Public Disclosure Authorized

Public Disclosure Authorized

Policy Research Working Paper 10532

Minding the Gap

Aid Effectiveness, Project Ratings and Contextualization

Diana Goldemberg
Luke Jordan
Thomas Kenyon

WORLD BANK GROUP
Independent Evaluation Group
July 2023

Policy Research Working Paper 10532

Abstract

This paper applies novel techniques to long-standing questions of aid effectiveness. It first replicates findings that donor finance is discernibly but weakly associated with sector outcomes in recipient countries. It then shows robustly that donors' own ratings of project success provide limited information on the contribution of those projects to development outcomes. By training a machine learning model on World Bank projects, the paper shows instead that the strongest predictor of these projects' contribution to outcomes is their degree of adaptation to country context, and the largest differences between ratings and actual impact occur in large projects in institutionally weak settings.

This paper is a product of the Independent Evaluation Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at dgoldemberg1@worldbank.org, lukej@mit.edu, and tkenyon@worldbank.org.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Produced by the Research Support Team

Minding the Gap: Aid Effectiveness, Project Ratings and Contextualization

Diana Goldemberg[1], Luke Jordan[2,], Thomas Kenyon[1]

Keywords: aid effectiveness, machine learning, World Bank projects
JEL Codes: O12, O15, O19

⋆The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

⋆⋆The authors would like to acknowledge the valuable contributions of Lily Chu, Stéphane Guimbert, Jed Friedman, Michael Woolcock, Lily Tsai, Daniel Honig, Justin Shenk, Jos Vaessen, Christopher Nelson, Stephen Francis Pirozzi, as well as seminar participantsat the MIT GOV/Lab Seminar and the RMES Results Peer Learning Series. All errors are our own.

Email addresses: dgoldemberg1@worldbank.org (Diana Goldemberg), lukej@mit.edu (Luke Jordan), tkenyon@worldbank.org (Thomas Kenyon)

[1]World Bank Group

[2]MIT

1. Introduction

Numerous empirical studies have investigated whether foreign aid effectively improves development outcomes in recipient countries. This literature has relied mainly on two levels of analysis. One focuses on the aggregate country-level impacts of aid, typically on economic growth or sector outcomes. Another takes a micro-level approach, with development projects as the unit of analysis, most often using donors' own ratings of project outcomes as a measure of effectiveness.

This study bridges those two strands of research by focusing on the association between donor-financed projects and observable development impact, treating project ratings as intermediating variables. This enables us to ask whether project ratings convey information about those outcomes. We use ratings from projects undertaken in 183 developing countries by eight donors since the 1990s, concentrating on a few service delivery sectors with readily available data on beneficiary-level outcomes. We succeed in replicating previous findings of small positive effects of aid on sector outcomes. However, our results suggest that the project ratings convey little information about impact.

The second and more important contribution of this study is to describe and analyze the correlates of projects' contributions to improvements in sector outcomes. Focusing on projects undertaken by the World Bank, for which more granular information and extensive text documentation are available, we use state of the art methods to assess what aspects of a project's production process are associated with stronger outcomes. We first create what are called "text embeddings" of project documents using the latest generation of transformer models,[3] turning texts into numerical representations of their similarity and differences. Then, we train machine learning models to predict projects' sector outcomes, and probe what features of the projects the model paid most attention to. We find that projects with what appear to be high degrees of tailoring to country context and concentration of funds in fewer sectors are associated with stronger outcomes. In doing so, we use newly available data on project characteristics and draw on methodological advances at the intersection of causal inference and econometrics with machine learning. To our knowledge, this is the first attempt to quantify the importance of project contextualization to development effectiveness.

Our findings have actionable implications for the system through which the World Bank and other development institutions evaluate project performance, by offering a cautionary tale against the over-reliance on project ratings as impact metrics. It also has implications for the design and staffing of these projects.

[3]The same class of models that power all state-of-the-art translation, search engines, AI text generators, as well as most plagiarism detectors.

2. Literature and Theory

2.1. Development Effectiveness

A burgeoning literature on aid effectiveness has focused on development projects as the unit of analysis, examining the association between project characteristics and country-level variables on the one hand and project success on the other (for a summary, see Ashton et al. 2023). The most commonly used measure of project success is donors' ratings of project outcomes, which this literature considers a noisy but valid measure of project performance (Denizer, Kaufmann, and Kraay 2013). Explanatory factors for project success cluster around: (i) country characteristics, such as institutional quality, political and economic stability, and regime type; (ii) project characteristics, such as duration, size, sector, and lending instrument; (iii) aspects of project design, such as the clarity of results frameworks and the number of components; and (iv) aspects of project supervision, such as the intensity, timing, and continuity of oversight.

The utility of this literature is limited in two respects: first, in that no study of which we are aware succeeds in explaining more than 30 percent of variance of project outcomes ratings; second, in that there are grounds for questioning the meaning of these ratings in the first place. Individual donor-financed projects often anticipate and rate only local impacts, seldom claiming a linkage to economic growth or country-level outcomes. Yet, the local and national effects of aid projects are linked by definition; the total impact of foreign aid upon sector outcomes must be associated with the cumulative effect of the individual projects. Nevertheless, it is an almost entirely neglected empirical question as to whether these ratings are indeed correlated with the contribution of external financing to development. We address this gap by investigating whether project ratings convey information on observable development impact.

A related literature strand focuses on the aggregate country-level impacts of aid. A few studies have attempted to estimate the relationship between donor financing and sector outcomes in education, energy, health, sanitation, and water. They employ similar strategies: panel data estimation techniques controlling for country-specific effects and potential endogeneity of regressors, with sector outcomes as the dependent variable and aid flows as the central explanatory variable. Mishra and Newhouse (2009) measure the reduction in the infant mortality rate associated with increases in health aid per capita. Birchler and Michaelowa (2016) examine the effect of education aid per capita on net primary school enrolment rates. And Ndikumana and Pickbourn (2017) investigate whether foreign aid to the water and sanitation sector has helped to expand access to water and sanitation services in Sub-Saharan Africa. These studies do not measure or estimate the relevance of the various characteristics of development financing identified in the broader literature beyond its volume; nor do they reflect the contribution of non-lending assistance, such as analytical work, to development outcomes. They do nonetheless provide a point of departure for our analysis, and we replicate their results of small but positive effects of aid on sector outcomes.

2.2. The Role of Projects

Projects are essential vehicles of development assistance, functioning as an intervening and determinative structure between individual interventions and sector outcomes.[4] Their role encompasses not just the financing but also the adaptation of interventions to local context, their implementation, evaluation, and replication through the policy cycle. In doing so, they remediate the 'implementation gap' between what is planned, or conceived on the basis of what might have worked elsewhere, and what is achieved. Variations in project characteristics can also undermine the generation of inferences from randomized control trials and limit the extent to which they may be extrapolated across sites.[5]

It is not unreasonable to expect that projects should cumulatively be associated with sector outcomes. At least since the early 2000s, aid agencies have increasingly combined project financing with technical assistance, strengthening government systems and claiming to improve the quality of a program of government expenditures beyond their own financing. Their effect should be detectable not just on the management of interventions financed by the project but, through institutional spillovers, on other areas of government activity.

Qualitative analyses of the effectiveness of development projects have emphasized goodness of fit with country circumstances as a critical determinant of success. This is partly because, to be successful, any policy has to be not only technically correct but also politically supportable and administratively feasible (Moore 1995); partly because technical correctness itself requires judgment as to the similarity between local context and the factors that determined the outcome of an intervention elsewhere. To the extent that this dance of contextualization occurs in World Bank projects, it is mostly during project preparation. But it has largely been altogether ignored in the quantitative literature on project outcomes, for want of measurability.

The scope of prior analyses has instead been constrained by the ready availability of publicly-disclosed data on aspects of project design and supervision. These are for the most part either only weakly linked in theory to project effectiveness or only rough empirical proxies for theoretically-relevant variables. Thus, for example, while a few studies have attempted to evaluate the contribution of economic analysis or clear results frameworks to project effectiveness, most have restricted themselves to easily observable characteristics like size, duration, sector and sources of financing, often with inconsistent findings. To the extent that they have examined the role of donor agency staff, this has been limited to the project manager, with little attention to other participants in the process. Similarly researchers have depended on country-level measures of institutional quality, even though familiarity with and capacity to implement

[4] For a fuller description of the role of projects and their place within a broader conceptual framework, see Section 1 of Ashton et al. (2023).

