When Private Firms Provide Public Goods: The Allocation of CSR Spending

Kim Fe Cramer, Lucie Gadenne, and Noémie Pinardon-Touati*

This paper studies how firms allocate Corporate Social Responsibility (CSR) spending to understand the welfare effects of corporate contributions to public goods. Using new data on the quasi-universe of CSR projects in India, we document key facts on the allocation of CSR across social topics (e.g., health, education) and locations. We use natural language processing to measure the technological proximity between firms' production technology and CSR topics, and we find that firms spend more on topics in which they have a technological advantage. This is consistent with an efficient allocation of CSR across topics and the main rationale for CSR in the theoretical literature. Across locations, however, we find that firms spend more in areas where social returns are likely lower. Overall, our results suggest that CSR mandates may be an efficient but inequitable way to increase public goods provision.

Keywords: private provision of public goods, CSR, textual analysis, India
JEL Codes: D64, L21, H41, M14

*Kim Fe Cramer: LSE (k.f.cramer@lse.ac.uk), Lucie Gadenne: Queen Mary University of London, IFS, and CEPR (l.gadenne@qmul.ac.uk), and Noémie Pinardon-Touati: Columbia University (np2842@columbia.edu). We thank Vimal Balasubramaniam, Giorgia Barboni, Michael Best, Nicolas Bonneton, Hans Christensen, Michele Fioretti, Simon Franklin, Maitreesh Ghatak, Moqi Groen-Xu, Sean Higgins, Jessica Jeffers, Anders Jensen, Cynthia Kinnan, Nicola Limodio, Karthik Muralidharan, Jordan Nickerson, Yanos Zylberberg, as well as seminar audiences and participants at the Zurich Public Finance Conference on Developing Countries, CEPR Public Finance Workshop, EBRD, City University, Indian School of Business, Northwestern Kellogg Innovations in Sustainable Finance Conference, King's College London, London Junior Finance Workshop, LSE, Oxford University, Paris School of Economics, QMUL, Reichman University, Sciences Po Sustainable Finance Workshop, Tinbergen Institute, University of London, WAPFIN, WEFIDEV, WashU Annual Finance Conference, and Young Scholars' Webinar, for helpful comments and suggestions. We also thank Marco Guttierez Chavez, Aristomenis Chrysafis-Progopoulos, Cécile Delcuvellerie, Nithin Mannil, Francine Montecinos, and Bruno Yzeiri for excellent research assistance.

1 Introduction

Firms around the globe spend large amounts on corporate social responsibility (CSR) activities (Benabou and Tirole, 2010; Hart and Zingales, 2017; Fioretti, 2022). In the developing world, governments are increasingly delegating public good provision to firms by mandating them to spend on CSR. In India, the first country to implement such a mandate, CSR expenditures represent 0.1% of GDP. But whether firms' CSR activities are ultimately beneficial for society - or merely a diversion from their core economic role - has remained contested since the early debates on the objectives of the firm (Lund, 2023).

The standard economics view on CSR, defined as the allocation of some profits to social causes, is that contributions to such causes are best done by individuals, not firms (Friedman, 1970). Firms could moreover engage in CSR purely for strategic reasons that lead them to maximize their private returns, potentially at the expense of social returns(Baron, 2001). A more positive view of CSR can be found in a theoretical literature that argues that CSR can be welfare-improving if firms have a *technological advantage* in producing public goods relative to the public or non-profit sector (Besley and Ghatak, 2007; Hart and Zingales, 2022). This occurs when the public good is 'bundled' with the production of the private good: a healthcare firm may for example have a technological advantage in running a health screening campaign. There is however no systematic evidence on how firms allocate their CSR expenditures that could help understand the welfare effects of CSR.

This paper seeks to fill this gap by studying how Indian firms allocate their CSR expenditures. India is a particularly good context in which to study this, for two main reasons. First, India is a large emerging economy facing substantial demand for public goods with limited tax capacity (Das et al., 2023; Muralidharan, 2024). Whether CSR can be an effective mechanism for delivering public goods is thus a question of major interest. Second, in 2013, India mandated that all large firms spend a minimum share of their profits on a specified list of social causes and report on all their CSR projects. This enables us to construct the first dataset on the quasi-universe of CSR activities in any economy: we observe all CSR expenditures of the 6,500 largest Indian firms over the period 2015–2019, and a detailed description of each CSR project.

We begin by presenting key facts about CSR in India. First, the allocation of CSR across social topics (e.g. education, health) is similar to how other public good

providers allocate their expenditures. Second, firms specialize in topics, and firms producing similar products make similar CSR choices. This suggests a potential relationship between firms' technologies and their CSR allocation. Third, CSR expenditures per capita are highly unequally distributed across space.

Motivated by these facts, we build a conceptual framework in which firms allocate their CSR expenditures across project types, defined by a topic and a location. Firms differ in the technology they use to provide projects in each type to capture the possibility that they have different technological advantages across types; we propose one micro-foundation for why these technological advantages arise. They also have heterogeneous preferences across types that can differ from those of the social planner, allowing for wedges between private and social returns. We contrast the socially and privately optimal allocations to clarify what can be learned by our empirical exercises.

We then consider whether CSR is efficiently allocated across topics: do firms spend more on projects they have a technological advantage in? Our key methodological innovation is to use Natural Language Processing (NLP) to construct an index of technological proximity between the firms' for-profit activity and CSR topics. Our method compares two types of texts: we use the text in the industry classification guidelines to describe firms' for-profit technologies and the description of CSR projects in a social topic to describe the technology required to produce CSR projects. We measure the proximity between industries and social topics using the cosine similarity between the vectors of word embeddings for these two types of texts. We provide support for our key assumption: similarity in the word embedding space reflects technological similarity, where technologies are used to produce both for-profit goods and CSR projects.

We find that firms' technological advantage is correlated with how they allocate CSR across topics: a one standard deviation increase in the proximity between a firm's industry and a topic increases the amount that the firm spends on the topic by 16%.[1] These results are robust to a wide range of robustness checks, including using alternative NLP models or textual sources, and are not driven by a particular topic or industry. Findings are also very similar when obtained on the sample of firms that spend more on CSR than the amount required by the law; this suggests our

[1] Our conceptual framework clarifies that what we are interested in is the *correlation* between technological advantage and CSR allocation, not a causal effect. Whether firms spend more on topics they have a technological advantage in because of their technology or because they have a high preference for these topics is irrelevant from an allocative efficiency perspective.

conclusions may not hinge on CSR being mandated and not voluntary. Our results are thus consistent with the idea that firms use their technological advantage when deciding how to allocate their CSR spending. Seen through the lens of the theoretical CSR literature (Besley and Ghatak, 2007; Hart and Zingales, 2022), this implies that CSR has the potential to be welfare-improving.

Firms allocating expenditures on public goods may however matter for equity as well as efficiency. Turning to to the allocation of expenditures across locations, we find that CSR expenditures in an area are positively correlated with that area's level of development. Assuming that public goods have higher social returns in poorer areas, this indicates a wedge between private and social returns. A key mechanism behind this finding is that firms concentrate their CSR spending in areas where they are headquartered. However, CSR expenditures are positively correlated with local development even when spending in headquarters is excluded. Overall, our findings indicate that the spatial distribution of CSR spending is regressive, including when compared to the distribution of government expenditures.

Our first contribution is to the empirical literature on CSR. Most of this literature focuses on the relationship between firms' CSR activities and their financial outcomes (see Margolis et al., 2007; Christensen et al., 2021; Gillan et al., 2021; Hong and Shore, 2023; Starks, 2023, for reviews).[2] More recent contributions focus on measuring firms' social impact or characterize the diverse stakeholder preferences underpinning firms' prosocial stances.[3] We build on this literature by leveraging data on CSR projects to inform the welfare properties of corporate contributions to public good provision. Our approach in particular complements that in Fioretti (2022), who uses detailed data on all activities of one firm to show that it acts prosocially beyond profit maximization. In contrast, we consider the quasi-universe of CSR spending in a context in which the amount of prosocial spending is given, and consider whether its *allocation* is consistent

[2]Previous contributions have investigated CSR in the Indian context in particular. In the accounting literature, Manchiraju and Rajgopal (2017); Dharmapala and Khanna (2018); Mukherjee et al. (2018); Bhattacharyya and Rahman (2019) investigate the effect of the CSR mandate on firm value, focusing on listed firms. In the strategy literature, Gatignon and Bode (2023) provide a descriptive analysis of Indian firms' CSR strategies. Chhaochharia et al. (2025) consider the mandate's effect on school enrollment. Rajgopal and Tantri (2023) examine the impact on firms spending voluntarily on CSR.

[3]See e.g., Flammer and Luo (2017); Bertrand et al. (2020); Gibson Brandon et al. (2022); Allcott et al. (2023); Cheng et al. (2023); Christensen et al. (2023); Colonnelli et al. (2023); Fioretti et al. (2023); Hartzmark and Shue (2023); Kahn et al. (2023); Conway and Boxell (2024); Green and Vallee (2024); Ferreira and Nikolowa (2025).

with welfare maximization.

In particular, we propose an empirical test for a key assumption in the theoretical literature on CSR: that firms allocate their CSR spending according to the technological advantage that stems from their for-profit production technology (see Kitzmueller and Shimshack, 2012, for a review). In most models, it is a necessary condition for CSR to increase welfare – that of shareholders (Hart and Zingales, 2017, 2022) or of society (Besley and Ghatak, 2007; Magill et al., 2015; Broccardo et al., 2022).[4] This paper is, to the best of our knowledge, the first to test and validate this assumption.

