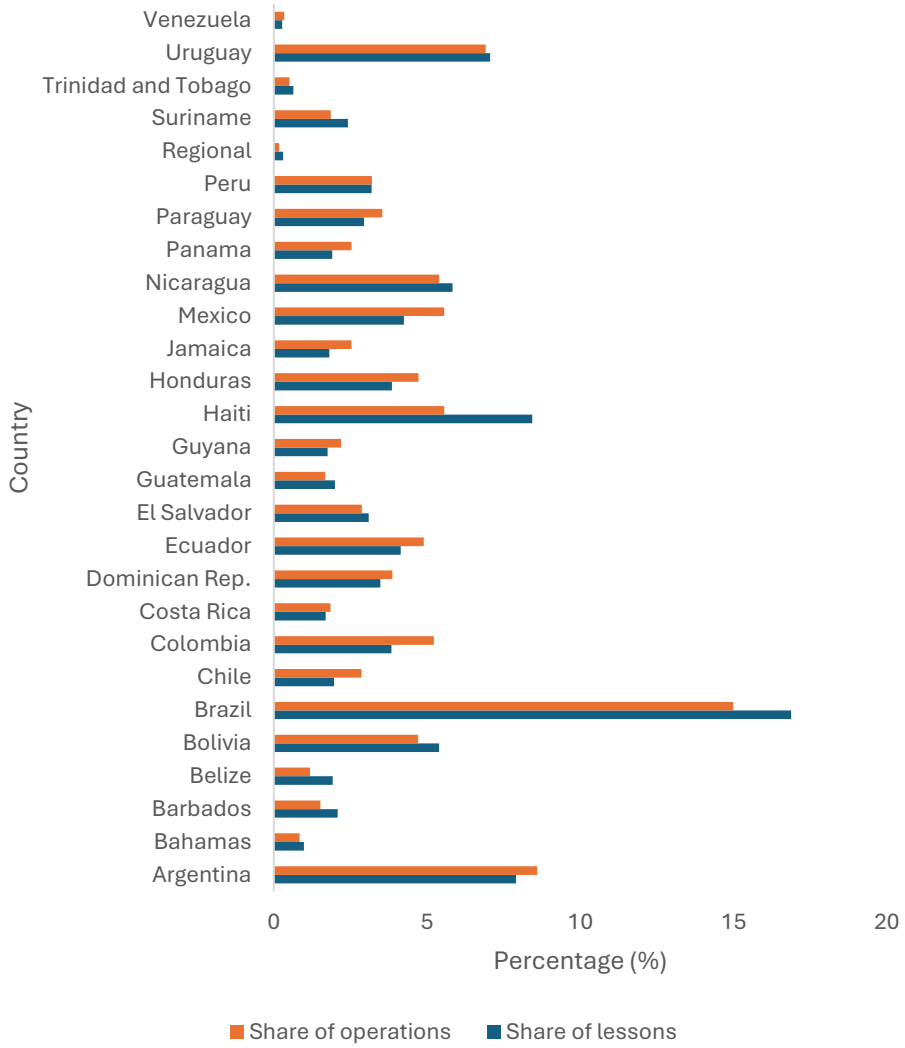
Figure A2. Number of Lessons by Operation



Source: Author's calculations, based on Project Completion Reports published by the Inter-American Development Bank between 2014 and 2024.

Figure A3 presents the geographical distribution of lessons and operations. The corpus is broadly representative of the Bank's country portfolio, but some countries contribute disproportionately to documented lessons. Brazil accounts for the largest share at 16.9 percent, followed by Haiti at 8.4 percent, Argentina at 7.9 percent, and Uruguay at 7.1 percent.

Figure A3. Share of Lessons and Operations by Country



Source: Author's calculations, based on Project Completion Reports published by the Inter-American Development Bank between 2014 and 2024.

Table A1 reports descriptive statistics on lesson length, measured in words, for the full corpus and by sector. Across all 6,619 observations, lessons average 109.3 words, with a median of 93 and a standard deviation of 70.4. The distribution is right-skewed; the 10th percentile falls at 45 words and the 90th at 192, with a few very long observations extending to 1,182 words at the maximum. Length is broadly consistent across sectors. IFD and CSD lessons are somewhat longer on average (123.6 and 124.7 words respectively) than INE lessons (99.6), reflecting differences in the nature of operations. Policy-based and institutional reform operations in IFD and urban operations in CSD tend to require more contextual framing than capital-intensive infrastructure investments. These differences are modest and do not systematically affect the embedding-based analysis. These multilingual language models are relatively robust to moderate variation in text length, and the length

range observed in this corpus, with 80 percent of lessons falling between 45 and 192 words, is well within the range for which they produce stable and reliable semantic representations.

Table A1. Sector Distribution of Lesson Length in Words

Sector	N	Mean	Median	SD	P10	P90	Min	Max
All sectors	6619	109.3	93	70.4	45	192	4	1182
INE	2382	99.6	85	65.1	41.1	176	4	727
IFD	1445	123.6	107	68.1	52	205	5	489
PTI	1345	107.6	93	72.8	41	186	6	989
SCL	1106	109.3	94	69.6	49	181	4	1144
CSD	335	124.7	103	93.1	50.4	215	5	1182

Source: Author's calculations, based on Project Completion Reports published by the Inter-American Development Bank between 2014 and 2024.
Notes: Word counts are computed on cleaned lesson text after encoding correction and whitespace normalization. Sector assignments follow the division-to-sector mapping described in the analytical strategy section, above. Observations with unmapped division codes are excluded from sector-level rows but included in the All-sectors row. INE = Infrastructure and Energy; IFD = Institutions for Development; PTI = Productivity, Trade and Innovation; SCL = Social Sector; and CSD = Climate Change and Sustainable Development.