[5] For more on implementation gaps see Williams (2019) and for a discussion on similar issues in evidence-based medicine, see Ford and Norrie (2016).

donor-financed projects varies significantly within countries.

We would expect projects to be more effective when they (i) incorporate prior analysis of the conditions under which an intervention functioned elsewhere and awareness of any material differences between it and the context to which it is to be transplanted; (ii) identify any necessary adaptations and resist external pressure towards over-rapid or unthinking replication; (iii) provide the financial and human resources needed to implement the project (Honig 2018). The likelihood of their doing so depends on a process involving not just the project manager, but the leadership and other team members on the donor side, and a project implementation unit generally staffed by civil servants on the government side. All investment projects also depend to a greater or lesser degree on the effectiveness of government procurement and financial management systems. These inputs are often poorly captured by standard indicators of bureaucratic quality (Blum 2014).

2.3. The Project Evaluation Process

The Development Assistance Committee of the Organisation for Economic Co-operation and Development (OECD-DAC) has long spearheaded an agenda on evaluation practice, encouraging analysis of aid effectiveness and results (instead of only inputs and activities), publishing its first set of principles for evaluation of development assistance in 1991. Nowadays, most bilateral and multilateral donors have an established process for evaluating their development effectiveness, aligned with OECD-DAC's normative framework that consists of six evaluation criteria – relevance, coherence, effectiveness, efficiency, impact and sustainability.[6]

At the World Bank, project evaluations are overseen by the Independent Evaluation Group (IEG). IEG rates several aspects of project performance, but the focal metric - reported most saliently to its Board and most commonly used by researchers - is the 'outcome' rating, which assesses whether the project achieved its stated objectives. The ratings are the culmination of a two-stage process: first the project management's own self-evaluation – the Implementation Completion and Results Report (ICR) – and subsequently the ICR Review (ICRR), in about 20 percent of cases followed two years later by a more detailed report, the Project Performance Assessment Report (PPAR), both conducted by IEG. Together these lead to a six-point outcome rating, ranging from highly unsatisfactory to highly satisfactory.

The other seven donors in our database - the Asian Development Bank (ADB), the Global Fund to Fight AIDS, Tuberculosis and Malaria (GFATM), the German Society for International Cooperation (GiZ), the International Fund for Agricultural Development (IFAD), the Japan International Cooperation

[6]Together they describe the desired attributes of interventions: all interventions should be relevant to the country context, coherent with other interventions, achieve their objectives, deliver results in an efficient way, and have positive impacts that last.

Agency (JICA), the German Development Bank (KfW) and the United Kingdom's Department for International Development (DFID) - similarly summarize their self-evaluations in a single 'outcome' rating. Precisely due to their widespread availability and ease of harmonization, the use of such project outcome ratings is prevalent in the aid effectiveness literature.

However, there are several grounds for doubting whether these outcome ratings capture either donor contribution or the likelihood of sustained improvements in development outcomes. First, they are an aggregation of several subratings and therefore mask variance in the contribution of individual components or interventions. Second, by assessing primarily whether a project has achieved its stated objectives, they may encourage project designers to limit their ambition to what can be easily, and sometimes already has been, achieved.[7] Third, they can reorient time horizons to short-term outputs, which are easier to measure within the project life-cycle, over longer-term efforts to resolve core problems (Andrews 2021).

3. Methods

3.1. Research Questions

We seek to advance the existing literature by considering four questions. First, can any general statements be made about the impact of development aid on sector outcomes? Second, does consideration of aggregate project outcome ratings within each sector mediate the terms of that relationship? In other words, do outcome ratings provide information on the relationship between aid and outcomes? Third, can the application of novel econometric and machine learning techniques better detect associations between project characteristics and development outcomes? And finally, what can such methods illuminate about the characteristics of development projects associated with positive sector outcomes?

3.2. Data

Data on sector outcomes and country characteristics are obtained from the World Development Indicators (WDI, World Bank 2021), covering 1990−2015. For official development assistance (ODA) flows, we used the AidData Core

[7]This may explain the inability of previous researchers to explain more than 30 percent of the variance in outcome ratings. If every project defined objectives to achieve the highest rating, the correlation between ratings and project/country characteristics would be zero. IEG does assess the 'Bank's contribution', defined as 'the extent to which the services provided by the World Bank ensured quality at entry of the project and supported effective implementation through appropriate supervision.' Its evaluation of project efficacy, or the extent to which outcomes were achieved, is also required to examine 'whether the achieved outcomes can plausibly be attributed to the government program or project.' But the first focuses largely on compliance with fiduciary and reporting requirements; while the second does not evaluate whether the objectives would have been achieved in the absence of the Bank's involvement.

Research Release (Tierney et al. 2011).[8] For project outcome ratings, we used the Project Performance Dataset (PPD), a consistent six-point project outcome score based on donor-reported outcome data (Honig, Lall, and Parks 2022). We aggregated from AidData to produce measures of total ODA per country-year, and used the PPD in combination with AidData to construct simple and size-weighted averages of ratings for projects completed in a given year.

For World Bank projects we use a scraper to download the three main documents of every project - the project information document (PID), the project appraisal document (PAD), and the implementation completion report (ICR). Respectively, they contain the information available at the beginning and end of project preparation, and at project closure. The documents had already been subject to plain text extraction by the World Bank, and we performed a minimal amount of post-processing to clean the files.[9] For project characteristics, we used the data compiled for Ashton et al. (2023). This includes much that is publicly available via the World Bank's project portal, as well as some internal data on team and management characteristics, and project preparation and supervision steps.

We concentrate on projects in five sectors: health, education, water and sanitation (WASH), energy and fiscal management, due to their combination of prior literature, data availability and size. Together, they account for 35% of aid flows, and one third of PPD projects. When combining sector aid flows and projects with outcomes and country characteristics from the WDI we conducted standard smoothing for noise and minimal interpolation for variables with high missingness.

3.3. Linear Methods: Replication, Sectoral Extension and Ratings

We first replicate the specifications in health, education, and WASH from Mishra and Newhouse (2009) for health, Birchler and Michaelowa (2016) for education, and Ndikumana and Pickbourn (2017) for WASH. Using identical model forms, we are able to reproduce the results in each paper closely (see Appendix A). Since our primary interest is in replicating these models as a baseline, we do not extend them through instrumentation or other techniques, nor do we make more than associative claims based on them.

We then extend this prior analysis by varying controls to check for robustness and adding two sectors: energy, and fiscal management. For energy, we use as a baseline the same controls as in WASH. For fiscal management, we use a limited set of controls, for income level and institutional quality. The general form of the regression equations was as follows:

[8] The AidData is based on the Organisation for Economic Co-operation and Development (OECD) Creditor Reporting System (CRS) donor-reported data, with added granularity on purpose and activity coding.

[9] The scrapped dataset of public documents related to World Bank development projects is openly available at HuggingFace (https://huggingface.co/datasets/lukesjordan/worldbank-project-documentS).

$$Y_{cts} = \gamma_{0,s} + \gamma_X X_{c,t-L,s} + \gamma_W W_{cts} + f_{cs} + f_{ts} + \epsilon_{cts} \quad (1)$$

Where Y_{cts} is the relevant sector s outcome in country c at time t, $X_{c,t-L,s}$ is the relevant aid variables with a lag of L (e.g., volume of aid to sector s), W_{cts} is the set of controls taken for each sector from the cited literature, and f_{cs} and f_{ts} are sector-specific sets of country and period fixed effects. The controls include macro-economic (e.g., GDP per capita), demographic (e.g., youth share of population), and institutional (e.g., Freedom House ratings) measures. The regression tables in Appendix A provide the exact outcome and controls used for each sector.

We conduct our primary extension for each model by adding the project outcome ratings as exogenous variables (in $X_{c,t-L,s}$), varying the specification for robustness. First, we construct a weighted average each year with the weights provided by the relative size of the rated projects in aid flows. We then take the mean and max of those ratings over rolling five year periods. As well as the rating itself (6-point scale) we use a binary variable according to whether the average rating was "moderately satisfactory" and higher, or "moderately unsatisfactory" and below. We also restricted the volume of aid to only that from donors in the PPD, or only World Bank Group projects.