Our second contribution is to the literature on the private provision of public goods. This literature focuses mostly on private provision via privatization or outsourcing (see Hart et al., 1997; Kotchen, 2006; Behaghel et al., 2014; Mukherjee, 2021; Knutsson and Tyrefors, 2022) and has thus far not studied CSR expenditures. It emphasizes tradeoffs between the efficiency gains of private provision and adverse effects in quality or distributional outcomes. Our results point to a similar efficiency-equity tradeoff for CSR activities. This paper's scope is also reminiscent of work on charitable giving studying the universe of charitable giving by individuals via administrative data, though this literature has so far focused on rich countries (see List, 2011, for a review).[5]

Third, our results speak to debates regarding how to finance development. A large literature considers how governments can raise more resources in low- and middle-income countries (LMICs) given tax capacity constraints (see for example Besley and Persson, 2009; Best et al., 2015; Gadenne, 2017; Jensen, 2022; Bergeron et al., 2024). Our results suggest that mandating CSR spending can complement tax-raising efforts, and indeed, several LMICs recently implemented CSR mandate laws similar to India's (Lin, 2021). Our aim is not to compare the CSR mandate to an increase in taxes on large Indian firms. We show, however, that the mandate was well enforced, with an economically significant increase in CSR expenditures. The public framing of the

[4]Note that this condition is necessary but not sufficient for CSR expenditures to increase social or shareholder welfare – one also needs to assume government under-provision of the public good and, for shareholder welfare, shareholder preferences for being socially responsible. CSR expenditures could also increase shareholder welfare if shareholders look to management to solve their free-riding problem (Morgan and Tumlinson, 2019), common ownership leads shareholders to want to maximize industry profit, not firms' profit, etc. Hart and Zingales (2022) however, argue that such considerations are second-order explanations compared to the technological advantage motivation.

[5]Within this literature Card et al. (2010) find that charitable contributions from individuals increase substantially in areas where a firm's headquarters are located. Our results suggest corporate contributions exhibit a similar type of home bias.

law as asking firms to contribute to development goals, together with the reporting requirements, de facto led to a transfer of resources from the private sector to public good provision in a context where tax enforcement itself is relatively weak.

Finally, our methodology builds on a growing literature using semantic distance to capture distance in economically relevant space (Gentzkow et al., 2019; Ash and Hansen, 2023) and in particular Hoberg and Phillips (2010, 2016) who use textual analysis to characterize the product space in which firms compete. We derive model-based tests to validate that word embeddings adequately capture technological properties of firms' production processes.

The paper is organized as follows. Section 2 describes our context of study, data, and provides evidence on the implementation of India's CSR mandate. Section 3 provides key stylized facts regarding the allocation of CSR expenditure in our context that motivate the simple conceptual framework that defines our hypotheses of interest in section 4. Section 5 considers the efficiency properties of the allocation of CSR expenditures across topics, whilst section 6 studies the equity characteristics of the allocation across locations.

2 Context and Data

2.1 Corporate Social Responsibility in India

In August 2013, India passed into law section 135 of the Companies Act, which mandates that large firms spend at least 2% of their average profits over the last three years on CSR activities. It came into effect in April 2014. Large firms are defined as those with profits above INR 50 million, income above INR 10 billion, or a net worth above INR 5 billion in any of the three preceding financial years.[6] These firms represent a large share of the Indian economy, corresponding to approximately 60% of formal sector activity. The act specifies the activities that qualify as CSR expenditures, clarifies that spending occurring within the 'normal course of business' (e.g., employee welfare) does not qualify, and imposes the formation of a CSR committee with at least one independent director. Importantly for our purposes, it also makes reporting of all CSR activities to the Ministry for Corporate Affairs (MCA) compulsory. During our study period (2015–2019), the mandate was enforced on a

[6]These thresholds are not associated with any other requirements in Indian law.

comply-or-explain basis, and since 2019 fines have been imposed for non-compliance. For more details on the provisions of the mandate, see Appendix D.1. Over the period 2015–2019, the total annual CSR expenditure is 142,669 million INR (2,283 million USD) on average, equivalent to 0.1% of GDP.[7] We return to discussing the effects of the law on CSR expenditures after describing our data.

2.2 Data

CSR data. Our main data source comes from the compulsory reporting of CSR activities to the MCA. Since the fiscal year 2014–2015 (hereafter 2015), all liable firms report on each of their CSR projects. The data is available on the MCA website and contains, for each CSR project, the amount spent on the project, the CSR topic this project belongs to (from a pre-specified list defined in the law), and a textual description of the project. After cleaning procedures outlined in Appendix C, the CSR data contains information on 124,813 projects conducted by 11,487 firms over the period 2015–2019. From the 28 CSR topics specified by the law and available in the CSR data, we group similar topics to obtain the 16 topics considered in our analysis.[8] Because projects often span multiple years, we aggregate data across years in what follows. To the best of our knowledge, this is the most comprehensive dataset on CSR activities for any country in the world. It is comparable in scope to data on charitable giving by individuals in the US compiled by the Giving USA Foundation (see List, 2011) but contains more information on project types and, crucially for our purposes, provides the official Corporate Identification Number (CIN) of each firm.

Accounting data. We combine CSR data with accounting data to obtain additional information on firms. We use the Prowess database from the Center for Monitoring the Indian Economy, which includes information from the income statements and balance sheets of all publicly traded firms as well as a large number of private firms. From this data, we obtain information on firms' industries at the 2-digit level,

[7] This is similar to the share observed in the US where charitable giving by corporations represented 27.36 billion USD in 2023, just under 0.1% of GDP (The Giving Institute, 2023). Throughout the paper, we denominate in 2015 INR and apply an INR to USD exchange rate of 0.016.

[8] This re-classification is done by using the information provided by the project descriptions to group together topics that contain few projects and are conceptually very similar, e.g., 'environmental sustainability' and 'conservation of natural resources'. See Appendix C.1.4 for a detailed description of our method.

which follows the National Industry Classification (NIC). We use information on CSR expenditures reported in balance sheet statements to examine the implementation of the reform, using data from 2007 onward.

We merge the CSR and accounting data at the firm level using firms' CINs. The accounting data does not, by design, include all Indian firms and has better coverage of large firms. We match 61% of firms and 91% of CSR expenditures in the CSR data as well as 99% of the post-2015 CSR expenditures reported in the accounting data. After the merge and a second set of cleaning steps outlined in Appendix C, our main analysis sample consists of 86,334 projects by 6,573 firms.

Table A.1 shows descriptive statistics of the variables used in our analysis. Firms are large, and the distribution has a long right-tail. We systematically present results with and without weighting by firm size (proxied by total CSR expenditures) in what follows. Figure A.1 plots the distribution of CSR expenditures by industry for the 20 largest industries in our data. The CSR shares follow a distribution similar to that of value-added per industry for India, as expected given that CSR expenditures are a function of profits.

Textual data. Our main analysis exploits the description of projects in the CSR data. To use this text, we perform a number of data cleaning steps described in Appendix C.1.2. In particular, we filter out uninformative tokens (words) and observations. After cleaning, the average project observation contains 4.3 informative tokens (standard deviation is 4.0) and the average CSR topic contains 23,509 tokens (standard deviation is 30,763).[9] Figures C.2–C.4 provide a visualization of this data by showing word clouds for the project descriptions by topic.

We encode the textual data using word embeddings. Word embeddings are a natural language processing method in which individual words are represented as real-valued vectors in a high-dimensional space. These vectors are meant to capture the meaning of words so that similar words have similar vectors. In addition, an internally consistent geometry on the vector space allows words to be related. We use the word embeddings provided by the pre-trained Word2Vec model released by Google. The model contains 300-dimensional vectors for 3 million words. We obtain a vector representation of the text describing each CSR project p, denoted $\mathbf{v}_p$. The details of the implementation of Word2Vec are in Appendix C.1.3. We use Word2Vec

[9]Before filtering out uninformative tokens, the average is 7.4 tokens per observation.

embeddings as our baseline due to their methodological transparency: representing documents as the mean of token-level embeddings enables intuitive and interpretable exploration of semantic patterns. We present results using OpenAI's NLP model as a robustness check.

In section 5, we additionally exploit textual data characterizing the firms' industries. For each 2-digit industry, the Handbook of the National Industrial Classification provides a description of the products and production technologies common to firms in the industry. After cleaning, this text yields an average of 250 informative tokens per industry (standard deviation is 225). Table C.4 shows an example of text for one industry. Using word embeddings, we obtain a vector representation of the text describing each industry i, denoted $\mathbf{v}_i$. In robustness tests, we exploit SEC 10-K filings for US firms to produce an alternative corpus describing industries (see Appendix C.2 for details). The magnitude of embedding vectors does not encode meaningful information so we normalize all embedding vectors to have norm 1.