Corpus Coverage by Year

Table A2 reports the number of lessons and operations by approval year across the full 2014–24 window. The corpus is heavily concentrated in the post-2017 period, with the four years from 2014 to 2017 together accounting for only 157 lessons across 13 operations, or 2.4 percent of the total corpus. Annual volumes in the latter period are too low for stable inference: 2014 and 2015 each contribute fewer than 36 lessons from one to three operations, and even 2017, the highest pre-2018 year, contributes only 66 lessons from six operations. The sharp increase in 2018 to 713 lessons from 85 operations, or 10.8 percent of the corpus, marks the beginning of the high-coverage period and most likely reflects a change in PCR submission rates or reporting requirements rather than in the underlying operations. From 2018 onward, annual lesson counts are sufficient for stable inference, ranging from 713 to 1,217 lessons per year, and the number of lessons per operation stabilizes between 8.4 and 13.4. The main analysis is, therefore, restricted to the period 2018–24, which covers 97.6 percent of the corpus. Pre-2018 observations are retained in the descriptive analysis but excluded from regression models and discursive feature analyses.

Table A2. Corpus Coverage by Year

Year	N_lessons	N_operations	Lessons_per_operation	Share_of_corpus_pct	Cumulative_share_pct
2014	35	1	35	0.5	0.5
2015	30	3	10	0.5	1
2016	26	3	8.7	0.4	1.4
2017	66	6	11	1	2.4
2018	713	85	8.4	10.8	13.2
2019	920	101	9.1	13.9	27.1
2020	989	90	11	14.9	42
2021	1217	91	13.4	18.4	60.4
2022	933	79	11.8	14.1	74.5
2023	958	78	12.3	14.5	89
2024	732	61	12	11.1	100.1
Total	6619	594	11.1	100	100

Source: Author's calculations, based on Project Completion Reports published by the Inter-American Development Bank between 2014 and 2024.
Notes: N_operations counts unique operation IDs (oper_num) with at least one lesson in the given year. Lessons per operation is the ratio of lessons to operations within the year. Share of corpus and cumulative share are computed over the full 6,619-observation sample. The cumulative share in the Total row reflects rounding of the annual shares.

Model and Hyperparameter Choices

The embedding, dimensionality reduction, and clustering steps each involve parameter choices that affect the resulting partition. Table A3 documents all key choices and the rationale for each.

The baseline embedding model is paraphrase-multilingual-MiniLM-L12-v2, a multilingual sentence transformer, that is, a language model designed to convert text into numerical representations that capture semantic meaning. This model was selected because it produces normalized sentence-level embeddings suitable for similarity-based clustering, performs well on bilingual Spanish–English text without language-specific preprocessing, and generates compact 384-dimensional representations that are computationally efficient for a corpus of this size. Sensitivity to this choice is assessed in Appendix A.11 using LaBSE, an alternative multilingual encoder trained on parallel translation data.

Dimensionality reduction is implemented using UMAP, a technique that preserves the relative proximity of semantically similar observations while reducing noise from high-dimensional representations. Clustering is then performed using HDBSCAN, a density-based algorithm that identifies groups of observations more similar to each other than to the rest of the corpus. Parameter values are chosen to balance the identification of substantively interpretable thematic areas with robustness to small, idiosyncratic groupings.

Table A3. Hyperparameter Choices

Step	Parameter	Value	Rationale
Embedding	Model	paraphrase-multilingual-MiniLM-L12-v2	Multilingual sentence transformer; handles bilingual corpus without language-specific preprocessing
Embedding	Normalization	L2 (unit norm)	Required for cosine similarity to be equivalent to dot product; standard for SBERT models
UMAP	n_neighbors	15	Controls the balance between local and global structure; 15 is the recommended default for corpora of this size and preserves both local clusters and global topology
UMAP	n_components	10	Retains sufficient structure for HDBSCAN to identify density modes while discarding high-dimensional noise; lower values (5) collapsed minority clusters, higher values (15) produced no meaningful change
UMAP	metric	cosine	Appropriate for normalized embeddings; consistent with the similarity measure used at the embedding stage
UMAP	random_state	42	Fixed for reproducibility
HDBSCAN	min_cluster_size	25	Minimum number of observations required to form a cluster; set to exclude very small groupings that are unlikely to represent substantively coherent learning thematic areas at corpus scale
HDBSCAN	min_samples	10	Controls the conservativeness of cluster assignment; lower values increase cluster membership but reduce cohesion of minority clusters; 10 produces stable results across the range 5–15
HDBSCAN	metric	euclidean	Applied to the UMAP-reduced space, where Euclidean distance is appropriate; cosine distance is applied at the embedding stage prior to reduction

Source: Author's elaboration, based on the embedding, dimensionality reduction, and clustering specifications used in the analysis.