3.4. Machine Learning Methods: Residual Outcomes and Text Embeddings

We then focus our analysis on projects undertaken by the World Bank, for which more granular information and extensive text documentation are available. At the project level, we applied debiased machine learning (Chernozhukov et al. 2018) to estimate treatment parameters for project effects by utilizing a linear model to partial out fixed effects and controls, then utilizing linear and non-linear models to estimate the residual using only project-level characteristics.

We denote by $\bar{Y}_{cts}$ the predicted value in country c at time t for the relevant outcome in sector s, estimated using only the controls and fixed effects in Equation 1. In other words, $\bar{Y}_{cts} = \gamma_{0,s} + \gamma_W W_{cts} + f_{cs} + f_{ts}$. We then removed this prediction to generate, in each sector, in a residual term $\tilde{Y}_{cts} \equiv Y_{cts} - \bar{Y}_{cts}$. Together:

$$\tilde{Y}_{cts} = Y_{cts} - (\gamma_{0,s} + \gamma_{Ws} W_{cts} + f_{cs} + f_{ts}) \quad (2)$$

The coefficients in equation 2 are estimated independently for each sector, and the resulting $\tilde{Y}_{sct}$ are each residual terms (and hence normalized scalar values). These residual outcomes are then the targets for project-level prediction.

We then extend our analysis to encompass numerical representations of text related to the project that, we argue, capture the degree to which project content is tailored to country and sector context (see the Appendix B for details of their construction). These representations are known as "text embeddings". In theory, such embeddings can be very simple: for example, a vector representing counts of key words in a document is an embedding. The embeddings

we utilize are several orders of magnitude more powerful than such counts, or similar statistical measures of topic frequency, because they capture not only the relative presence of key words and terms, but the interrelationship among words. These embeddings capture not only *what* language is used but *how* it is used. The same word in different parts of a block of text, or surrounded by different language, will be embedded differently in the high-dimensional space.

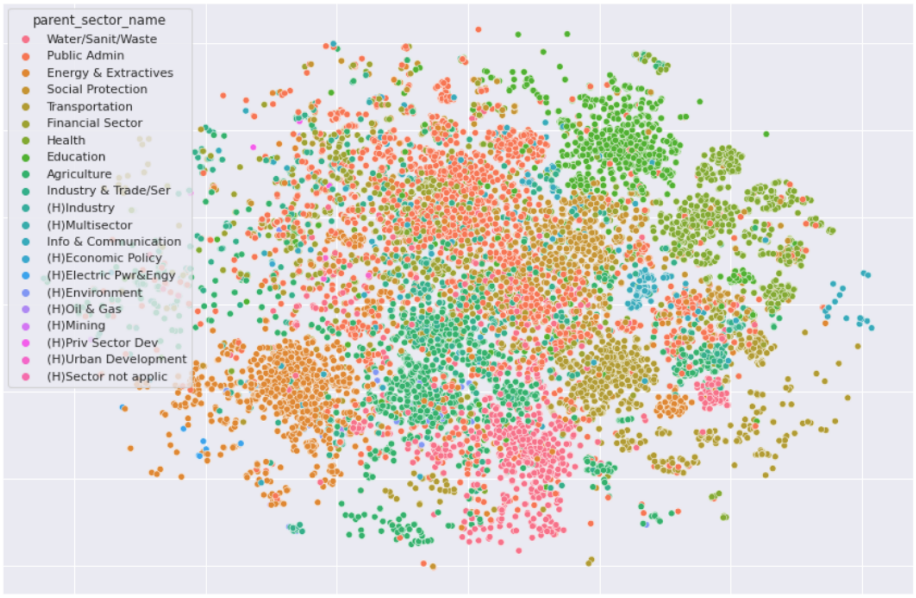

Figure 1: Dimensionality-reduced contextual embeddings of Project Development Objectives

We find strong indication that these embeddings are capturing meaningful interrelationships among projects. To visualize this, we reduce these high-dimensional embeddings to two dimensions. Though the axis have no direct interpretability, the plot of embeddings for all World Bank projects shows clear separation by sector, even though their area of focus was not included in any of the information provided to the embedding pipeline (Figure 1). Moreover, when considering health projects, the mean embedding is stable throughout the period 1990−2019, with slight fluctuations in standard deviation per decade (see Table 1). In 2020−21, however, the mean shifts dramatically, and variance collapses - as we would expect, given the COVID-19 pandemic and consequent focus on emergency response.

We then use these embeddings to construct a novel measure for the degree of tailoring of a project to its context, which we will call *project contextualization*. To do so, we calculate the mean embeddings for each decade in each sector and in each country, and then compute the Euclidean distance between each project's embedding and the mean for its sector and country in the decade it was approved. Formally, for each sector s and decade d, with projects in the

Table 1: Health projects embeddings evolution over time

Period	N	Mean X	Mean Y	Mean Distance
1990−99	195	7.84	9.64	5.01
2000−09	275	7.75	7.65	5.55
2010−19	308	7.65	9.89	2.79
2020−21	203	9.44	11.76	1.19

Note: Embeddings are projections into abstract high-dimensional vector space expressing inter-relationships, so lack physical units. This table reports the reduction of health projects embedding in two dimensions (X and Y), centered on the 1990−1999 mean.

sector and decade P_{sd} and $|P_{sd}| = N_{sd}$, we calculate:

$$E^*_{sd} = \frac{1}{N_{sd}} \sum_{n=1}^{N_{sd}} E_p, p \in P_{sd} \tag{3}$$

and similarly construct E^*_{cd} for each country c. We then define the *sector distance*, denoted T^s_p for a given project p, as $\|E_p - E^*_{sd}\|$. Similarly, the *country distance*, denoted T^c_p, is defined as $\|E_p - E^*_{cd}\|$. The sector distance can be interpreted as the degree to which the project document is, in the deep texture of its language, adjusting sectoral knowledge to local country realities, and the country distance a similar measure of how sectoral peculiarities are being brought to bear on local problems.

Further work is needed to determine what design aspects, as reflected in the text of the project documents, are most clearly associated with development outcomes. In this paper, we restrict ourselves to embeddings generated from the project development objectives (PDO), results framework indicators and implementation completion report (ICR). But even these appear to be capturing the degree of specificity to country circumstances. To illustrate, here is the PDO of a health project with very low (sector distance more than 1 standard deviation smaller) contextualization:

> The revised PDOs are to: (i) improve coverage, utilization and quality of health care services in the territory of the Recipient, and (ii) strengthen the Government's stewardship functions in the health sector.

By contrast, more contextualized PDOs are more specific with respect to population groups, periods and outcomes. Here are another two examples with high (more than one standard deviation above average distance to sector mean) contextualization:

> The Project's development objective is to ensure access to improved and sustained water and sanitation services in rural communities in [redacted country name]. This would be accomplished through the implementation of the new Rural Water Supply and Sanitation

> (RWSS) sector policy and the preparation of a National RWSS program. To this end, the Project would support a decentralized and demand responsive delivery mechanism and help build the institutional foundation for implementing the National RWSS Program both at the central and local governments levels.

And:

> The specific objective of this project is to support programs designed to halt transmission of HIV/AIDS among vulnerable populations (PLWHA, IDUs, CSWs, and their clients and sexual partners) and between these vulnerable populations and the general population. Key outcome indicators include: Percent of vulnerable groups in participating provinces reporting safer injection practices (from an estimated 20% at baseline to 70% at the project end); Percent of vulnerable groups in participating provinces reporting condom use in sexual intercourse (from an estimated 40% at baseline to 80% at project end).

Finally, we seek to identify relationships between the contextualization variable and project preparation and supervision characteristics. These characteristics comprise project region, recipient income levels, fragility, and institutional strength; the time and cost of project preparation; the location (headquarters or country office) and experience of management; whether or not analytical work in the sector and country was conducted in the years prior to the project; and various characteristics of the project manager, including education level (PhD or not), age, experience, and prior work in the sector, analytical or lending. Since we have embeddings for all historical World Bank projects, we are able to probe for these relationships across a somewhat larger dataset ($N = 4,260$).

3.5. Non-Linear Models

Having constructed the residual outcomes and text embeddings, we set two prediction tasks:

1. Binary classification. For each project, the target prediction is whether the residual outcome in the project sector is positive five years after project completion.
2. Regression. For each project, the target prediction is the precise residual outcome in the project sector five years after project completion.