Notations. Throughout the text, we use the notation $\|\mathbf{x}\|$ to denote the Euclidean norm of vector $\mathbf{x}$ and $\cos(\mathbf{x}, \mathbf{y})$ for the cosine similarity between vectors $\mathbf{x}$ and $\mathbf{y}$:

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}.$$

Additional variables. We use two additional project-level variables from the data. The first is an indicator for whether the project was implemented directly by the firm or indirectly via a third party (typically an NGO). The second is information on project location: we observe the state in which a project occurs for projects totaling two-thirds of the CSR spending; among those we obtain the district for 27% of expenditures. We observe CSR expenditures in all 35 states and 496 districts (78% of all districts).[10]

2.3 Implementation of the CSR Mandate

This section briefly describes evidence on the implementation of the CSR mandate. In Figure 1(a), we plot the evolution of total CSR expenditures in India over time, as

[10] When we do not observe the state this is either because the variable isn't filled in the original data or because the project is specified as occurring in the whole of India. In our regression analysis in section 6 below, we exclude four states with a population that is lower than one million, as well as the small state of Chandigarh, which does not have government spending data.

reported in the accounting data. We see a large increase from 2015 onward: aggregate CSR spending roughly tripled since the mandate was implemented. Figure 1(b) plots the evolution of CSR spending as a share of profits separately among liable firms (defined as firms whose income, profits, or net worth are above the thresholds defined in the law) and all other firms in the accounting data. All the aggregate increase in CSR spending comes from liable firms. Appendix D investigates the evolution of CSR expenditures in liable and non-liable firms over the period more formally by conducting a difference-in-differences exercise (see in particular Figure D.1) and considers whether pre-existing expenditures were re-labeled as CSR after 2015. Results suggest the mandate led to an increase in CSR spending as a share of profits of 1 percentage point in the first year of its implementation, and up to nearly 1.5 points at the end of the period. The average profit share of CSR among liable firms is 2.3% at the end of the period, suggesting the mandate was well respected overall (see also Chhaochharia et al., 2025).[11]

Figure 1(b) also shows that many firms already spend on CSR prior to the mandate: 14% of firms in our main sample spend more than 1% of profits on CSR in 2014, 20% spend more than 2.5% after 2015. Together, these two sets of firms represent 26% of our sample and constitute what we call the 'voluntary CSR' sample. Their behavior may indicate intrinsic preferences for CSR activities. In what follows, we systematically consider whether these firms allocate their expenditures differently.

3 Key Facts About CSR in India

This section documents four key facts on the allocation of CSR spending by firms in India. These facts motivate our analysis of the efficiency-equity trade-off associated with firms deciding on the allocation of public goods. We obtain them on our main analysis sample and systematically reproduce them on i) the sample of all firms in the CSR data and ii) the voluntary CSR sample; we find very similar patterns in Appendix B.[12]

[11] We also test whether the increase in CSR expenditures crowded-out government expenditures at the state level, but are under-powered to conclude with certainty on that question (see Table D.3).

[12] Facts 1 and 4 on the allocation across topics and states have overlaps with official CSR reports (Ministry of Corporate Affairs, 2021) and the descriptive analysis in Gatignon and Bode (2023).

Fact 1: CSR spending is concentrated in health and education. Table 1 shows the allocation of CSR spending across topics. This table also clarifies the meaning of the topics by listing the three most common project types within each topic. For each topic, the most common project types are identified by partitioning projects into types by estimating a k-means clustering algorithm on project embeddings.[13] We see that firms finance a wide range of projects.

The largest social topic in terms of spending is education (32% of the total). Common education projects involve school construction or renovation and the promotion of education for differently-abled children. The second largest topic is health (17% of spending), with projects focused on preventive healthcare, patient care, medical equipment, or mobile health camps. Infrastructure and environmental sustainability follow, with 8% of spending each. Infrastructure involves mostly small-scale infrastructure in rural areas (e.g., rural roads, street lights); for environmental sustainability, conservation and tree plantation are the most frequent types of project. The other social topics all receive equal to or less than 6% of spending. Section 5 below considers the determinants of firms' allocation of CSR across topics.

Fact 2: Firms' allocation across topics correlates with the allocation of other public good providers. Figure 2 compares the allocation of CSR spending across topics to the allocation of spending by other key public goods providers: the government and NGOs. To make this comparison feasible, we aggregate several topics together. We leave the details of the mapping between the CSR topics and the spending categories for other providers, as well as the respective data sources, to Appendix C.2.

The allocation of CSR spending across topics is significantly correlated with that of government spending and NGO activity. In both cases, the pairwise correlation is around 0.8 and statistically significant at the 1% level. Notably, the three types of public goods providers allocate almost precisely the same share to education. Firms differ from the government and NGOs in that they allocate less to vulnerable populations and more to industry and technology, vocational skills, and water and sanitation projects. Overall, Figure 2 suggests that different public good providers agree to a large extent on the relative valuation of public goods across topics. We return to

[13] See the full list of clusters and implementation details in Appendix C.3. In addition, Figures C.2–C.4 show word clouds for each topic.

this when we compare the role of technological advantage in explaining the allocation of CSR across topics to that of preferences for topics common to firms and the government.

Fact 3: Firms specialize in topics, and similar firms specialize similarly. Firms' CSR spending is highly concentrated across topics. The firm-level distribution of spending shares across topics has an average Herfindahl-Hirschman Index equal to 0.63, and 34% of firms allocate more than 90% of their spending to only one topic. This is not only the result of indivisibilities: in the sample of firms reporting multiple projects, 20% of firms allocate more than 90% of their spending to only one topic.

Moreover, firms that sell similar products tend to allocate their CSR expenditures across topics similarly. We regress the cosine similarity between firms' CSR shares across topics on the cosine similarity between firms' sales shares across products and find a statistically significant correlation (see Table A.2). These specialization patterns suggest a link between firms' for-profit production processes and their choice of CSR spending. This is a key building block of our conceptual framework in section 4.

Fact 4: CSR spending is highly concentrated geographically. Figure 3 shows the distribution of CSR expenditures across states. Almost 30% of CSR spending funds projects in the state of Maharashtra. Six states (Maharashtra, Karnataka, Gujarat, Tamil Nadu, Andhra Pradesh, and Delhi) receive 66% of the spending. This does not simply reflect the distribution of the population: Maharashtra represents only 9% of the population and these six states 34%. This concentration of CSR spending in a few states thus leads to large discrepancies in CSR spending per capita. In section 6, we explore both the determinants and the implications of the geographical allocation of CSR expenditures.

4 Conceptual Framework

This section provides a simple conceptual framework that compares firms' privately optimal CSR allocation to the socially optimal allocation to guide our empirical analysis. We are interested in the allocation across project types, which we define in our empirical analysis as either topics or locations. Firms differ in their preferences across

types and in the type-specific production function they use to produce projects from CSR expenditures. The latter captures the idea that firms may have a technological advantage in producing some public goods. We start with a general model where this technological advantage is exogenously given, then propose a micro-foundation in which it arises because firms are endowed with technologies used in the production of both private goods and CSR projects.

4.1 General Model with Exogenous CSR Productivities

Set-up. Our object of interest is how firms f allocate an exogenous CSR amount E across project types $p \in \mathcal{P}^{pub}$. We denote s_{fp} the share that firm f allocates to type p. The amount of project type p produced by firm f is given by:

$$y_{fp} = \exp(\alpha_{fp})(s_{fp}E)^{\rho} \tag{1}$$

where $\rho < 1$. The parameter α_{fp} captures firm f's technological advantage in providing type p.

Firms obtain utility U from their projects y_{fp}, defined in the following way:

$$U_f = \sum_p \zeta_{fp} y_{fp} \tag{2}$$

where the ζ_{fp} terms capture firm preferences across project types p which could reflect both private returns (strategic considerations) and/or warm-glow utility.

Social welfare is an increasing function of the projects funded by all firms and is defined as follows:

$$W = \sum_p \mu_p \sum_f y_{fp} \tag{3}$$

where the μ_p terms capture the social returns to project type p.

Socially optimal allocation. Maximizing social welfare in expression (3) subject to $\sum_p s_{fp} = 1, \forall f$ yields:

$$s^*_{fp} = \frac{\left[\mu_p \exp(\alpha_{fp})\right]^{1/(1-\rho)}}{\sum_q \left[\mu_q \exp(\alpha_{fq})\right]^{1/(1-\rho)}} \tag{4}$$

The socially optimal amount firm f allocates to a project type p is increasing in its (relative) technological advantage in this type, α_{fp}, and in the social returns parameter μ_p. We define a CSR allocation as *allocatively efficient* if firms with a higher technological advantage on a project type spend more on that type. The socially optimal allocation satisfies allocative efficiency.

Privately optimal allocation. Each firm maximizes its utility in expression (2) subject to $\sum_p s_{fp} = 1$. This yields:

$$s_{fp} = \frac{\left[\zeta_{fp} \exp(\alpha_{fp})\right]^{1/(1-\rho)}}{\sum_q \left[\zeta_{fq} \exp(\alpha_{fq})\right]^{1/(1-\rho)}} \tag{5}$$

When firms internalize social welfare ($\zeta_{fp} = \mu_p, \forall f, \forall p$) or have no preferences across project types ($\zeta_{fp} = \zeta_f, \forall p$), allocative efficiency holds, and firms spend more on project types they have a technological advantage in. However, when firms' preferences across types are different from those of the social planner, allocative efficiency may not hold. In particular, if the correlation between firms' preferences across types and their technological advantage across types is negative and large, allocative efficiency does not hold.

4.2 Micro-Foundation of CSR Productivities

Why are some firms more productive at some CSR project types? We propose a simple micro-foundation in a model where firms are endowed with multi-dimensional technology vectors used to produce both for-profit and CSR projects, and different projects require different combinations of technologies.[14] We provide a succinct description of the model, and leave details and derivations to Appendix E.