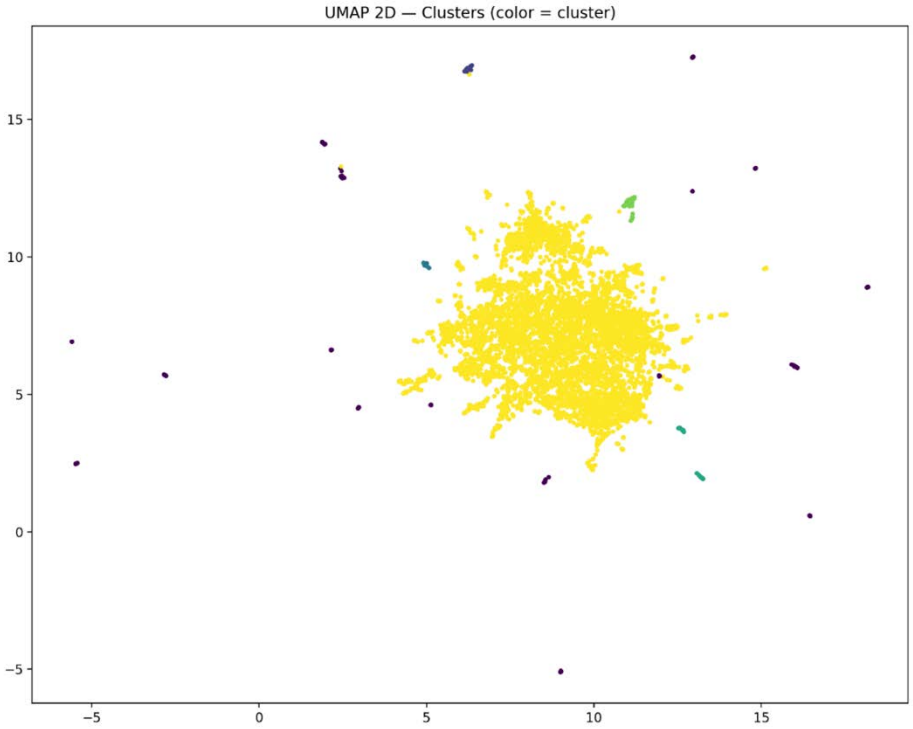
Visual Validation

Figure A4 presents a two-dimensional UMAP projection of the full corpus, with observations colored by cluster assignment. UMAP is a visualization technique that places lesson statements with similar meaning closer together in a low-dimensional space. The figure is generated by applying UMAP with n_components = 2 directly to the original sentence embeddings. This differs from the clustering procedure, which uses a 10-dimensional reduction. The figure, therefore, provides an intuitive visualization of the overall semantic structure of the corpus rather than an exact representation of the space in which clusters are identified.

The dominant cluster (shown in the largest color region) occupies a broad and continuous area of the projection. This pattern is consistent with its interpretation as a wide thematic area encompassing diverse but related forms of implementation learning rather than a narrowly defined topic. The four minority clusters appear as more compact concentrations at the periphery of this dominant region, with relatively limited overlap given their size.

Observations not assigned to any cluster (noise points, cluster = −1) are dispersed across the projection rather than forming a coherent grouping. This suggests they do not represent an additional latent thematic area but instead correspond to lessons whose content overlaps with multiple areas of institutional learning.

Figure A4. UMAP 2D Projection of Lesson Corpus, Colored by Cluster



Source: Author's elaboration based on the lesson corpus and clustering results.
Notes: Two-dimensional UMAP projection computed independently from the 10-dimensional reduction used for clustering, with identical hyperparameters except n_components=2. Colors correspond to HDBSCAN cluster assignments. Each point is one lesson statement (N = 6,619).

Semantic Distance Validation

Table A4 reports mean intracluster cosine distances for each substantive cluster, alongside a random pairs benchmark computed from 1,000 randomly sampled lesson pairs drawn from the full corpus. Cosine distance is a measure of semantic dissimilarity between lesson statements based on their sentence embeddings: Values close to zero indicate very similar content, while values closer to one indicate greater thematic dispersion.

The four minority clusters display substantially lower intracluster distances, ranging from 0.07 for the Integrated Tourism Development cluster to 0.40 for the Integrating Gender cluster. This indicates lessons within these clusters are semantically concentrated around relatively narrow topical cores. By contrast, the dominant cluster has a mean intracluster distance of 0.52, nearly identical to the random-pairs benchmark of 0.53. This pattern indicates the dominant cluster does not represent a tightly bounded topic but, rather, a broad thematic area encompassing diverse forms of implementation-related learning.

This result is consistent with the internal differentiation documented in the section "The Dominant Cluster," above, which identifies multiple coherent subthemes within the dominant thematic area. It is also consistent with the alternative partitioning results presented in Table A.10, where k-means

clustering subdivides the dominant cluster into smaller groupings rather than eliminating it. Taken together, these findings suggest the appropriate validation criterion for the dominant cluster is not semantic compactness but, rather, the stability of its boundaries across alternative models and specifications.

Table A4. Mean Intracluster Cosine Distance by Cluster

Cluster	N	Mean cosine distance
Integrated Tourism Development and Destination Governance	44	0.07
Integrity and Oversight in Procurement	32	0.32
Technology Systems and Data Capacity	43	0.39
Integrating Gender into the Design of Operations	73	0.40
Project Design and Implementation Management (dominant)	6,172	0.52
Random pairs baseline	—	0.53

Source: Author's calculations based on L2-normalized sentence embeddings from the lesson corpus.
Notes: Cosine distances computed on L2-normalized sentence embeddings. Intra-cluster distances are computed on a random sample of up to 1,000 observations per cluster using the upper triangle of the pairwise distance matrix. The random pairs baseline draws 1,000 observations uniformly from the full corpus regardless of cluster assignment (random_state=42).