The lag-time for prediction follows the replicated studies and is used because projects may not target the specific sector outcome during the project period itself, or may take place in only a few districts at a time. However, as discussed in section 2.2, this is justified in that almost all development projects aim for systemic effects, whether via demonstration, capacity building or other channels.

As inputs to the non-linear models used for each prediction we construct vectors for each project p consisting of:

1. Basic quantitative data on the project, such as the size of the loan (in 2010-adjusted US dollars), its duration (in months), and the percentage of its budget allocated to its primary sector. We extend several of these features, for example calculating the Herfindahl-Hirschman Index (HHI) for budget allocation across sectors (see Table D.11).
2. Categorical features, such as the funding source (IBRD or IDA[10]) and the financing instrument (the loan or grant type). The features are one-hot encoded and described in Table D.12.
3. Text embeddings for the project title and project development objective (PDO), as well as for the implementation completion report (ICR) and the results framework indicators, where they exist.
4. Project contextualization features, i.e., sector-distance and country distance for each of the embeddings.[11]

We concatenate the numeric and categorical features, the text embeddings, and the distance measures, to generate the combined project feature vector X_p. Combining across health, education, WASH and energy, this results in $n = 1,457$ projects as inputs, with the lagged residual outcomes (as described above) as targets for predictions. Following standard practice we construct a test set of $N_{test} = 146$ projects and train on $N_{train} = 1,311$.

We then utilize standard techniques to search among model architectures and among hyper-parameters for the models. In each case we trivially use classification and regression variants of the model architectures. We use a linear model as a baseline, in both modern variants (Lasso and Ridge). We also consider support vector machines, decision tree ensembles (random forest), gradient boosted trees (XGBoost), and fully-connected neural networks (small in size, given the limited data). The full list of architectures and hyper-parameters is provided in Table D.13.

To measure predictive performance, we use the receiver-operator area under curve (ROC AUC) metric for the binary classification, measured on the test set. ROC AUC can be interpreted as the probability that the model will rank more highly a random project associated with a positive residual than one associated with a negative lagged residual of being positive. A ROC AUC of 0.5 means the model is only as good as random choice in distinguishing between positive and negative projects, and a ROC AUC of 1 means it distinguishes such projects perfectly. We also report the r^2 of the corresponding regression models on the training set, in order to compare results to more traditional regression techniques in the development literature, which do not use train-test splits.[12]

[10] International Bank for Reconstruction and Development or International Development Association.

[11] We include both these distance features and each project's raw embedding, since the embedding on its own can (and by our empirical results does) contain information about a given project's relationship to others not captured in the sector- and country-distances alone.

[12] We do not use results on the training set to select models or make claims for them, following standard practice.

3.6. Robustness and Interpretation

We perform multiple checks for robustness of both the linear and non-linear models. For the linear models, we search over multiple possible linear specifications by adding and removing controls and adjusting lags and observe the effects on significance measures and coefficients for exogenous variables.[13]. We also test for orthogonality between X and W in equation 1, to test for the possibility that W is a function in part of past X and hence that the residual in equation 2 is prematurely purged of the influences of X, weakening the association unintentionally. In other words, we test that associations between past aid and present controls are not muddying the results. The non-linear models are all tested using the standard practice of K-fold cross-validation.[14]

We interpret the models in part using standard methods specific to model types, such as coefficient magnitudes and significance for linear models and impurity-based feature importance in decision-tree ensembles, supplemented by Shapley Additive Explanations (SHAP values, see Lundberg and Lee 2017). SHAP values use a game theoretic approach to explain the output of a model by attributing contributions to the final prediction to model features, analogously to attributing the contribution of individual players within a team to the final result of a game.

4. Results

4.1. Sectoral Aid Effects: Can any general statements be made about the impact of development aid on sector outcomes?

The linear model regression results are reported in Appendix A. In each case, the volume of aid per capita had a statistically significant effect on sector outcomes, appropriately lagged and smoothed. The coefficients were, though, modest in each sector:

- Doubling per capita education aid is associated with an 8 percentage point increase in net primary school enrolment.
- Doubling per capita health aid is associated with a 2 percentage point reduction in the infant mortality rate.
- A 1 percentage point increase in WASH aid as percentage of GDP is associated with between 1-5 percentage point increase in rural access to water and sanitation.

[13] In other words, we introduce causal perturbation, following the practice in the DoWhy library (Sharma and Kiciman 2020).

[14] The training set is itself divided five times into a "hold-out" (or validation) set and a training set proper, with a candidate model trained on the training set proper and scored on the validation set. After the five runs both the scores and the models themselves are averaged, and excessive variance between each "fold" is examined.

- Doubling per capita energy aid is associated with a 2 percentage points increase in access to energy.
- Doubling per capita aid to fiscal management is associated with a 4 percentage points increase in the tax (net of social contributions) to GDP ratio.

These results were largely robust to causal perturbation. Coefficients remained significant with only minor changes in magnitude when controls were added or removed, with the partial exception of education, where the addition of a lagged prior enrollment figure resulted in the aid coefficient becoming insignificant. The consistency of the results gives us confidence in saying that aid is associated with improved sector outcomes, but the effect is generally modest, and dwarfed by other variables.

As a robustness check, we also find that the aid variable has low-to-trivial correlation with the controls in almost all cases, and the cases of moderate correlation argue more for W causing X in equation 1 than vice-versa. For example, HIV prevalence (Pearson coefficient of 0.39 with period-average mean per-capita health commitments) and fertility (0.26 on the same measure) are the only health controls with more than a 0.1 correlation with per capita aid, and the coefficient is positive, i.e, greater levels of HIV and fertility result in more aid.

These results do not change when longer lags are introduced to X. For example, the correlation of last-five-years' aid and pupil-teacher ratios is -0.16 and that between 5-to-10-years' aid and the same ratio is -0.17. More aid is then extremely weakly correlated to lower pupil-teacher ratios, but that, and Freedom House ratings (-0.3), are the strongest correlation between the controls and treatments and those correlations do not strengthen (even trivially) when lagging X. This strongly suggests that X and W are largely orthogonal, and that aid's effects are not stronger through some lagged or cumulative effect on the controls.[15]

On the other hand, the coefficients on aid are probably an underestimate of its true effect. First, we are comparing "flows" of aid to an effective "stock" of sector outcome performance. Second, some proportion of the aid flows will be unrelated to the specific outcome used as the dependent variable (e.g. to higher rather than primary education, or to learning outcomes as opposed to enrollment). But the degree of underestimation is likely limited by the intentional alignment between official aid flows and the outcome-level indicators to

[15] Correlations remain small when lagging total aid as far back as a decade, and, while aid within the sectors summed country-wise over the period are moderately to strongly correlated, such aid is not so correlated when disaggregated over time. In other words, there is little reason to believe that the effect of sectoral aid is being weakened by a relationship between overall aid and growth followed by growth and sectoral outcomes. Further avenues to try to increase the size of the effect of aid's effectiveness are beyond this paper, whose principal purpose is not to investigate this relationship in itself.

measure progress towards the Millennium Development Goals (MDGs) that we use as our dependent variables.

4.2. Project Rating Significance: Do project outcome ratings provide information on the relationship between aid flows and sector outcomes?

The results for project ratings, reported in detail in Appendix A and summarized in Table 2, are also clear. Only for fiscal management outcomes do the weighted average ratings convey information about outcomes. In the other sectors, the ratings are not significant: the coefficients are near zero and their inclusion makes no difference to the coefficients on aid volume. These results are robust to using the alternate measures of ratings and to the restriction of aid flows to particular donors or the World Bank Group alone. The one exception is in a specification for sanitation, but with a small sample size, a small coefficient and a negative sign.

The more sensitive "debiased/double machine learning" techniques confirm the absence of effects seen in the traditional regressions. Table 2 shows, for each sector, the r^2 of the partialing out step, the r^2 of the treatment test, the coefficient on the treatment (weighted average rating) in the treatment test and the p-value of the treatment test.

It might be argued that ratings' absence of information in four out of five sectors is a result of projects focusing on other outcomes than those we are testing against. But this cannot explain why aid volume enters significantly against the sector outcomes, and, when examined in detail, requires implausible assumptions to account for the results. Assume that some proportion X of the aid in a sector targeted the MDG sector outcome, and the rest targeted entirely unrelated outcomes. Then the "true" coefficient on aid volume would be $1/X$ times the coefficient detected in our regression. If overall project ratings did convey information, then they should convey information on the X proportion targeting the outcome, and hence should modify the coefficient on volume, even if attenuated. The ratings on the $1 - X$ share of aid explicitly targeted to non-MDG outcomes would diminish the rating effect. But the coefficient could only be reduced to insignificance if the ratings on the $1 - X$ proportion were negatively correlated to the X share and neutralized them precisely.