Production. Production occurs across projects, indexed by $p \in \mathcal{P}$. $\mathcal{P}$ can be partitioned into $\mathcal{P}^{pub}$ for CSR projects and $\mathcal{P}^{pri}$ for for-profit goods sold in competitive markets. Projects are produced by combining tasks $\tau \in \mathcal{T}$. The relative importance of tasks varies across projects as characterized by the vector $\boldsymbol{\phi}_{\mathbf{p}} = [\phi_{p\tau}]_{\tau \in \mathcal{T}}$ such

[14] Our modeling choices borrow heavily from the task-based production framework (e.g., Acemoglu and Restrepo, 2018), but the key insights are applicable to a general multi-dimensional sorting framework (e.g., Lindenlaub, 2017).

that $\forall p,\ \sum_{\tau} \phi_{p\tau} = 1$. To perform tasks, firms hire workers and are endowed with a task-specific productivity vector $\mathbf{z_f} = [z_{f\tau}]_{\tau \in \mathcal{T}}$. Firm f's output in project p is:

$$y_{fp} = \left[\prod_{\tau \in \mathcal{T}} (\exp(z_{f\tau}) \ell_{fp\tau})^{\phi_{p\tau}} \right]^{\rho}$$

where $\rho < 1$ as above, and $\ell_{fp\tau}$ is the labor assigned to task τ and project type p. The wage w is taken to be exogenous.

$\boldsymbol{\phi_p}$ captures the technological requirements of project p. We interpret tasks in a broad sense as any type of work, material, or immaterial input that are required to complete a project. $\mathbf{z_f}$ captures firm productivity at each of the tasks, and can capture physical capital, organizational capital, any type of specialized knowledge, know-how, or information. We assume for convenience that $||\boldsymbol{\phi_p}|| = \phi, \forall p$ and $||\mathbf{z_f}|| = 1, \forall f$, and relax these assumptions in Appendix E.

Project-specific productivity. Conditional on producing project p, the firm allocates labor ℓ_{fp} across tasks in a way that maximizes project-specific returns:

$$\pi_{fp} = \max_{\{\ell_{fp\tau}\}} \zeta_{fp} y_{fp} - w \ell_{fp}$$

subject to: $\sum_{\tau \in \mathcal{T}} \ell_{fp\tau} = \ell_{fp}$. When p is a for-profit good, ζ_{fp} is the market price of good p. When p is a CSR project ζ_{fp} captures how much firm f values project type p, as above. The optimal labor allocation satisfies: $\frac{\ell_{fp\tau}}{\ell_{fp}} = \phi_{p\tau}$. Project-level profit maximization yields an expression for project output as a function of firm-project productivity α_{fp} :

$$y_{fp} = \exp(\alpha_{fp}) \ell_{fp}^{\rho}, \tag{6}$$

with:

$$\alpha_{fp} = \rho \phi \cos(\boldsymbol{\phi_p}, \mathbf{z_f}) + \kappa_p \tag{7}$$

$\cos(\boldsymbol{\phi_p}, \mathbf{z_f})$ is the cosine similarity between the vector of returns to tasks for that project type $\boldsymbol{\phi_p}$ and the firm's vector of task-specific technologies $\mathbf{z_f}$. κ_p is a project-level constant.

Privately and socially optimal allocations. The firm allocates labor across projects to maximizes total returns:

$$\max_{\{\ell_{fp}\}} \sum_{p \in \mathcal{P}^{pri}} (\zeta_{fp} y_{fp} - w\ell_{fp}) + \sum_{p \in \mathcal{P}^{pub}} (\zeta_{fp} y_{fp} - w\ell_{fp})$$

where we now define the exogenous CSR expenditure requirement as $\sum_{p \in \mathcal{P}^{pub}} w\ell_{fp} = E$. We obtain that firms' privately optimal CSR shares follow the expression in equation (5), with α_{fp} now defined in (7).

The social planner maximizes the social welfare function given in (3) above. This yields the same socially optimal allocation across CSR project types p as above (equation (4)).

Sources of technological advantage. Equation (7) shows that firms are more productive in projects that put a large weight on their most productive tasks. Productivity at the firm×project level increases in the proximity between firms' own technologies and projects' technological requirements. This is in line with resource-based theories of the firm proposing that firms diversify into products that use common capabilities, supported by empirical evidence in Boehm et al. (2022) and Koh and Raval (2025).

This framework implies that firm's projects, both private and public, are determined by a common firm-level productivity vector. A direct implication is that firms making similar for-profit product choices have a similar vector $\mathbf{z_f}$, and hence make similar CSR allocation choices. This is in line with the evidence in Table A.2 discussed in the previous section.

Implications for CSR productivities across industries. Our empirical tests exploit variation in firms' technologies at the level of their industry. We define an industry as a set of private goods $\mathcal{P}_i \subset \mathcal{P}^{pri}$ centered around a technological vector $\boldsymbol{\phi_i}$: $\forall p \in \mathcal{P}_i, \boldsymbol{\phi_p} = \boldsymbol{\phi_i} + \boldsymbol{\epsilon_p}$, with $\boldsymbol{\epsilon_p}$ mean-zero and i.i.d. In addition, we assume that firms belonging to industry i have a productivity vector centered around $\boldsymbol{\phi_i}$: $\mathbf{z_f} = \boldsymbol{\phi_i} + \boldsymbol{\epsilon_f}$. This ensures that firms in industry i have $p \in \mathcal{P}_i$ as their main product, in line with how firms are classified across industries in standard datasets.

For a firm f in industry i, we can then write:

$$\alpha_{fp} = \rho\phi \cos(\boldsymbol{\Phi_p}, \boldsymbol{\Phi_i}) + \kappa_p \tag{8}$$

Firms are more productive in CSR projects that have technological requirements similar to that of their for-profit industry.

4.3 Hypothesis Taken to the Data

In what follows, we start by considering how CSR expenditures are allocated across one dimension of project type, that of social topic indexed by d (section 5). We do not impose any shape on the distribution of social returns across topics and test whether the private allocation is allocatively efficient by considering whether $\frac{ds_{fd}}{d\alpha_{fd}} > 0$. Our object of interest is thus the correlation between firms' technological advantage across topics and the share they spend on topics: allocative efficiency holds if this correlation is positive. In line with our micro-foundation, we measure technological advantage as the proximity between a firm's for-profit technology and the requirements of CSR projects, as detailed below. We allow for any pattern of aggregate firm preferences towards topics by including topic fixed effects and consider the extent to which these can be explained by preferences for topics that are shared by firms and the government as a second step.

We then consider in section 6 how CSR spending is allocated across the other dimension of project type, locations, indexed by s. There, we use a location's economic development to proxy for the social returns to spending, and test whether $\frac{ds_{fs}}{d\mu_s} > 0$ by looking at the correlation between economic development and CSR expenditures across locations.

5 Allocative Efficiency: Do Firms Use Their Technological Advantage?

Motivated by our definition of allocative efficiency above, we consider whether firms spend more on topics for which their for-profit production processes give them a technological advantage. We begin by explaining how we proxy for technological advantage, then outline our empirical strategy and present our results.

5.1 Construction of a Proxy for Technological Advantage

Testing for allocative efficiency requires a measure of firm×project-specific CSR productivities. We exploit the insight that firms will be more productive in CSR projects that are technologically close to their for-profit production process. This insight has been repeatedly discussed in the literature under the term 'bundling' (see e.g. Besley and Ghatak, 2007) and is formalized by our micro-foundation for firm×project-specific CSR productivity in section 4.2. We operationalize it by building a measure of technological proximity for each pair of industry i and social topic d, meant to proxy for the model object $\cos(\boldsymbol{\phi_d}, \boldsymbol{\phi_i})$.

To proxy for $\cos(\boldsymbol{\phi_d}, \boldsymbol{\phi_i})$, we make the assumption that if projects in a given CSR topic require a technology that is close to the firm's for-profit technology, this *technological* proximity will be reflected in a *semantic* proximity between descriptions of the CSR topic and descriptions of the firm's production function. For instance, consider whether pharmaceutical firms are more efficient at undertaking CSR projects in health than financial firms. Our premise is that this would be reflected in a higher semantic proximity between the description of a pharmaceutical firm's production function and the description of health CSR projects than the semantic proximity between the description of a financial firm's production function and the description of these projects.

To construct this measure, we exploit embeddings of textual data describing CSR topics and industries, as described in section 2.2. For each industry i, we use the description in the NIC handbook to obtain the embedding $\mathbf{v}_i$. For each topic d, we define $\mathbf{v}_d$ as the average of the project embeddings $\mathbf{v}_p$ for projects p belonging to topic d. We proxy for $\cos(\boldsymbol{\phi_d}, \boldsymbol{\phi_i})$, the technological proximity between industry i and topic d, as:

$$\text{Proximity}_{id} = \cos(\mathbf{v}_d, \mathbf{v}_i) \tag{9}$$

The key assumption underlying our measurement exercise is that the semantic proximity between texts describing industry i and topic d appropriately proxies for the proximity of the technological requirements of industry i and topic d. Formally, we assume $\cos(\mathbf{v}_d, \mathbf{v}_i) \approx \cos(\boldsymbol{\phi_d}, \boldsymbol{\phi_i})$.[15] After describing our measure, we propose

[15] Let $\mathcal{T}$, $\mathcal{W}$ be the dimensions of the task and the embedding space, respectively. A sufficient condition is that $\mathbf{v} = \mathbf{A}\boldsymbol{\phi}$ where $\mathbf{A} \in \mathbb{R}^{\mathcal{W}\times\mathcal{T}}$ is a linear isometry. This requires $\mathcal{T} \leq \mathcal{W}$, and ensures $\cos(\mathbf{v}_d, \mathbf{v}_i) = \cos(\boldsymbol{\phi_d}, \boldsymbol{\phi_i})$. Alternatively, when $\mathcal{T} \geq \mathcal{W}$, $\mathbf{A}$ must be a random projection and we apply the Johnson–Lindenstrauss lemma to obtain $\cos(\mathbf{v}_d, \mathbf{v}_i) \approx \cos(\boldsymbol{\phi_d}, \boldsymbol{\phi_i})$.

three tests to validate this assumption below.