Selection of Number of Topics

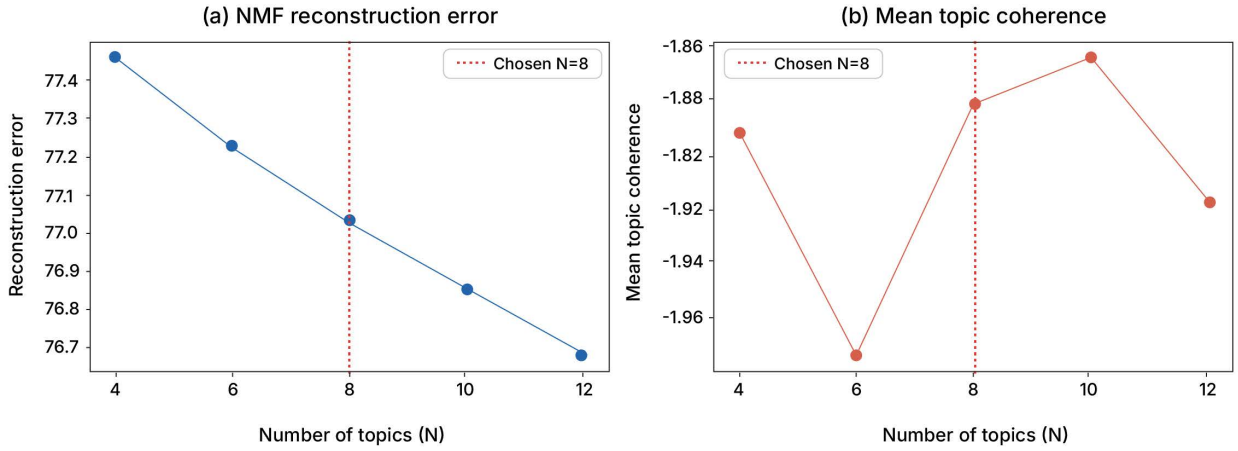
The number of subtopics within the dominant cluster was determined by estimating a series of Nonnegative Matrix Factorization (NMF) models across alternative specifications (N = 4, 6, 8, 10, 12). NMF is a topic-modeling technique that identifies recurring patterns of word use across documents and groups observations into interpretable thematic components. Model selection was based on two complementary diagnostics.

The first is reconstruction error, which captures how closely the simplified topic structure reproduces the original distribution of terms across lesson statements. As expected in factorization models, reconstruction error decreases steadily as the number of topics increases, and therefore does not by itself provide a clear criterion for choosing the optimal specification.

The second diagnostic is topic coherence, which measures the extent to which the most representative terms of each topic tend to appear together in the same lesson statements. Higher coherence indicates a topic reflects a meaningful thematic pattern rather than an arbitrary statistical grouping. Figure A5 reports both diagnostics across the range of N values considered.

Topic coherence is highest for specifications with eight and ten topics, while lower-dimensional (N = 6) and higher-dimensional (N = 12) solutions display weaker internal consistency. The preferred specification of eight subtopics was selected because it achieves near-maximal coherence while maintaining parsimony. Inspection of the 10-topic solution revealed that the additional topics largely subdivided existing themes without adding substantively distinct areas of institutional learning. Under the 8-topic specification, each subtopic collects a sufficiently large number of observations and displays a distinct and interpretable vocabulary, with the exception of subtopic 7, discussed below.

Figure A5. NMF Diagnostics: Reconstruction Error and Topic Coherence across N



Source: Author's calculations, based on NMF models estimated on the dominant cluster.
Notes: Reconstruction error is the NMF objective value (Frobenius norm). Mean topic coherence is computed as the average pairwise UMass-style log co-document frequency score across the top 10 terms of each topic. Higher (less negative) coherence indicates greater within-topic term cooccurrence. Red dashed line marks the chosen specification (N = 8).

Top Terms by Subtopic

Table A5 reports the most representative terms associated with each of the eight subtopics identified within the dominant learning thematic area. These terms are derived from the NMF topic model and reflect recurring vocabulary patterns that characterize each thematic grouping. Subtopic labels used in the main text were assigned inductively through qualitative inspection of these term lists rather than imposed a priori. The table, therefore, provides a transparent basis for assessing whether the resulting thematic interpretation corresponds to substantively meaningful areas of institutional learning.

Subtopic 0 (Operational Sustainability and Service Viability) is characterized by vocabulary related to social services, maintenance, community engagement, health, and environmental sustainability. These lessons focus on whether project-supported services remain operational and socially appropriated after closure, highlighting challenges in long-term functionality and beneficiary ownership.

Table A5. Top 15 NMF Terms by Subtopic

Subtopic	Label	Top 15 terms
0	**Operational Sustainability and Service Viability**	social [social], servicios [services], mantenimiento [maintenance], comunidades [communities], salud [health], población [population], ambiental [environmental], acciones [actions], recursos [resources], gestión [management], sociales [social], sistemas [systems], sostenibilidad [sustainability], intervenciones [interventions], comunicación [communication]
1	**Institutional Architecture and Fiduciary Capacity**	gestión [management], procesos [processes], personal [staff], unidad [unit], equipo [team], adquisiciones [procurement], capacidad [capacity], adquisición [procurement], ejecutora [executing], capacitación [training], unidad ejecutora [executing unit], unidades [units], procedimientos [procedures], institucional [institutional], equipos [teams]
2	**Procurement and Execution**	obras [works], construcción [construction], contratos [contracts], obra [work], contrato [contract], diseños [designs], diseño [design], no [no], costos [costs], estudios [studies], infraestructura [infrastructure], supervisión [supervision], licitación [bidding], retrasos [delays], calidad [quality]
3	**Governance and Political Sustainability of Reforms**	reformas [reforms], diseño [design], política [policy], coordinación [coordination], gobierno [government], políticas [policies], apoyo [support], no [no], reforma [reform], medidas [measures], actores [actors], cambios [changes], país [country], sector [sector], institucional [institutional]
4	**Monitoring and Information Systems**	evaluación [evaluation], información [information], monitoreo [monitoring], seguimiento [follow-up], monitoreo evaluación [monitoring and evaluation], datos [data], sistema [system], impacto [impact], evaluación impacto [impact evaluation], no [no], plan [plan], sistemas [systems], evaluaciones [evaluations], informes [reports], plan monitoreo [monitoring plan]
5	**Results, Attribution, and Evaluability**	indicadores [indicators], no [no], medir [measure], productos [outputs], objetivos [objectives], matriz [matrix], base [baseline], impacto [impact], medición [measurement], indicadores impacto [impact indicators], número [number], indicadores no [indicators no], línea [line], específicos [specific], metas [targets]
6	**Risk Management**	riesgos [risks], sesgo [bias], impacto [impact], gestión riesgos [risk management], análisis riesgos [risk analysis], análisis [analysis], mitigación [mitigation], actores baja [actors low], baja gobernabilidad [low governability], sesgo estratégico [strategic bias], hechos actores [actor facts], operativo riesgos [operational risks], sesgo operativo [operational bias], impacto sesgo [impact bias], hechos [facts]
7	**Operational Rules and Execution Routines**	regulación operativa [operational regulation], años [years], profisco [PROFISCO], regulación [regulation], operativa [operational], coordinador [coordinator], ii [II], profisco ii [PROFISCO II], especificaciones técnicas [technical specifications], referencia especificaciones [specifications reference], términos principales [main terms], principales referencia [main reference], técnicas [technical], especificaciones [specifications], operativas [operational]