We also note that, during this period, projects were predominantly MDG related and that dividing the period into two, one peak MDG period and one after, does not alter the estimate. Further, we do detect a relationship for fiscal management, even though not all fiscal projects explicitly aim to increase the tax share of GDP. Finally, as noted in section 3.5, the use of lags makes it even more reasonable to expect effects on the dominant MDG outcome from projects in the sector. In all, it seems far more plausible that ratings are not providing information than that ratings on a small share of non-primary-outcome projects precisely cancel out information in the primary-outcome projects, even though the results are unchanged in periods where that share was trivial and even accounting for lags in within-sector spillovers, and not least because this neutering effect would have to mysteriously vanish in one out of five sectors.

Table 2: Significance and Magnitude of Coefficient on Ratings

Sector	N	rR_c^2	r_t^2	Rating coefficient
Education	731	0.48	0.00	-0.01
Health	250	0.80	0.00	-0.01
WASH	406	0.84	0.01	-0.04**
Energy	317	0.87	0.01	-0.02
Fiscal	539	0.62	0.80	0.07***

Notes: Sector outcome ratings were incorporated as the dollar-weighted average rating in the prior period. r_c^2 denotes the adjusted r^2 on the initial regression of the sector outcomes against the controls, and r_t^2 denotes the adjusted r^2 for the regression of the residual outcomes. Significance of rating coefficients indicated as ***p<0.01, **p<0.05, *p<0.10.

4.3. Contextualization and Sector Outcomes: Can we detect associations between project characteristics and development outcomes?

We used the strongest performing non-linear model to test for feature importance. The results are shown in Figure 2. The embedding features are dominant, followed by the measures of sector concentration, and only afterwards country institutional quality and other project characteristics more commonly used in the literature.

The most important embeddings are those associated with the PDO, especially the distances of that embedding to sector and country mean, followed by results framework indicators and the ICR. The importance of the contextualization features indicate that the model is learning to detect from the embeddings the degree to which a project has been contextualized to its country, via its distance to its sector mean, and to which country considerations have been adjusted in light of the sector's characteristics, via the distance to the country mean.[16]

We consider feature importance when adding in during-project and at-review features. We find that the embeddings of the ICR report itself then join the PDO embedding and contextualization measures as one of the most important features, and on some specifications becomes **the** most important feature. The actual project length (as opposed to its proposed length) is similarly important. However, neither fully displaces those from approval, and the PDO embeddings and contextualization and concentration measures retain high importance. In keeping with our other results, project ratings are unimportant.

Roughly half of the projects in the dataset had a positive residual outcome, which we would generally expect given their construction. Ensemble based

[16]It is important to remember here that these are measures of relative contribution to the performance of a non-linear model on the entire dataset and cannot be used in the trivial manner of a coefficient in a simple linear regression, to read off that, for example, a proportional increase in contextualization leads immediately, *ceteris paribus*, to a certain increase in predicted performance. More simply, nothing in these results should be taken to imply that writing a longer PDO with the names of some local programs will lead to improved sector outcomes (or even that simply an increase in intellectual effort across the PAD will do so).

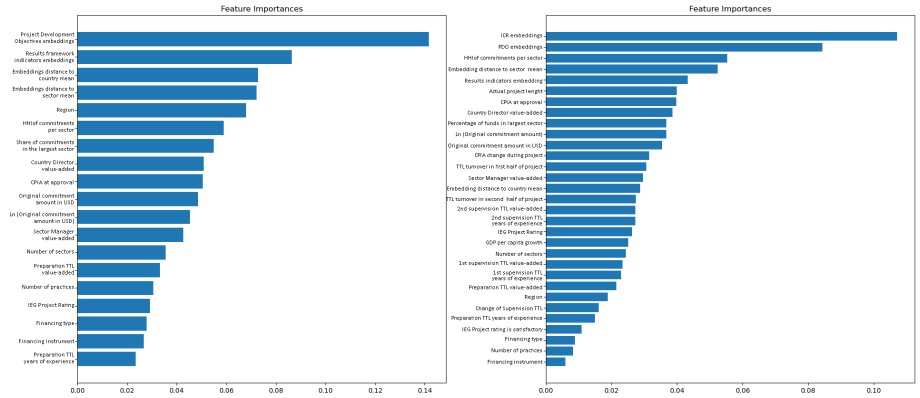

Figure 2: Feature importance with only at approval (left) and at review (right) features, measured in relative weight (i.e., with all features' importance summing to 1)

methods achieved ROC AUCs approaching 0.7, indicating that the model correctly predicted a positive or lagged residual outcome 7 out of 10 times. Similarly, ensemble models' regression performance on the training set was high, with an adjusted r^2 of 0.76. Further details and full results are provided in Appendix B.2.

We also attempt to identify the correlates of the gap between project ratings and outcomes by training a non-linear model to predict this distance, which might be characterized as the degree to which ratings have been gamed. We find that loan size and region are important in explaining the distance between the rating and the residual outcome (see Table C.10). Specifically, gaming appears more likely for large loans, particularly those in countries with weak institutional assessment scores at project approval.

4.4. Determinants of Contextualization: What do we know about the characteristics of development projects associated with positive sector outcomes?

We cannot rule out that contextualization and sector outcomes are both driven by other unobservable factors. It is plausible, for example, that very effective actors from donor agencies and government counterparts design better and more contextualized projects, and that these actors also drive better results. Nevertheless, given the importance of contextualization in explaining sector outcomes, we investigate what aspects of the project preparation process might be conducive to it. Here our findings are only tentative. We find some evidence that it is associated with longer preparation times and existence of prior analytical work, but our results do not lend themselves to robust interpretation. A random forest model is able to explain 30% of the variance in embedding sector distance (as described in section 3.6), while linear models explain less than 10% of the variance.

5. Conclusion

We present two main findings. The first is that in four of the five sectors for which we have data, donor agencies' project outcome ratings provide no information as to the long-run impact of their projects. It follows that they are a poor measure of aid effectiveness, though they may still be useful for monitoring other aspects of project performance. Our second finding is that the single most important correlate of impact is what appears to be a proxy for the degree of contextualization of project design to country circumstances, far ahead of country institutional quality, project size or other commonly identified factors. Our methods enable us to explain around 70% of the residual variance in development outcomes, a much higher proportion than previous analyses of the determinants of project outcome ratings.

These findings have significant implications for how we think about project preparation. Further work is needed to establish what the embeddings mean. But our results at least suggest that greater attention be paid to country contextualization. This does not appear to correlate with the standard determinants of project quality identified in the literature, such as project manager's age, education or prior experience. Nor does it correlate with whether the staff is based in headquarters or in the field - consistent with IEG's own assessment that corporate field staffing targets have failed to ensure that decentralization is tailored to country and program needs or applied to areas where it can bring the most benefits (Independent Evaluation Group 2016). While we find some evidence that the length of project preparation and existence of prior analytical work are positively associated with impact, we are unable to say to what extent they matter.

Our analysis also has implications for how we think about project evaluation. On the one hand, it is encouraging that ICRs provide sufficient information to accurately predict the likely contribution of projects to long-run outcomes. On the other hand, the incorporation of this information into summary ratings appears to have resulted in them becoming disassociated from development impact. This may warrant a more careful reading of the qualitative evidence in ICRs, as well as more attention to how project outcome targets are calibrated, perhaps by requiring teams to specify an ex ante counterfactual absent the World Bank's involvement.

References

Andrews, Matthew (2021). "Successful Failure in Public Policy Work". In: *CID Faculty Working Paper Series No. 402*. URL: https://www.hks.harvard.edu/centers/cid/publications/faculty-working-papers/successful-failure-public-policy.

Ashton, Louise et al. (2023). "A Puzzle with Missing Pieces : Explaining the Effectiveness of World Bank Development Projects". In: *The World Bank Research Observer* 38 (1), pp. 115–146. URL: https://doi.org/10.1093/wbro/lkac005.