Higher values indicate higher similarity, but the variable has no cardinal interpretation. We standardize it to have a mean of zero and a standard deviation of one to ease the interpretation of our results. In robustness checks, we use alternative measures for $\mathbf{v}_d$ and $\mathbf{v}_i$.

Description of Proximity$_{id}$. Figure 4 depicts the distribution of the proximity variable across topics and industries in a heatmap for all 16 topics and the 16 largest industries in our data. Deeper blue colors indicate higher proximity, and light grey indicates lower proximity. We see that the distribution of the variable is reasonably intuitive: for example, the health topic has a particularly high proximity with the pharmaceutical industry, and the civil engineering industry has a high proximity with the sanitation, safe drinking water, infrastructure, and environmental topics – all topics which require some degree of engineering. We also see that some topics/industries have consistently low or high proximity with many industries/topics (see, for example, the 'safe drinking water' topic). This may reflect true technological patterns, or be due to less desirable characteristics of our textual corpus, such as the recurrence of some non-technical terms in project descriptions. Our regression results below control for topic and firm fixed effects throughout to allow for this possibility, and we consider the robustness of our results to the exclusion of each topic or industry in turn.[16] Figure A.3 plots the distribution of the variable at the firm×topic level, with some examples. One standard deviation in proximity corresponds roughly to the difference in proximity between the topics 'hunger and malnutrition' and 'health' for the pharmaceutical industry.

In Tables C.5–C.8, we investigate which tokens lead a topic×industry pair to have a high or low similarity. The patterns are very intuitive. For instance, for the CSR topic 'infrastructure', the closest industry is 'civil engineering'. Looking at the sets of tokens closest to the centroids of the embeddings vector $\mathbf{v}_d$ and $\mathbf{v}_i$ for this pair, we see that both sets have significant overlap, with tokens related to construction driving their high cosine similarity.

[16] Figure A.2 plots the heatmap of the distribution of the residuals of the proximity variable after removing firm and topic fixed effects.

Validation. We propose three tests of our assumption that $\cos(\mathbf{v}_d, \mathbf{v}_i) \approx \cos(\boldsymbol{\phi}_\mathbf{d}, \boldsymbol{\phi}_\mathbf{i})$. We provide the intuition for these tests below and leave formal derivations to Appendix E.

Test 1. Industry predicts semantic proximity across firms. From our definition of industries, it follows that the technology vectors of firms in the same industry are more similar than those of firms in different industries. If word embeddings appropriately represent technology vectors, then this property should hold for embeddings. We implement this test using the firm-level business descriptions contained in 10-K filings for US firms (there is no equivalent firm-level data for India). Table A.3 shows that firms in the same 2-digit industry have a higher cosine similarity than firms in different industries by 0.118 (compared to an average across-industry value of 0.679). This number rises to 0.146 and 0.155 when we compare firms in the same 3-digit and 4-digit industries, respectively.

Test 2. Semantic proximity across industries predicts firms' sales shares across industries. Our model predicts that a firm in industry i will have higher sales share in products belonging to industry $i' \neq i$ if industry i has high technological proximity with i' (high $\cos(\boldsymbol{\phi}_\mathbf{i}, \boldsymbol{\phi}_\mathbf{i'})$). If the industry embedding vectors $\mathbf{v}_i$ appropriately represent technological requirements, this property should hold for semantic proximity between i and i' (high $\cos(\mathbf{v}_\mathbf{i}, \mathbf{v}_\mathbf{i'})$). Table A.4 shows that this is indeed the case.

Test 3. Semantic proximity across industries predicts input-output proximity. Finally, if industry embedding vectors capture technological requirements we should see that industries with similar embeddings also use similar inputs, sell similar products, or develop supplier-customer relationships. Using the input-output matrix for India, Table A.5 shows that industry pairs (i, i') with similar products or inputs, or strong supplier-customer links, indeed have high textual proximity $\cos(\mathbf{v}_\mathbf{i}, \mathbf{v}_\mathbf{i'})$.

The approach consisting in using textual data and semantic proximity to characterize the proximity of firms in economically-relevant dimensions has been pioneered by Hoberg and Phillips (2016). They interpret semantic similarity as product portfolio similarity, and hence proximity in the competitive space. While test 1 is in line with their interpretation of semantic proximity, tests 2 and 3 suggest that semantic similarity also captures technological proximity well.[17,18]

[17] In our model, firms product portfolio jointly reflects their technology (α_{fp}) and their valuation of products (ζ_{fp}) which could differ across products and firms due to, e.g., imperfect competition.

[18] In particular, test 1 would work if the text describing firms' activity was simply listing the

5.2 Empirical Strategy

We consider whether firms f's technological advantage is correlated with how they allocate CSR expenditures across topics d using the following specification:

$$y_{fd} = \beta \text{Proximity}_{i(f)d} + \gamma_f + \gamma_d + \varepsilon_{fd} \tag{10}$$

where y_{fd} is an increasing function of CSR expenditures at the firm f and topic d level, $\text{Proximity}_{i(f)d}$ is our proxy for technological advantage defined at firm's industry $i(f)$ and topic level, defined above, and standard errors are clustered at the topic×industry level. We include firm fixed effects γ_f to ensure β captures a correlation with the firm's relative technical advantage on a topic, motivated by specification (5), and topic fixed effects γ_d to capture preferences across topics that are common to all firms (we investigate what could be driving such common preferences below).

This specification tests for allocative efficiency, as defined above: finding $\beta > 0$ would indicate that firms spend more on topics in which they have a technological advantage in. Note that our definition of allocative efficiency does not require that firms spend more on topics on which they have a technological advantage *because* they choose to leverage this advantage: allocative efficiency still holds if the allocation is due to firms having, for example, high preferences for topics they have an advantage in (if the α_{fd} and ζ_{fd} terms in the conceptual framework above are positively correlated). Our object of interest is thus the correlation between CSR expenditures and technological advantage.

Our first outcome variable is the share of CSR expenditures a firm spends on the topic. We then consider the extensive margin decision by using an indicator for whether the firm spends any amount on the topic, and finally the intensive margin decision using the share of CSR expenditures spent on the topic, conditional on this share being positive. Our baseline specification gives equal weight to all firms but we also present results obtained by weighing each firm by its total CSR expenditures to consider how technological proximity affects the aggregate CSR allocation.

products produced by firms. Tests 2 and 3 suggest that the embeddings capture information about production processes that allow them to have predictive power for patterns of firm scope across industries and input-output relatedness.

5.3 Results

Table 2 presents the result of estimating specification (10).[19] The correlation between the proxy for technological advantage and how much firms spend on a topic is positive regardless of the outcome variable used. In Panel A, column 1, we see that a one standard deviation in technological proximity between a firm's industry and a topic increases the share that the firm spends on that topic by one percentage point, a 16% increase relative to the mean. A one standard deviation also increases the probability that the firm spends on the topic by two percentage points (9% relative to the mean) and the share spent, conditional on spending a positive amount, by two percentage points (8%). The effects of proximity on CSR expenditure outcomes are larger in Panel B, where we weigh each firm by its total CSR expenditures (the effect on the unconditional spending share is 29%), suggesting larger effects for larger firms.[20]

Figure A.5 presents a series of robustness checks. We first consider results when changing choices made to obtain word embeddings: using OpenAI's NLP model to construct the proximity variable instead of Word2Vec or using text from 10-K filings in the US instead of from the NIC handbook, following Hoberg and Phillips (2016). We then change the method we use to aggregate information from projects at the topic level. Our baseline averages the cosine similarity at the project level across topics. We consider using the median similarity instead, or placing more weights on projects whose descriptions have more informative content (weighing each project by the number of tokens or aggregating at the topic level before computing the cosine similarity). We consider results obtained by grouping together topics that are conceptually similar into 10 topics instead of 16 (see Table C.3 for the new classification), and cluster standard errors at the topic or industry level. Estimates are remarkably similar across these specifications, with the exception of that obtained when grouping together similar topics: it is 60% larger, but less precisely estimated and not statistically different from our baseline estimate. Figure A.6 shows that results are stable when we exclude each topic, or each of the 20 largest industries, in turn.[21] Finally,

[19] Figure A.4 depicts the relationship graphically.

[20] The effect on the spending probability is smaller relative to the mean (11%) when we weigh by total CSR spending because larger firms (that by definition spend more on CSR) spend on a larger number of topics.

[21] The only exception is when we exclude the health topic. The coefficient drops slightly, suggesting technological advantage is particularly important for this topic, but remains statistically significant and indistinguishable from the coefficient obtained using our baseline specification.