Source: Author's elaboration, based on the NMF model estimated on the dominant cluster.
Notes: Terms are the top 15 TF-IDF-weighted features from the NMF component matrix H for each subtopic, listed in descending order of weight. Because the NMF analysis was conducted on the Spanish-language version of the lesson text, the model-generated terms are in Spanish; English translations are provided in square brackets. The NMF model uses n_components=8, random_state=42, init=nndsvda, and max_iter=400, applied to a TF-IDF matrix with unigrams and bigrams, min_df=5, max_df=0.90, and a combined Spanish-English stopword list with thematic area-specific exclusions. Subtopic labels were assigned by inspection of these term lists.

Subtopic 1 (Institutional Architecture and Fiduciary Capacity) loads strongly on executing-unit terminology, including *unidad ejecutora*, *adquisiciones*, *procedimientos*, and *capacitación*. This pattern reflects lessons concerning the organizational and administrative infrastructure required to manage project execution effectively, including procurement processes, staffing, and institutional capacity development.

Subtopic 2 (Procurement and Execution) is defined by construction and contracting vocabulary, such as *obras*, *contratos*, *diseño*, *costos*, *supervisión*, and *licitación*, alongside references to delays and quality issues. These lessons capture operational experiences related to the technical governance of investment projects, including contract management, cost control, and implementation sequencing.

Subtopic 3 (Governance and Political Sustainability of Reforms) centers on reform-related terminology, including *reformas*, *política*, *coordinación*, *gobierno*, and *actores*. This thematic grouping reflects lessons drawn primarily from policy-based operations, emphasizing the political durability of reforms, interinstitutional coordination, and stakeholder alignment.

Subtopic 4 (Monitoring and Information Systems) is characterized by vocabulary associated with data collection, monitoring processes, and reporting systems, such as *monitoreo*, *evaluación*, *seguimiento*, *datos*, and *sistema*. These lessons point to implementation challenges in generating reliable information on performance, including the development of information systems and institutional routines for tracking results during execution.

Subtopic 5 (Results, Attribution, and Evaluability) focuses on the conceptual architecture of results frameworks rather than on operational monitoring systems. Since its vocabulary (*indicadores*, *matriz*, *línea de base*, *metas*, and *medir*) reflects lessons about how outcomes are defined, measured, and attributed at the design stage, this thematic grouping captures evaluability concerns upstream in the project cycle, distinguishing it from the implementation-focused measurement issues represented in Subtopic 4.

Subtopic 6 (Risk Management) is defined by terminology related to risk analysis and mitigation, including *riesgos*, *gestión riesgos*, *mitigación*, and *sesgo estratégico*. These lessons concern the identification and governance of operational risks, particularly in environments characterized by institutional fragility or political uncertainty.

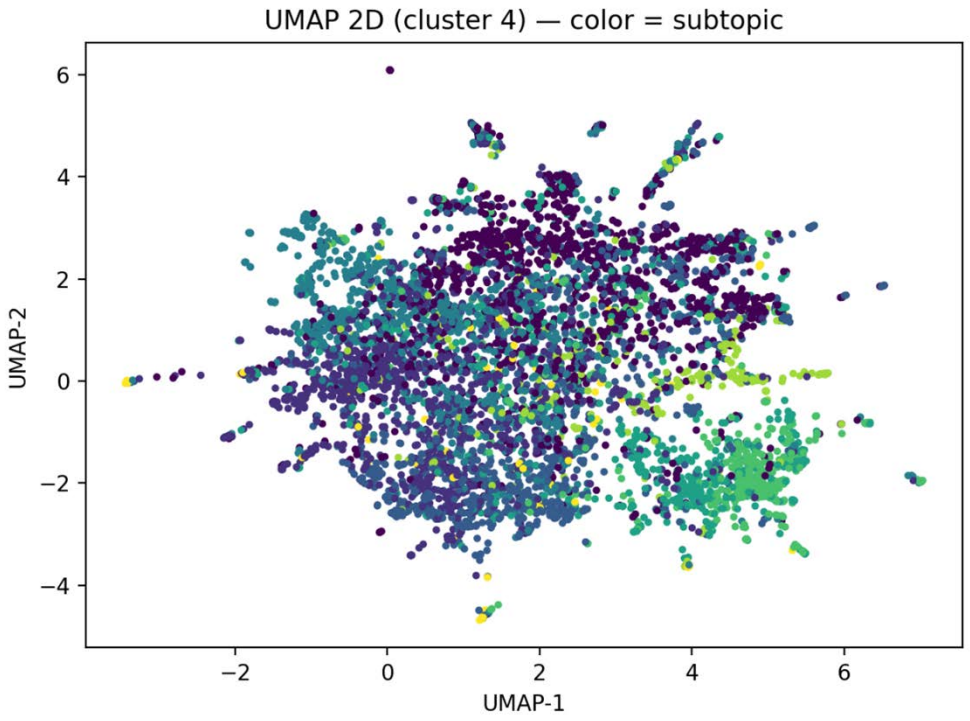
Subtopic 7 (Operational Rules and Execution Routines) is dominated by vocabulary specific to a small number of operations, including references to particular regulatory frameworks, project names, and technical specifications. Unlike the other subtopics, this grouping does not reflect a broad thematic area of institutional learning but instead captures project-specific procedural language. It represents approximately 1 percent of observations within the dominant cluster and should therefore be interpreted with caution.