Birchler, Kassandra and Katharina Michaelowa (2016). "Making aid work for education in developing countries: An analysis of aid effectiveness for primary education coverage and quality". In: *International Journal of Educational Development* 48, pp. 37–52. URL: https://doi.org/10.1016/j.ijedudev.2015.11.008.

Blum, Jurgen Rene (2014). "What factors predict how public sector projects perform? A review of the World Bank's public sector management portfolio". In: *World Bank Policy Research Working Paper* 6798. URL: http://hdl.handle.net/10986/17299.

Chernozhukov, Victor et al. (Jan. 2018). "Double/debiased machine learning for treatment and structural parameters". In: *The Econometrics Journal* 21.1, pp. C1–C68. ISSN: 1368-4221. URL: https://doi.org/10.1111/ectj.12097.

Denizer, Cevdet, Daniel Kaufmann, and Aart Kraay (Nov. 1, 2013). "Good countries or good projects? Macro and micro correlates of World Bank project performance". In: *Journal of Development Economics* 105, pp. 288–302. ISSN: 0304-3878. DOI: 10.1016/j.jdeveco.2013.06.003. URL: https://www.sciencedirect.com/science/article/pii/S0304387813000874.

Ford, Ian and John Norrie (2016). "Pragmatic trials". In: *New England journal of medicine* 375.5, pp. 454–463. URL: https://www.nejm.org/doi/full/10.1056/NEJMra1510059.

Honig, Daniel (2018). "Navigation by Judgment: Why and When Top Down Management of Foreign Aid Doesn't Work". In: DOI: 10.1093/oso/9780190672454.001.0001. URL: https://oxford.universitypressscholarship.com/10.1093/oso/9780190672454.001.0001/oso-9780190672454.

Honig, Daniel, Ranjit Lall, and Bradley C Parks (2022). "When does transparency improve institutional performance? Evidence from 20,000 projects in 183 countries". In: *American Journal of Political Science*. DOI: https://doi.org/10.1111/ajps.12698.

Independent Evaluation Group (2016). *Behind the Mirror: A Report on the Self-Evaluation Systems of the World Bank Group*. World Bank. URL: http://hdl.handle.net/10986/24956.

Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 4765–4774.

URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

McInnes, Leland, John Healy, and James Melville (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. DOI: 10.48550/ARXIV.1802.03426. arXiv: 1802.03426 [stat.ML]. URL: https://arxiv.org/abs/1802.03426.

Mishra, Prachi and David Newhouse (2009). "Does health aid matter?" In: *Journal of health economics* 28.4, pp. 855–872. DOI: 10.1016/j.jhealeco.2009.05.004. URL: https://www.sciencedirect.com/science/article/pii/S0167629609000563.

Moore, Mark H (1995). *Creating public value: Strategic management in government*. Harvard university press. ISBN: 9780735100046.

Ndikumana, Léonce and Lynda Pickbourn (2017). "The impact of foreign aid allocation on access to social services in sub-Saharan Africa: The case of water and sanitation". In: *World Development* 90, pp. 104–114. DOI: 10.1016/j.worlddev.2016.09.001. URL: https://www.sciencedirect.com/science/article/pii/S0305750X1530543X.

Reimers, Nils and Iryna Gurevych (2019). "Sentence-bert: Sentence embeddings using siamese bert-networks". In: *arXiv preprint arXiv:1908.10084*.

Sharma, Amit and Emre Kiciman (2020). *DoWhy: An End-to-End Library for Causal Inference*. arXiv: 2011.04216 [stat.ME].

Tierney, Michael J et al. (2011). "More dollars than sense: Refining our knowledge of development finance using AidData". In: *World Development* 39.11, pp. 1891–1906. DOI: https://doi.org/10.1016/j.worlddev.2011.07.029. URL: https://www.sciencedirect.com/science/article/pii/S0305750X1100204X.

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems*, pp. 5998–6008.

Williams, Martin J (2019). "External Validity and Policy Adaptation: From Impact Evaluation to Policy Design". In: *The World Bank Research Observer* 35.2, pp. 158–191. ISSN: 0257-3032. DOI: 10.1093/wbro/lky010. URL: https://doi.org/10.1093/wbro/lky010.

World Bank (2021). *World Development Indicators*. URL: https://databank.worldbank.org/source/world-development-indicators.

Appendix A. Linear Regression Results

Table A.3: Education Aid Volumes, Ratings and Sector Outcomes

	I	II	III	IV	V
Education Aid[1]	0.08**	0.08**			
	(0.03)	(0.03)			
PPD-only Education Aid[1]			-0.01	-0.01	-0.01
			(0.02)	(0.02)	(0.02)
Avg Education Rating[2]		0.02	0.01		
		(0.03)	(0.03)		
Max Education Rating[2]				-0.04	
				(0.03)	
Binary Education Rating[2]					0.01
					(0.12)
Young Population[3]	-0.08***	-0.08***	-0.03	-0.03	-0.03
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Pupil-teacher ratio[4]	-0.02	-0.02*	-0.02*	-0.02*	-0.02*
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
GDP per capita PPP[5]	1.15***	1.15***	1.64***	1.69***	1.64***
	(0.28)	(0.28)	(0.31)	(0.31)	(0.31)
Cash surplus/deficit[6]	0.00	0.00	0.01*	0.02*	0.01*
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Inflation (%)	0.00	0.00	0.00***	0.00***	0.00***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Trade (% of GDP)	-0.01**	-0.01**	-0.00*	-0.01*	-0.00*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Freedom House[7]	0.07	0.07	0.09	0.08	0.09
	(0.06)	(0.06)	(0.07)	(0.07)	(0.07)
R-squared	0.53	0.53	0.56	0.56	0.56
R-squared Adj.	0.49	0.49	0.51	0.51	0.51
No. observations	1243	1243	999	999	999

Notes: The dependent variable is primary school enrolment (% net) in all specifications. Specification (I) closely matches Birchler and Michaelowa (2016). Independent variables: (1) Aid = log, average of $ commitments per capita in the prior 5y period; (2) Avg = $-weighted average project outcome rating in the prior 5y period, Max = maximum project outcome rating achieved in the prior 5y period, Binary = dummy for any satisfactory project outcome ratings in the prior 5y period; (3) Share of population ages 0-14; (4) in primary; (5) constant 2017 international $; (6) Government cash surplus/deficit as % of GDP; (7) Average of Freedom House Political Rights and Civil Liberties scores. Constant, fixed effects, and missing value indicators for imputed variables are included but not shown. Robust standard errors in parentheses. Significance of coefficients indicated as ***p<0.01, **p<0.05, *p<0.10.

Table A.4: Health Aid Volumes, Ratings, and Sector Outcomes

	I	II	III	IV	V	VI
Health Aid[1]	0.00*	0.10***	0.01**	0.00	0.00	0.01
	(0.00)	(0.02)	(0.00)	(0.00)	(0.00)	(0.01)
Avg Health Rating[2]			0.08	0.05		
			(0.07)	(0.07)		
Max Health Rating[2]						-0.11
						(0.14)
Binary Health Rating[2]					0.00***	
					(0.00)	
HIV prevalence[3]	0.03***	0.02***	0.02***	0.02***	0.02***	-0.01
	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.02)
Fertility rate[4]	0.38***	0.34***	0.38***	0.37***	0.37***	0.30***
	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.03)
GDP per capita PPP[5]	-0.05***	-0.07***	-0.05***	-0.07***	-0.07***	-0.04*
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Population total	0.11***	0.11***	0.11***	0.10***	0.10***	0.05***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Conflict (UCDP/PRIO)	0.02	0.00	0.03	-0.02	-0.02	0.14
	(0.09)	(0.08)	(0.09)	(0.08)	(0.08)	(0.11)
Access to water		0.02***		0.02***	0.02***	0.01***
		(0.00)		(0.00)	(0.00)	(0.00)
Access to sanitation		-0.01***		-0.01***	-0.01***	-0.01***
		(0.00)		(0.00)	(0.00)	(0.00)
Physicians rate[6]		-0.03		-0.02	-0.02	-0.05
		(0.03)		(0.03)	(0.03)	(0.06)
R-squared	0.83	0.86	0.83	0.86	0.86	0.81
R-squared Adj.	0.83	0.86	0.83	0.85	0.85	0.79
No. observations	609	609	585	585	585	213

Notes: The dependent variable is under-5 mortality rate (per 1,000 live births) in all specifications. Specifications (I) and (II) closely match Mishra and Newhouse (2009). Independent variables: (1) Aid = log, average of $ commitments per capita in the prior 5y period; (2) Avg = $-weighted average project outcome rating in the prior 5y period, Max = maximum project outcome rating achieved in the prior 5y period, Binary = dummy for any satisfactory project outcome ratings in the prior 5y period; (3) Share of population ages 15-49; (4) births per woman; (5) constant 2017 international $; (6) Number of physicians per 1,000 people. Constant, fixed effects, and missing value indicators for imputed variables are included but not shown. Robust standard errors in parentheses. Significance of coefficients indicated as ***p<0.01, **p<0.05, *p<0.10.