Table A.6 presents results obtained by aggregating our data at the industry and topic level, in levels and in logs. Results are similar to those obtained in Panel B of Table 2 regardless of the specification used.

Our results thus suggest that firms allocate more CSR expenditures to topics in which they have a technological advantage. How much of the overall allocation across topics does this explain? Given the similarities between the allocation of funds across topics by firms and the government described above, we compare how much technological proximity and government preferences explain the aggregate allocation. In Table A.7 we estimate a version of equation (10) in which we replace topic fixed effects, that capture common preferences for topics across all firms, with the government's expenditure share on each topic. This is equivalent to assuming that firms' preferences for topics are identical to the government's and using government expenditure shares to proxy for the latter.[22] Interestingly, this specification explains nearly as much of the variation in the data as our baseline specification, suggesting firms' preferences across topics are indeed similar to those of the government. Comparing across coefficients, we find that technological proximity explains roughly one-fifth to one-fourth as much of the variation in our data as the government's spending share. Technological advantage thus plays a meaningful role in explaining the overall allocation, though it is smaller than the role played by common preferences across topics.

5.4 Mechanisms and Heterogeneity

Seen through the lens of our conceptual framework our results suggest the allocation of CSR expenditures across topics is allocatively efficient. But do firms spend more on topics they have a technological advantage in because of this technological advantage or because they have an intrinsic preference for spending on those topics?

We fundamentally cannot disentangle unobserved firm preferences from their technological advantage across topics. That said, we use information on the mode of implementation of projects to provide suggestive evidence. Firms that outsource CSR projects to third-parties likely make less direct use of their own technology than those implementing projects themselves. If preferences are the only determinant of firms' allocation across topics, the decision to outsource should be orthogonal to technolog-

[22] To do this, we re-classify topics so that government and CSR expenditures are comparable, in line with the approach in section 3 above. This is the same classification into 10 topics as that in the robustness checks.

ical advantage across topics. If, however, firms choose at least some projects because they want to leverage their for-profit production technologies, we should see that they are less likely to outsource projects in topics they have an advantage in. In Table A.8, we see that firms are indeed less likely to outsource projects on topics for which they have a technological advantage. These results, while suggestive, are subject to two caveats. First, the test hinges on a narrow view of firms' technology: firms use their technological advantage even when working with third-party implementers if this advantage consists in choosing better projects or implementers. Second, the definition of third-party implementation is imprecise in the reporting requirements, giving rise to measurement concerns for this variable (see Appendix C.1.5).

In Table B.2 we consider whether the correlation with technological proximity is different among firms that spend (substantially) more than the amount prescribed by the law. Unweighted results are extremely similar to those obtained on the main sample but weighted results are slightly smaller. This suggests technological advantage matters less for large firms with a strong preference for CSR, perhaps because they also have strong intrinsic preferences for some topics.

In Figure A.7 we ask whether firms that are particularly beholden to different types of stakeholders behave differently, perhaps in response to stakeholder pressure. We find no evidence that firms with different ownership structures (publicly listed firms or firms with dominant stakeholders), firms in which employees may have more bargaining power (proxied by the average wage, labor share, or training expenses), or firms that rely more on their reputation with final consumers (proxied by advertising expenses or downstreamness) behave differently. The correlation with technological proximity is smaller, but not significantly so, among firms that may need a good relationship with the government (because they operate in heavily regulated industries or compete with government-owned firms), perhaps because of strategic considerations.

6 Is CSR Allocated Equitably?

We now turn to the allocation of CSR across locations to consider its equity characteristics. A natural proxy for the potential social returns to CSR expenditures in an area (μ_p) is its level of economic development; in what follows, we think of an allocation as more equitable if the correlation between area-level expenditure shares and GDP per capita is more negative. As shown above, however, CSR expenditures

are concentrated in a few states, with almost 30% going to just Maharashtra, the richest state (in total GDP) in our data. To consider more generally how equitable the CSR allocation is, we run the following specification at the firm f and state s level:

$$y_{fs} = \beta \text{GDP}_s + \gamma_f + \varepsilon_{fs} \tag{11}$$

where GDP_s is the state's gross product per capita in logs, γ_f are firm fixed effects and we control throughout for state population.

Results are presented in Table 3 (columns 1, 3, and 5). We see that CSR expenditures shares are positively correlated with state GDP per capita: $\beta > 0$. This is also true when considering the extensive and the intensive margin separately. This suggests that there is a wedge between firms' private returns to CSR projects and their social returns: if firms' private returns followed social returns ($\zeta_{fs} = \mu_s$), and unless technological advantage across states α_{fs} is strongly negatively correlated with social returns (a possibility we discuss below), we should observe $\beta < 0$.

Figure 5 plots CSR spending as a function of state GDP (both per capita, blue dots) as well as the linear fit of a regression of CSR spending on GDP per capita using states' population as weight. A 10% increase in state GDP per capita increases CSR spending in that state by 19%, and results are similar if we focus on firms that voluntarily spend more on CSR than the law requires (Figure B.2 and Table B.3).[23] Firms could be allocating their CSR expenditures to the poorest areas in rich states, making the allocation less regressive, so in Figure A.9, we replicate this analysis at the district level. We see a very similar pattern, with more CSR expenditures in richer districts, regardless of the proxy for district development used. In Figure A.10, we consider the correlation between CSR expenditures at the state×topic level and a proxy for state-level needs on that CSR topic. For four topics, we can define a plausible proxy for need (e.g., for education, we consider the state-level literacy rate). This analysis shows that CSR expenditures tend to flow to areas with relatively low need, even when considering topic-specific expenditures.

The allocation of CSR across space is thus a priori inequitable but is it more or less inequitable than alternative uses of CSR funds? One comparison point is the allocation of government expenditures per capita, as the government could have

[23]We also obtain very similar results when we substitute state-level GDP per capita with a multidimensional poverty index (Figure A.8) or consider the sample of all firms in the CSR data (Figure B.2 and Table B.3).

increased taxes instead of imposing a CSR mandate. In Figure 5, we see that state-level government expenditures per capita (restricted to the topics covered by the CSR data, green dots) increase slightly with state development, but the slope is only one-fifth of the slope for CSR expenditures.[24] The allocation of CSR across space is thus more inequitable than the allocation of government expenditures.[25]

Why are CSR expenditures concentrated in rich states? One reason could be that firms concentrate their spending where they are headquartered, and corporate headquarters are concentrated in rich states. To test this, columns 2, 4, and 6 in Table 3 control for an indicator for whether the firm is headquartered in the state.[26] The coefficients for the indicator are very large, reflecting the fact that around 60% of CSR spending occurs in firms' headquarter states. The coefficient for state GDP per capita indeed falls and becomes null for the intensive margin.

This concentration of spending where firms are headquartered could be driven by firms' preferences for their local area. It may also reflect efficiency considerations if it is due to firms having particularly good information on the needs of their local areas or the technology required to meet them. Note however that when we exclude expenditures in headquarter states in Figure 5, the aggregate slope falls but remains statistically significant and more than twice as large as that for government expenditures. This suggests efficiency considerations linked to firms' locations alone are unlikely to explain why firms spend more in richer states.

7 Conclusion

In this paper, we use a novel dataset on the quasi-universe of the CSR expenditures of Indian firms to shed light on the potential welfare effects of CSR. We reach two main conclusions. First, we provide evidence consistent with the idea that firms spend more on CSR projects they have a technological advantage in, i.e., projects they may be particularly good at providing because of the technology they use in their for-

[24] The small positive slope reflects the fact that richer states both get less inter-governmental transfers from the central government and collect more tax revenues.

[25] Another potential benchmark is that CSR expenditures could have been redistributed to shareholders instead. We cannot locate shareholders, but they are likely much richer than the average Indian citizen and located in richer states. A counterfactual allocation of CSR expenditures to profits could thus have led to an even more inequitable allocation of these funds across locations.

[26] We use information on where the firm is registered in the accounting data to proxy for headquarter state.

profit production processes. We do so by constructing a proxy for the technological proximity between firms' industries and CSR topics (e.g., health, education), using the textual proximity between the descriptions of industries and topics. Seen through the lens of the theoretical literature on CSR, this suggests CSR can efficiently contribute to public good provision. Second, we find that firms spend substantially more on CSR in richer locations, in part because they spend more in states where their headquarters are located. In summary, our results suggest that mandating CSR may be an efficient way to increase expenditures on public good provision, but it will come at an equity cost.

References

Acemoglu, Daron and Pascual Restrepo, "The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment," *American Economic Review*, June 2018, *108* (6), 1488–1542.

Allcott, Hunt, Giovanni Montanari, Bora Ozaltun, and Brandon Tan, "An Economic View of Corporate Social Impact," NBER Working Papers 31803, National Bureau of Economic Research, Inc October 2023.

Antràs, Pol, Davin Chor, Thibault Fally, and Russell Hillberry, "Measuring the upstreamness of production and trade flows," *American Economic Review*, 2012, *102* (3), 412–416.

Ash, Elliott and Stephen Hansen, "Text Algorithms in Economics," *Annual Review of Economics*, 2023, *15* (Volume 15, 2023), 659–688.

Asher, Sam, Tobias Lunt, Ryu Matsuura, and Paul Novosad, "Development research at high geographic resolution: an analysis of night-lights, firms, and poverty in India using the shrug open data platform," *The World Bank Economic Review*, 2021, *35* (4).