Taken together, the distribution of representative terms confirms that the dominant learning thematic area is internally differentiated into multiple operational subthemes. Most of these relate to institutional capacity, implementation governance, and results management rather than sector-specific technical design.

Visual Separation of Subtopics

Figure A6 presents a two-dimensional UMAP visualization of the observations belonging to the dominant cluster, with lesson statements colored according to their subtopic assignments. UMAP is a dimensionality-reduction technique that places observations with similar semantic content closer together in visual space, allowing patterns of thematic differentiation to be inspected intuitively.

Figure A6. UMAP 2D Projection of Dominant Cluster, Colored by Subtopic



Source: Author's elaboration, based on sentence embeddings and NMF subtopic assignments.
Note: Two-dimensional UMAP projection of sentence embeddings, restricted to the 6,172 observations in the dominant cluster. Colors correspond to NMF subtopic assignments (0–7). The projection is computed independently from the NMF model and serves as a visual check on subtopic separation in semantic space.

The projection is computed directly from the sentence embeddings of the dominant-cluster observations and is, therefore, independent of the NMF topic model, which identifies subtopics based on patterns of word use. This distinction is important: While the topic model captures differences in vocabulary emphasis, the visualization reflects broader similarities in meaning across lessons.

Subtopics with more specialized and concentrated vocabularies, particularly Subtopic 2 (Procurement and Execution) and Subtopic 3 (Governance and Political Sustainability of Reforms), appear as relatively distinct regions in the projection. The largest subtopics, particularly Subtopics 0 and 1, occupy areas that overlap more substantially, while Subtopic 2 remains comparatively differentiated. This pattern is consistent with their shared focus on implementation-management

challenges and reflects the often gradual rather than discrete boundaries between operational learning thematic areas.

Overall, the visualization supports the interpretation that subtopic assignments capture meaningful variation in the content of lessons within the dominant learning thematic area. The observed degree of overlap also reinforces the characterization of institutional learning as organized around a broad managerial space composed of related but internally differentiated themes.

Thematic Concentration within Sectors

To assess whether the large share of lessons in the dominant cluster reflects a general feature of institutional learning rather than the sectoral composition of the portfolio, Table A6 reports the degree of thematic concentration within each sector. Concentration is measured using Shannon entropy, a statistical indicator of how evenly observations are distributed across categories. In this context, entropy captures whether lessons within a sector are spread across several thematic areas or concentrated in a single one.

Table A6. Within-Sector Thematic Concentration

Sector	N	Share dominant (%)	Normalized entropy
INE	2,382	99.0	0.06
CSD	335	97.9	0.10
SCL	1,106	97.0	0.14
PTI	1,345	91.4	0.30
IFD	1,445	81.4	0.53

Source: Author's calculations based on cluster assignments and sector-level metadata.
Note: Sector assignments follow the division-to-sector mapping described in the analytical strategy section, above. Entropy is computed over the three cluster-group categories (dominant, other, unclassified) and normalized by $H_{max} = \log_2(3) = 1.585$ bits. INE = Infrastructure and Energy; CSD = Climate Change and Sustainable Development; SCL = Social Sector; PTI = Productivity, Trade and Innovation; and IFD = Institutions for Development.

Entropy is calculated over three cluster-group categories (dominant, other, and unclassified) and normalized by the theoretical maximum for three categories. A value close to one, therefore, indicates lessons are evenly distributed across thematic areas, while a value close to zero indicates strong concentration in a single thematic area.

Results show low entropy across all five sectors, ranging from 0.06 in INE to 0.53 in IFD. This confirms the predominance of the dominant cluster is not driven by the aggregate composition of the corpus but is, instead, observed within sectors individually. In practical terms, lessons learned in infrastructure, social protection, climate and sustainable development, productivity and innovation, and institutional capacity operations all display a similar tendency to emphasize implementation-management issues.

Even in IFD, the sector with the lowest share of dominant-cluster observations, more than four out of five lessons fall within the broad managerial learning thematic area. The comparatively higher entropy observed in this sector reflects a larger proportion of unclassified observations rather than

a substantive shift toward minority thematic clusters. This pattern is consistent with the distinct vocabulary of policy-based operations, which may be less easily captured by cluster boundaries estimated on the full corpus, as discussed in section on sectoral composition.

K-Means Comparison

Density-based clustering methods such as HDBSCAN can sometimes identify large connected regions in high-dimensional semantic spaces. To assess whether the dominant cluster reflects a genuine pattern in the data rather than an artifact of the clustering algorithm, robustness checks were conducted using k-means, a partition-based clustering method that forces observations into a predetermined number of groups of roughly similar size.