Table A.5: WASH Aid Volumes, Ratings and Sector Outcomes

	Water I	II	III	Sanitation I	II	III
WASH Aid[1]	0.01**	0.01**	0.02**	0.04***	0.04***	0.05*
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.03)
Avg WASH Rating[2]		0.00			-0.00	
		(0.01)			(0.01)	
Binary WASH Rating[2]			-0.02*			-0.06
			(0.01)			(0.07)
Adult literacy[3]	0.00	0.00	0.01**	0.01***	0.01***	0.00
	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Young Population[4]	0.02*	0.02*	0.02***	-0.02***	-0.02***	-0.01
	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)
GDP per capita PPP[5]	-0.72***	-0.72***	0.16***	0.12***	0.12***	0.46***
	(0.10)	(0.10)	(0.05)	(0.04)	(0.04)	(0.08)
Conflict (UCDP/PRIO)	-0.11*	-0.11*	-0.02	0.22***	0.22***	0.09
	(0.07)	(0.07)	(0.02)	(0.07)	(0.07)	(0.11)
Lag access to sanitation	0.04***	0.04***	0.00			
	(0.01)	(0.01)	(0.00)			
Lag access to water				0.02***	0.02***	0.01***
				(0.00)	(0.00)	(0.00)
R-squared	0.63	0.63	0.99	0.62	0.62	0.84
R-squared Adj.	0.60	0.60	0.98	0.61	0.61	0.82
No. observations	755	755	121	755	755	121

Notes: The dependent variable is access to improved water source (% of population) in Water I, II and III specifications, and access to improved sanitation facilities (% of population) in Sanitation I, II and III. Specifications Water I and Sanitation I closely match Ndikumana and Pickbourn (2017). Independent variables: (1) Aid = average of WASH commitments as % of GDP in the prior 5y period; (2) Avg = $-weighted average project outcome rating in the prior 5y period, Binary = dummy for any satisfactory project outcome ratings in the prior 5y period; (3) Literacy rate (% of people ages 15 and above); (4) Share of population ages 0-14; (5) log, constant 2017 international $. Constant, fixed effects, and missing value indicators for imputed variables are included but not shown. Robust standard errors in parentheses. Significance of coefficients indicated as ***p<0.01, **p<0.05, *p<0.10.

Table A.6: Energy Aid Volumes, Ratings and Sector Outcomes

	I	II	III
Energy Aid[1]	0.02***	0.02***	0.03
	(0.01)	(0.01)	(0.03)
Avg Energy Rating[2]		0.00	
		(0.00)	
Binary Energy Rating[2]			0.02
			(0.05)
Adult literacy[3]	0.01**	0.01**	0.00
	(0.00)	(0.00)	(0.01)
Young Population[4]	0.06***	0.06***	0.07**
	(0.01)	(0.01)	(0.03)
GDP per capita PPP[5]	0.34***	0.34***	0.55***
	(0.06)	(0.06)	(0.20)
Conflict (UCDP/PRIO)	-0.05*	-0.05*	-0.01
	(0.03)	(0.03)	(0.07)
R-squared	0.96	0.96	0.98
R-squared Adj.	0.96	0.96	0.97
No. observations	753	753	104

Notes: The dependent variable is access to electricity (% of population) in all specifications. Independent variables: (1) Aid = log, average of $ commitments per capita in the prior 5y period; (2) Avg = $-weighted average project outcome rating in the prior 5y period, Binary = dummy for any satisfactory project outcome ratings in the prior 5y period; (3) Literacy rate (% of people ages 15 and above); (4) Share of population ages 0-14; (5) log, constant 2017 international $. Constant, fixed effects, and missing value indicators for imputed variables are included but not shown. Robust standard errors in parentheses. Significance of coefficients indicated as ***p<0.01, **p<0.05, *p<0.10.

Table A.7: Fiscal Policy Support Volumes, Ratings and Sector Outcomes

	I	II
Fiscal Aid[1]	0.04***	0.00
	(0.01)	(0.02)
Avg Fiscal Rating[2]		0.07***
		(0.02)
Conflict (UCDP/PRIO)	0.38***	0.18*
	(0.03)	(0.09)
GDP per capita PPP[3]	0.01	0.07
	(0.01)	(0.07)
ODA (% of GNI)	0.01	-0.12***
	(0.01)	(0.02)
Freedom House[4]	-0.01	0.02
	(0.01)	(0.02)
R-squared	0.64	0.82
R-squared Adj.	0.62	0.80
No. observations	2893	539

Notes: The dependent variable is tax (net of social contributions) to GDP ratio in all specifications. Independent variables: (1) Aid = log, average of $ commitments per capita in the prior 5y period; (2) Avg = $-weighted average project outcome rating in the prior 5y period; (3) constant 2017 international $; (4) Average of Freedom House Political Rights and Civil Liberties scores. Constant, fixed effects, and missing value indicators for imputed variables are included but not shown. Robust standard errors in parentheses. Significance of coefficients indicated as ***p<0.01, **p<0.05, *p<0.10.

Appendix B. Machine Learning Methods

Appendix B.1. Text Embeddings

We extend our analysis to encompass numerical representations of text related to the project, known as "text embeddings". These embeddings are produced by complex functional forms ("transformer models") that rely on a mechanism called self-attention (Vaswani et al. 2017). The complexity of these functional forms and the use of machine learning to set their parameters by stochastic gradient descent means they are more difficult to interpret than simpler statistical measures, as will be discussed further below, but compensate with substantial gains in empirical results.[17] We use a two-step process to generate such embeddings for development projects. First, we use pretrained transformer models to generate embeddings of each word in the text. Specifically, we use an extension of these models to *sentence embeddings*, in which whole sentences are encoded using a transformer architecture trained to embed "close" sentences (measured by cosine-similarity of their word-level embeddings) close to each other (Reimers and Gurevych 2019). That first-stage model produces a very high ($n = 768$) dimensional vector, too high to be used downstream given the number of projects available. In our second step, therefore, we reduce the embeddings' dimension using a combination of principal-component analysis (PCA) and uniform manifold approximation and projection (UMAP).[18] UMAP is a state-of-the-art technique for dimensionality reduction that combines machine learning and algebraic topology to learn a low dimensional manifold projection of a high dimensional set of data (McInnes, Healy, and Melville 2020). We utilize PCA to reduce dimensionality to $n = 76$, then use UMAP to further reduce to 2-dimensions. The resulting 2-dimensional embedding vector we label E_p, for a given project p.

As a caveat, although the embeddings capture interrelationships among texts their absolute position is not in itself meaningful. That is even more the case when the embeddings are passed through UMAP, which is a stochastic process and therefore will result in random variation in the absolute position of any particular embedding in its dimensionality-reduced form. The reduced-form embeddings are meaningful only when used to construct intermediate relationships, such as distances to means, and when conjoined with other features of projects and fed through a training process as part of an entire dataset. As

[17] These models now power all state-of-the-art translation, search engines, AI text generators, as well as most plagiarism detectors.

[18] Dimensionality reduction is known to degrade the performance of downstream tasks utilizing sentence embeddings, and so is avoided in ML research where possible. However, given the limited size of our dataset, utilizing the full-width embeddings would have created its own difficulties with over-fitting. On balance, we decided to reduce the embedding width, but note that if a larger project-level dataset were constructed, more limited reduction might lead to significant gains in the downstream model performance reported in Section Appendix B.2. We explored alternate combinations of PCA, UMAP, as well as t-SNE for robustness, but found that as well as having the most appealing theoretical justification the pipeline used provided the most stable and accurate downstream results.

an obvious robustness test, we rerun our non-linear model pipelines end-to-end with different random instances of UMAP, and find that the results are stable.