Awasthi, Kshitij, Sai Yayavaram, Rejie George, and Trilochan Sastry, "Classification for regulated industries: A new index," *IIMB Management Review*, 2019, *31* (3), 309–315.

Baron, David P., "Private Politics, Corporate Social Responsibility, and Integrated Strategy," *Journal of Economics & Management Strategy*, 2001, *10* (1), 7–45.

Behaghel, Luc, Bruno Crépon, and Marc Gurgand, "Private and Public Provision of Counseling to Job Seekers: Evidence from a Large Controlled Experiment," *American Economic Journal: Applied Economics*, October 2014, *6* (4), 142–74.

Benabou, Roland and Jean Tirole, "Individual and Corporate Social Responsibility," *Economica*, 2010, *77* (305), 1–19.

Bergeron, Augustin, Gabriel Tourek, and Jonathan L Weigel, "The state capacity ceiling on tax rates: Evidence from randomized tax abatements in the drc," *Econometrica*, 2024, *92* (4), 1163–1193.

Bertrand, Marianne, Matilde Bombardini, Raymond Fisman, and Francesco Trebbi, "Tax-Exempt Lobbying: Corporate Philanthropy as a Tool for Political Influence," *American Economic Review*, July 2020, *110* (7), 2065–2102.

Besley, Timothy and Maitreesh Ghatak, "Retailing public goods: The economics of corporate social responsibility," *Journal of public Economics*, 2007, *91* (9), 1645–1663.

_ **and Torsten Persson**, "The Origins of State Capacity: Property Rights, Taxation, and Politics," *The American Economic Review*, 2009, *Vol. 99(4)*, 1218–1244.

Best, Michael, Anne Brockmeyer, Henrik Jacobsen Kleven, Johannes Spinnewijn, and Mazhar Waseem, "Production vs Revenue Efficiency With Limited Tax Capacity: Theory and Evidence From Pakistan," *Journal of Political Economy*, 2015, *123* (6).

Bhattacharyya, Asit and Md Lutfur Rahman, "Mandatory CSR expenditure and firm performance," *Journal of Contemporary Accounting & Economics*, 2019, *15* (3), 100163.

Boehm, Johannes, Swati Dhingra, and John Morrow, "The comparative advantage of firms," *Journal of Political Economy*, 2022, *130* (12), 3025–3100.

Brandon, Rajna Gibson, Simon Glossner, Philipp Krueger, Pedro Matos, and Tom Steffen, "Do Responsible Investors Invest Responsibly?*," *Review of Finance*, 09 2022, *26* (6), 1389–1432.

Broccardo, Eleonora, Oliver Hart, and Luigi Zingales, "Exit versus Voice," *Journal of Political Economy*, 2022, *130* (12), 3101–3145.

Card, David, Kevin F. Hallock, and Enrico Moretti, "The geography of giving: The effect of corporate headquarters on local charities," *Journal of Public Economics*, 2010, *94* (3), 222–234.

Cheng, Ing-Haw, Harrison Hong, and Kelly Shue, "Do managers do good with other people's money?," *The Review of Corporate Finance Studies*, 2023, *12* (3), 443–487.

Chhaochharia, Vidhi, Rik Sen, and Jing Xu, "Corporate Benevolence and Societal Impact: Evidence from India's CSR Reform," 2025. Mimeo, University of Georgia.

Christensen, H, Emmanuel T De George, Anthony Joffre, and Daniele Macciocchi, "Consumer responses to the revelation of corporate social irresponsibility," 2023.

Christensen, Hans B, Luzi Hail, and Christian Leuz, "Mandatory CSR and sustainability reporting: Economic analysis and literature review," *Review of accounting studies*, 2021, *26* (3), 1176–1248.

Colonnelli, Emanuele, Timothy McQuade, Gabriel Ramos, Thomas Rauter, and Olivia Xiong, "Polarizing Corporations: Does Talent Flow to Good Firms?," Working Paper 31913, National Bureau of Economic Research November 2023.

Conway, Jacob and Levi Boxell, "Consuming values," *Available at SSRN 4855718*, 2024.

Das, Satadru, Lucie Gadenne, Tushar Nandi, and Ross Warwick, "Does going cashless make you tax-rich? Evidence from India's demonetization experiment," *Journal of Public Economics*, 2023, *224*, 104907.

Dharmapala, Dhammika and Vikramaditya Khanna, "The impact of mandated corporate social responsibility: Evidence from India's Companies Act of 2013," *International Review of law and Economics*, 2018, *56*, 92–104.

Ferreira, Daniel and Radoslawa Nikolowa, "Polarization, purpose and profit," *Journal of Financial Economics*, 2025, *172*, 104147.

Fioretti, Michele, "Caring or Pretending to Care? Social Impact, Firms' Objectives, and Welfare," *Journal of Political Economy*, 2022, *130* (11), 2898–2942.

_ , Victor Saint-Jean, and Simon C. Smith, "The Shared Cost of Pursuing Shareholder Value," 2023.

Flammer, Caroline and Jiao Luo, "Corporate social responsibility as an employee governance tool: Evidence from a quasi-experiment," *Strategic Management Journal*, 2017, *38* (2), 163–183.

Friedman, Milton, "The social responsibility of business is to increase its profits," 1970.

Gadenne, Lucie, "Tax Me, but Spend Wisely? Sources of Public Finance and Government Accountability," *American Economic Journal: Applied Economics*, January 2017, *9* (1), 274–314.

Gatignon, Aline and Christiane Bode, "When few give to many and many give to few: Corporate social responsibility strategies under India's legal mandate," *Strategic Management Journal*, 2023, *44* (9), 2099–2127.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, "Text as Data," *Journal of Economic Literature*, September 2019, *57* (3), 535–74.

Gillan, Stuart L, Andrew Koch, and Laura T Starks, "Firms and social responsibility: A review of ESG and CSR research in corporate finance," *Journal of Corporate Finance*, 2021, *66*, 101889.

Government of India, "State-Wise Number and Percentage of Population Below Poverty Line in India: 2011-12," 2014. Planning Commission, Government of India, https://www.data.gov.in/resource/state-wise-number-and-percentage-population-below-poverty-line-india-2011-12-based.

_ , "State/UT-wise Prevalence of Underweight, Stunting, Wasting and Anaemia Among Children Under 5 Years of Age as per National Family Health Survey-4, 2015-16," 2020. Ministry of Women and Child Development, Government of India, https://ap.data.gov.in/resource/stateut-wise-prevalence-underweight-stunting-wasting-and-anaemia-among-children-under-5.

_ , "State-Wise Infant Mortality Rate, 2013," 2021. Office of the Registrar General and Census Commissioner, Ministry of Home Affairs, Government of India, https://www.rbi.org.in/scripts/PublicationsView.aspx?id=20670.

_ , "State-Wise Literacy Rate, 2011," 2021. Office of the Registrar General and Census Commissioner, Ministry of Home Affairs, Government of India, https://www.rbi.org.in/scripts/PublicationsView.aspx?id=20665.

_ , "State/UT-wise Details of Headcount Ratio, Intensity and Multi-Dimensional Poverty Index (MPI), 2015-16," 2024. https://www.data.gov.in/resource/stateut-wise-details-headcount-ratio-intensity-and-multi-dimensional-poverty-index-mpi.

Green, Daniel and Boris Vallee, "Measurement and effects of bank exit policies," 2024.

Hart, Oliver and Luigi Zingales, "Companies should maximize shareholder welfare not market value," *ECGI-Finance Working Paper*, 2017, (521).

_ , Andrei Shleifer, and Robert W. Vishny, "The Proper Scope of Government: Theory and an Application to Prisons*," *The Quarterly Journal of Economics*, 11 1997, *112* (4), 1127–1161.

Hart, Oliver D. and Luigi Zingales, "The New Corporate Governance," NBER Working Papers 29975, National Bureau of Economic Research, Inc April 2022.

Hartzmark, Samuel and Kelly Shue, "Counterproductive Impact Investing: The Impact Elasticity of Brown and Green Firms," *SSRN Electronic Journal*, 01 2023.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil, "A Bright Idea for Mesuring Economic Growth," *American Economic Review*, 2011.

Hoberg, Gerard and Gordon Phillips, "Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis," *The Review of Financial Studies*, 08 2010, *23* (10), 3773–3811.

_ **and** _ , "Text-based network industries and endogenous product differentiation," *Journal of political economy*, 2016, *124* (5), 1423–1465.

Hong, Harrison and Edward Shore, "Corporate Social Responsibility," *Annual Review of Financial Economics*, 11 2023, *15*, 327–350.

ICRISAT, "ICRISAT-District Level Data," 2020. International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), https://data.icrisat.org/dld/ .

Jensen, Anders, "Employment Structure and the Rise of the Modern Tax System," *American Economic Review*, January 2022, *112* (1), 213–234.

Kahn, Matthew E, John Matsusaka, and Chong Shu, "Divestment and Engagement: The Effect of Green Investors on Corporate Carbon Emissions," Working Paper 31791, National Bureau of Economic Research October 2023.

Kitzmueller, Markus and Jay Shimshack, "Economic Perspectives on Corporate Social Responsibility," *Journal of Economic Literature*, March 2012, *50* (1), 51–84.

Knutsson, Daniel and Björn Tyrefors, "The Quality and Efficiency of Public and Private Firms: Evidence from Ambulance Services*," *The Quarterly Journal of Economics*, 02 2022, *137* (4), 2213–2262.