Table A7. K-Means Robustness Check: Overlap with HDBSCAN Dominant Cluster

k	N in best-match cluster	Precision	Recall	F1	Vocab Jaccard
5	1,755	0.958	0.272	0.424	0.538
8	894	0.974	0.141	0.247	0.379
10	815	0.995	0.131	0.232	0.250
15	617	0.887	0.089	0.161	0.429
20	457	0.985	0.073	0.136	0.290

Source: Author's calculations, based on k-means and HDBSCAN cluster assignments.
Note: Precision is the share of observations in the best-matching k-means cluster that belong to the HDBSCAN dominant cluster. Recall is the share of HDBSCAN dominant cluster observations recovered by the best-matching k-means cluster. F1 is the harmonic mean of precision and recall. Vocab Jaccard is the Jaccard similarity between the top 20 TF-IDF terms of the HDBSCAN dominant cluster and the best-matching k-means cluster. K-means is estimated on the full sentence embedding matrix (not the UMAP-reduced space), with n_init=10 and random_state=42.

Table A7 reports overlap statistics between the HDBSCAN dominant cluster and the closest corresponding k-means cluster across alternative specifications (k = 5, 8, 10, 15, 20). For each value of k, the best-matching cluster is defined as the partition with the highest F1 score, a summary measure that combines precision (how pure the cluster is) and recall (how much of the dominant cluster it captures).

Results show consistently high precision across all specifications (0.887 to 0.995). This indicates the k-means cluster most similar to the dominant cluster is composed almost entirely of the same observations. Recall declines sharply, however, as k increases. This occurs because k-means divides the corpus into partitions of comparable size and, therefore, cannot represent a single thematic area that encompasses the large majority of observations without splitting it into several smaller groups.

This fragmentation pattern is substantively informative. Rather than eliminating the dominant thematic area, alternative specifications subdivide it into multiple high-purity clusters. This is consistent with the interpretation of the dominant cluster as a broad managerial learning space that contains internally differentiated subthemes. Vocabulary overlap between the dominant cluster and its k-means counterparts is moderate to high across specifications, confirming the stability of an implementation-focused lexicon, regardless of the clustering method used.

Embedding Model Sensitivity

The baseline analysis represents each lesson statement using paraphrase-multilingual-MiniLM-L12-v2, a multilingual sentence-embedding model that converts text into numerical vectors capturing semantic similarity. To assess whether the identification of a dominant thematic area depends on the specific geometric properties of this model rather than the underlying content of the corpus, the full analytical pipeline is replicated using LaBSE (Language-Agnostic BERT Sentence Encoder), an alternative multilingual encoder.

LaBSE differs substantially from the baseline model in both architecture and training strategy. While the baseline model is trained on paraphrase pairs to recognize semantically equivalent sentences, LaBSE is based on BERT (Bidirectional Encoder Representations from Transformers) and trained on parallel translation data across more than 100 languages. These differences provide a stringent robustness check: If the concentrated thematic structure were driven by model-specific representation choices, the clustering results would be expected to change materially under the alternative encoder.

Despite these differences, both models identify a single dominant cluster accounting for a nearly identical share of observations. Table A8 summarizes the comparison. Overlap between the dominant clusters identified by each model is very high (F1 = 0.966), indicating that the vast majority of lessons assigned to the dominant thematic area under the baseline specification are assigned to it under the alternative encoder, as well. Agreement in representative vocabulary is similarly strong: 19 of the 20 most distinctive terms are shared across models.

Taken together, these results indicate that the highly concentrated thematic structure documented in the main analysis reflects a substantive feature of the lesson corpus rather than an artifact of the specific embedding model used to represent textual content.

Table A8. Embedding Model Sensitivity: Baseline versus LaBSE

	MiniLM (baseline)	LaBSE (alternative)
Model	Paraphrase-multilingual-MiniLM-L12-v2	Sentence transformers/LaBSE
Training objective	Paraphrase pairs	Translation ranking, parallel corpora
Embedding dimension	384	768
Dominant cluster N	6,172	6,288
Dominant cluster share	93.2%	95.0%
F1 (overlap between dominant clusters)	—	0.966
Precision	—	0.957
Recall	—	0.975
Vocab Jaccard (top 20 TF-IDF terms)	—	0.905
Shared top terms	—	19 of 20

Source: Author's calculations, based on the MiniLM and LaBSE embedding specifications.
Notes: Both models use identical UMAP and HDBSCAN hyperparameters. F1, precision, and recall measure overlap between the HDBSCAN dominant cluster under each model. Vocab Jaccard is computed over the top 20 TF-IDF terms of each model's dominant cluster. The only term appearing in MiniLM but not LaBSE is "*operaciones**"; the only term appearing in LaBSE but not MiniLM is "*capacidad,**"

Alternative HDBSCAN Parameters

The baseline clustering specification uses a minimum cluster size of 25 observations (min_cluster_size = 25) and a conservativeness parameter of min_samples = 10, applied to the 10-dimensional UMAP representation of the lesson corpus. To assess sensitivity to this choice, the model was reestimated using a lower minimum cluster size (min_cluster_size = 15), holding all other parameters constant.

Under this alternative specification, the algorithm identifies a much larger number of clusters (71) and classifies nearly half of all observations as noise (47.0 percent), compared to five clusters and 3.9 percent noise under the baseline. No single dominant thematic grouping emerges. The largest cluster accounts for only 6.8 percent of observations, and lessons are distributed across many small clusters.

This pattern reflects a well-known feature of density-based clustering methods. When the minimum cluster size threshold is reduced, the algorithm begins to detect finer local variations in the data, effectively subdividing broader thematic regions into smaller groupings. In this case, the dominant learning thematic area identified in the baseline specification is decomposed into smaller clusters that correspond to more specific types of implementation experience.