Appendix B.2. Residual Outcome Predictions

In each sector roughly half of the projects in the dataset had a positive residual outcome (see Table B.8). Ensemble based methods achieved receiver-operator area under curve (ROC AUC) approaching 0.7, indicating that the model correctly predicted a positive or lagged residual outcome 7 out of 10 times. Similarly, ensemble models' regression performance on the training set was high, with an adjusted r^2 of 0.76. Prior techniques had been able to explain at most 30% of the variance in project ratings (which ratings are themselves, as above, of doubtful importance). Non-linear models are able to explain a substantially higher percentage of variation of a more meaningful target variable, with the positive test set performance and similar results across the model types and across folds giving confidence that this result is not simply the product of over-fitting or label leaking. Full results are reported in Table B.9.

One note is that performance collapses for linear models. All linear models had ROC AUCs below a coin toss, and explained little to no variance in the target. Tree-based models outperformed Support Vector Machines (SVMs) learning methods, although with minimal differences between ensemble methods and gradient boosting. This may lead to concerns that the tree-based ensemble methods are over-fitting. Such concerns should be alleviated by the relatively large hold-out set and the use within the training set of K-fold cross validation.

However, we conducted additional tests for robustness in several ways. First, we examined the scores for hyperparameter combinations with heavy regularization, that is, which significantly penalized over-fitting. When we did so, we found some decline in performance, but only moderately. For example, reducing the maximum tree depth from 100 to 3 reduced the testing set AUC score from 0.67 to 0.63, a modest reduction (and no reduction was observed at max depth 10). Second, we dropped all but the top 20% of features (by feature importance) and similarly saw declines of only 4 percentage points in the test set ROC AUC and B in the training set adjusted r^2. When we add the features found at review time to those at approval time, we find a slight performance increase, with an ROC AUC score of 0.7 and an explained variance of 0.86. One further concern might be that the models were simply detecting the presence or absence of sectoral outcome keywords. To check for that possibility, we tested for correlations between the presence of sectoral key words and loan size and the residual outcomes, and found none (see B.3).

Table B.8: Summary statistics for residual outcomes

Sector	N	Positive	Mean	StdDev
Education	352	218	0.17	0.89
Energy	225	99	-0.12	0.94
Health	580	242	-0.07	1.14
WASH	300	130	0.05	0.88
Total	1457	689	0.01	1.01

Notes: This table reports summary statistics of the residual sector outcomes for World Bank projects, estimated independently for each sector according to equation 2. The residual terms are normalized scalar values. Values for Fiscal projects are not reported as those were not included in the project-level non-linear models, given the positive result for the sector in the ratings models.

Table B.9: Prediction Results

Model	ROC AUC	R^2
Linear (Lasso)	0.500	0.000
Linear (Ridge)	0.603	0.076
Ensemble (RF)	0.672	0.564
Ensemble (XGB)	0.695	0.764
Neural Network	0.534	0.197
SVCs	0.589	0.273
Ensemble (RF, at approval)	0.700	0.861

Notes: RF = Random forest, XGB = gradient boosted trees, SVC = support vector classifier, ROC AUC = receiver-operator area under curve

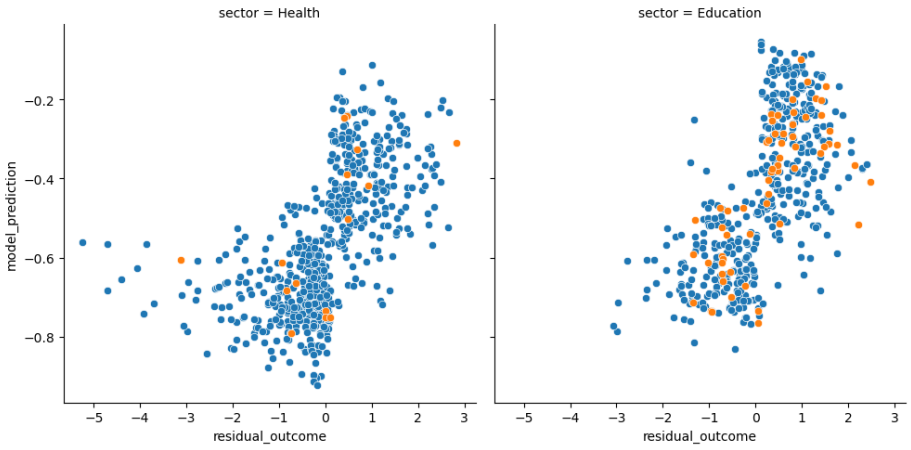

Figure B.3: Outcome keyword presence in PDOs compared to residual outcomes and model predictions

Appendix C. Non-linear Results

Table C.10: Divergence between Residual Outcomes and Normalized Ratings, Per Region and Loan Size Tertile

Region	Loan size	Avg gaming prediction	Avg prob gamed	Avg gaming
Africa East	large	0.90	0.70	0.75
Africa West	large	1.36	0.77	1.64
East Asia and Pacific	large	0.98	0.77	1.01
Europe and Central Asia	large	1.20	0.76	1.22
Latin America and Caribbean	large	0.76	0.72	0.69
Middle East and North Africa	large	0.05	0.53	-0.16
South Asia	large	0.62	0.65	0.62
Africa East	medium	0.94	0.68	0.94
Africa West	medium	1.03	0.73	1.11
East Asia and Pacific	medium	0.50	0.58	0.48
Europe and Central Asia	medium	1.04	0.73	1.09
Latin America and Caribbean	medium	0.71	0.68	0.73
Middle East and North Africa	medium	0.16	0.54	0.02
South Asia	medium	0.49	0.63	0.42
Africa East	small	-0.74	0.20	-0.85
Africa West	small	-0.80	0.17	-0.69
East Asia and Pacific	small	-1.01	0.11	-0.96
Europe and Central Asia	small	-0.75	0.10	-0.67
Latin America and Caribbean	small	-0.99	0.08	-1.04
Middle East and North Africa	small	-1.14	0.10	-1.64
South Asia	small	-0.95	0.13	-1.01

Notes: Loan size corresponds to observed tertiles of loan size. "Avg gaming prediction" = non-linear model's predicted difference between normalized average rating and normalized sector outcomes (lagged). "Avg prob gamed" = prediction of likelihood that a project has a larger than average difference between its normalized rating and normalized sector outcomes. "Avg gaming" = observed difference between normalized rating and normalized sector outcomes

Appendix D. Non-linear Methods

Table D.11: Numeric Features

Feature	Unit	Description
Original Commitment	USD	Size of loan or grant at approval (in constant 2015 dollars)
Project Duration	Months	Original intended duration of project
CPIA	$1-6$	WB Country Policy and Institutional Assessment for implementing country at project approval
GDP per capita	USD	GDP per capita in constant PPP (at approval FY), log scale
Prep TTL experience	Projects	Number of prior projects prepared by the project's task team leader
Prep TTL "value add"	(VA)	Preparing TTL "value add" in relation to project ratings
Country Director VA	(VA)	Project rating value addition of country director at time of approval
Sector Manager VA	(VA)	Project rating value addition of sector manager at time of approval
Sector Percentage	%	Project budget allocated to primary sector
Number Sectors		Number of sectors the project spans
Sector HHI	HHI	Herfindahl-Hirschmann Index of budget allocations across project sectors
Freedom House Index	Index	Freedom House index for implementing country at time of approval

Table D.12: Categorical Features

Feature	Categories	Description
Financing instrument	IPF, DPL, others	The type of financing used for the project
Funding source	IBRD, IDA, blend	Source of funding within World Bank
Region	Africa East, South Asia, etc.	World Bank region in which the project fell at approval
Primary Sector	Health, Education, etc.	The project's primary sector
Fragile/Conflict	Binary	Whether the implementing country was fragile or post-conflict at approval

Table D.13: Algorithms and hyper-parameters tested for project prediction

Algorithm	Varieties	Hyper-parameters
Linear Models	Linear, Logistic Lasso	L1 term multiplier
Support Vector Machines	Support Vector Classifier (SVCs) and Support Vector Regressor (SVRs)	Regularization term, kernel types
Ensemble Trees	Random Forest (RF)	Minimum samples in leaf, maximum depth
Gradient Boosting	XGBoost	Learning rate, minimum child weight
Neural Network	Multilayer Perceptron (MLP)	Hidden layer sizes, regularization