The resulting fragmentation is consistent with the internal subtopic structure documented in the section on the dominant cluster, above, and with the k-means robustness results presented in Appendix A.10. These exercises together suggest the dominant cluster represents a broad but internally differentiated semantic thematic area rather than an artificial aggregation produced by the clustering algorithm.

The baseline parameterization is, therefore, preferred because it captures this thematic area at a level of aggregation aligned with the research objective, which is to identify major patterns in institutional learning while still allowing substantively distinct minority clusters to emerge as separate thematic groupings.

Full Regression Output: Models A and B

Table A9 reports the estimated effects of all covariates from the two main logistic regression specifications discussed in Section 5.5. Results are presented as average marginal effects (AMEs), which indicate how the probability of a discursive feature (such as the use of normative language or the attribution of responsibility) changes when a given explanatory variable varies, holding other factors constant. Effects are expressed in percentage-point terms to facilitate interpretation.

Table A9. Full Regression Output: Average Marginal Effects, Models A and B, 2018–24

	Model A			Model B		
	AME	**SE**	**p**	**AME**	**SE**	**p**
Normative verb						
Dominant cluster	−0.009	0.035	0.786	−0.005	0.035	0.893
Year trend/Year FE	✓			✓		
Sector FE	✓			✓		
N	6,456			6,456		
Actor (specific)						
Dominant cluster	0.023	0.029	0.440	0.013	0.032	0.687
Year trend/Year FE	✓			✓		
Sector FE	✓			✓		
N	6,456			6,456		
Actor + responsibility						
Dominant cluster	0.074**	0.028	0.008	0.069**	0.029	0.017
Year trend/Year FE	✓			✓		
Sector FE	✓			✓		
N	6,456			6,456		

Source: Author's calculations, based on logistic regression models using lesson-level data from Project Completion Reports published by the Inter-American Development Bank between 2014 and 2024.
Notes: Entries are average marginal effects (AME) at the overall mean, in percentage-point terms. Standard errors are clustered at the operation level. Model A includes a continuous year trend. Model B includes year fixed effects (preferred specification). Sector fixed effects use Infrastructure and Energy (INE) as the omitted category. Full covariate-level output, including all year and sector fixed effect terms, is available in appendix_e1_full_regression.csv.
*** p < 0.01, ** p < 0.05, * p < 0.10.

Model A includes a linear time trend and sector fixed effects. Model B replaces the linear trend with year fixed effects and is the preferred specification reported in Table 3 of the main text. Both models are estimated on the 2018–24 sample (N = 6,456), with standard errors clustered at the operation level to account for correlation among lessons originating from the same project.

The coefficient on the dominant cluster indicator is the main parameter of substantive interest. Year and sector fixed effects are included to control for broader contextual influences on discursive patterns. In particular, year effects capture common shifts in reporting practices or institutional priorities over time, while sector effects account for systematic differences in writing conventions across operational areas. Individual coefficients for these controls are not discussed in the main analysis because they serve primarily to isolate the relationship between thematic area and discursive structure.

Per-Subtopic Marginal Effects: Basis for Figure 6

Table A10 reports the estimated effects underlying Figure 6 in the main text. Results are presented as average marginal effects (AMEs), which indicate how the probability of observing a given discursive feature changes when a lesson belongs to a specific subtopic rather than any other subtopic within the dominant learning thematic area.

Table A10. Per-Subtopic Average Marginal Effects on Discursive Features, Dominant Cluster, 2018–24

Subtopic	N	Normative verb		Actor (specific)		Actor + responsibility	
		AME	SE	AME	SE	AME	SE
Operational Sustainability and Service Viability	1,364	−0.054***	0.018	−0.156***	0.018	−0.106***	0.016
Institutional Architecture and Fiduciary Capacity	619	-0.061***	0.022	-0.073***	0.024	-0.049***	0.019
Procurement and Execution	1,018	0.074***	0.021	0.030	0.021	0.044***	0.016
Governance and Political Sustainability of Reforms	972	−0.028	0.022	0.137***	0.021	0.025	0.016
Results, Attribution, and Evaluability	1,220	0.017	0.018	0.212***	0.018	0.113***	0.016
Monitoring and Information Systems	460	0.160***	0.030	−0.287***	0.030	−0.101***	0.027
Risk Management	95	−0.032	0.034	−0.135***	0.034	−0.085**	0.033
Operational Rules and Execution Routines	75	−0.117***	0.045	0.022	0.046	0.049	0.034

Source: Author's calculations, based on separate logistic regression models estimated on lessons assigned to the dominant cluster, 2018–24.
Note: Entries are average marginal effects from separate logistic regressions estimated on the dominant-cluster sample. Effects indicate how the likelihood of a discursive feature differs for lessons belonging to the specified subtopic compared with other lessons in the dominant thematic area. Models include year and sector fixed effects. Standard errors are clustered at the operation level. N reports the number of observations in each subtopic.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Each row corresponds to a separate logistic regression estimated on the dominant-cluster sample (N = 6,028, 2018–24). In each specification, the key explanatory variable is a binary indicator for membership in the indicated subtopic. Year fixed effects and sector fixed effects are included to account for broader differences in reporting practices across time and operational areas. Standard errors are clustered at the operation level to reflect the fact that multiple lessons may originate from the same project.

The reported marginal effects, therefore, capture how discursive patterns vary across different types of implementation-related learning after controlling for temporal and sectoral influences. The reference category in each regression is the set of all other lessons within the dominant cluster.

To explore more IDB publications, visit
https://publications.iadb.org