Koh, Paul and Devesh Raval, "Economies of Scope from Shared Inputs," 2025.

Kotchen, Matthew J., "Green Markets and Private Provision of Public Goods.," *Journal of Political Economy*, 2006, *114* (4), 816 – 834.

Lin, Li-Wei, "Mandatory Corporate Social Responsibility Legislation around the World: Emergent Varieties and National Experiences," *University of Pennsylvania Journal of Business Law*, 2021.

Lindenlaub, Ilse, "Sorting Multidimensional Types: Theory and Application," *The Review of Economic Studies*, 2017, *84* (2 (299)), 718–789.

List, John A., "The Market for Charitable Giving," *Journal of Economic Perspectives*, June 2011, *25* (2), 157–80.

Lloyd, Stuart, "Least squares quantization in PCM," *IEEE transactions on information theory*, 1982, *28* (2), 129–137.

Lund, Dorothy S, "Toward a Dynamic View of Corporate Purpose," *European Corporate Governance Institute-Law Working Paper*, 2023, (746).

MacQueen, James, "Some methods for classification and analysis of multivariate observations," in "Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics," Vol. 5 University of California press 1967, pp. 281–298.

Magill, Michael, Martine Quinzii, and Jean-Charles Rochet, "A Theory of the Stakeholder Corporation," *Econometrica*, September 2015, *83* (5), 1685–1725.

Manchiraju, Hariom and Shivaram Rajgopal, "Does corporate social responsibility (CSR) create shareholder value? Evidence from the Indian Companies Act 2013," *Journal of Accounting Research*, 2017, *55* (5), 1257–1300.

Margolis, Joshua, Hillary Elfenbein, and James Walsh, "Does it pay to be good? A meta-analysis and redirection of research on the relationship between corporate social and financial performance," 01 2007.

Mikolov, Tomas, "Google code archive - Word2Vec," Jul 2013.

Ministry of Corporate Affairs, "Compendium on Corporate Social Responsibility in India," Technical Report, Ministry of Corporate Affairs, Government of India 2021.

Morgan, John and Justin Tumlinson, "Corporate provision of public goods," *Management Science*, 2019, *65* (10), 4489–4504.

Mosaik, "Global High Resolution Estimates of the United Nations Human Development Index," 2023. https://www.mosaiks.org/data-sets .

Mukherjee, Abhishek, Ron Bird, and Geeta Duppati, "Mandatory Corporate Social Responsibility: The Indian experience," *Journal of Contemporary Accounting & Economics*, 2018, *14* (3), 254–265.

Mukherjee, Anita, "Impacts of Private Prison Contracting on Inmate Time Served and Recidivism," *American Economic Journal: Economic Policy*, May 2021, *13* (2), 408–38.

Muralidharan, Karthik, *Accelerating India's Development: A State-Led Roadmap for Effective Governance*, Viking, 2024.

Rajgopal, Shivaram and Prasanna Tantri, "Does a government mandate crowd out voluntary corporate social responsibility? Evidence from India," *Journal of Accounting Research*, 2023, *61* (1), 415–447.

Rios, Anthony and Brandon Lwowski, "An empirical study of the downstream reliability of pre-trained word embeddings," in "Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)" 2020.

Rousseeuw, Peter J, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, 1987, *20*, 53–65.

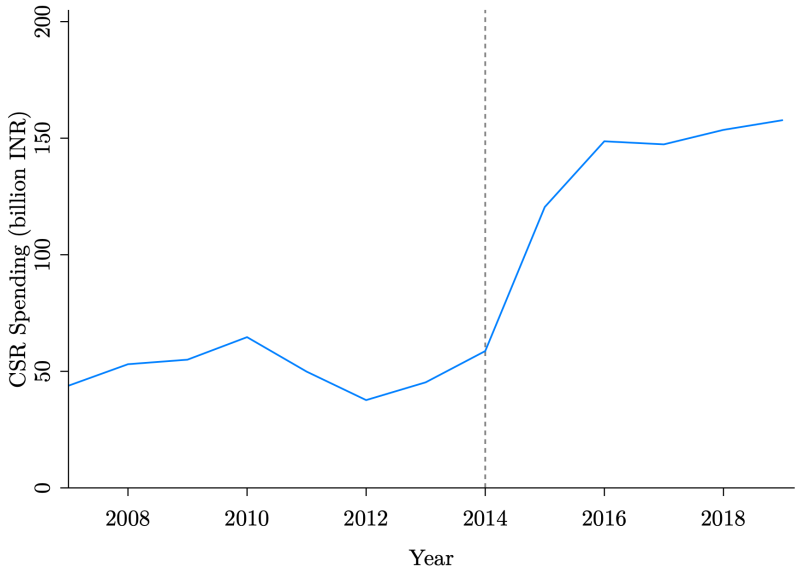
SHRUG, "The Socioeconomic High-resolution Rural-Urban Geographic Platform for India (SHRUG), Version 2.1," 2024. https://www.devdatalab.org/shrug.

Starks, Laura, "Presidential Address: Sustainable Finance and ESG Issues—Value versus Values," *The Journal of Finance*, 2023, *78* (4), 1837–1872.

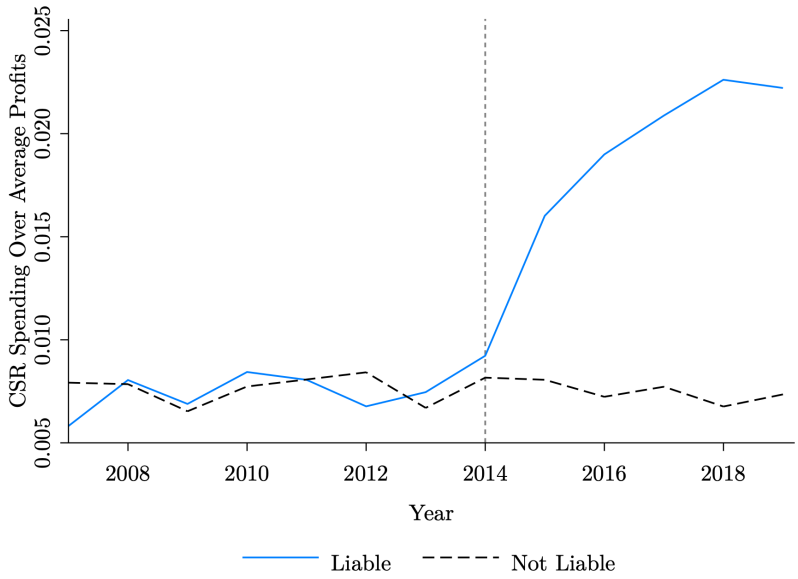
The Giving Institute, *Giving USA: The Annual Report on Philanthropy in America 2023*, Chicago, IL: Giving USA Foundation, 2023. Comprehensive report on charitable giving trends and data in the United States.

Figure 1: CSR Spending Over Time

(a) Aggregate CSR Spending

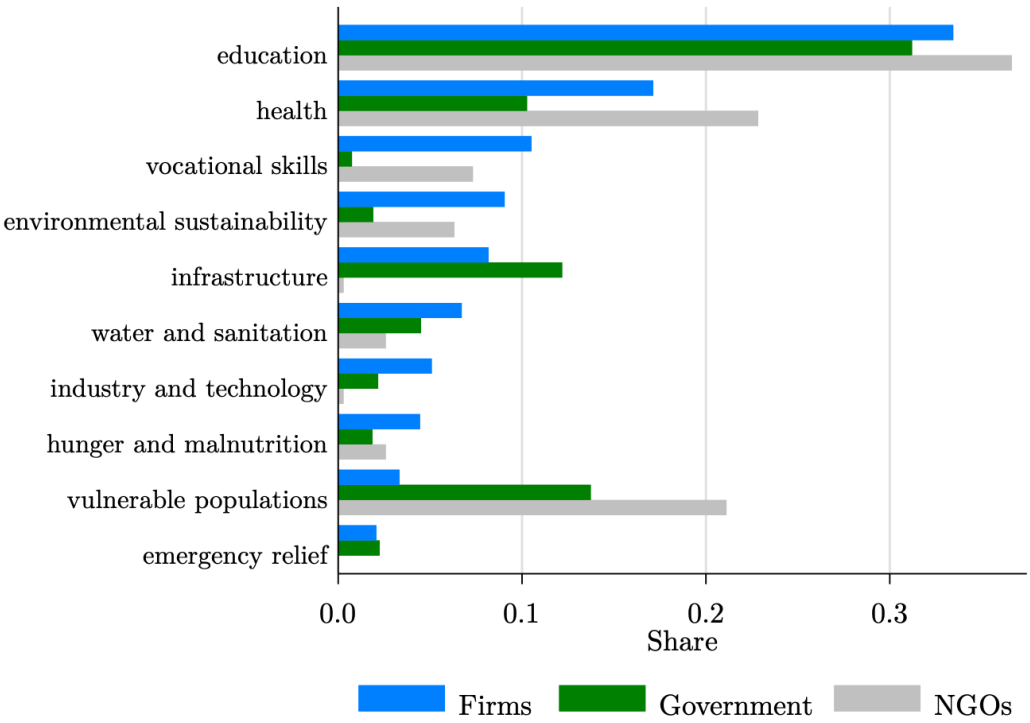


(b) CSR Spending by Firms' Liability Status



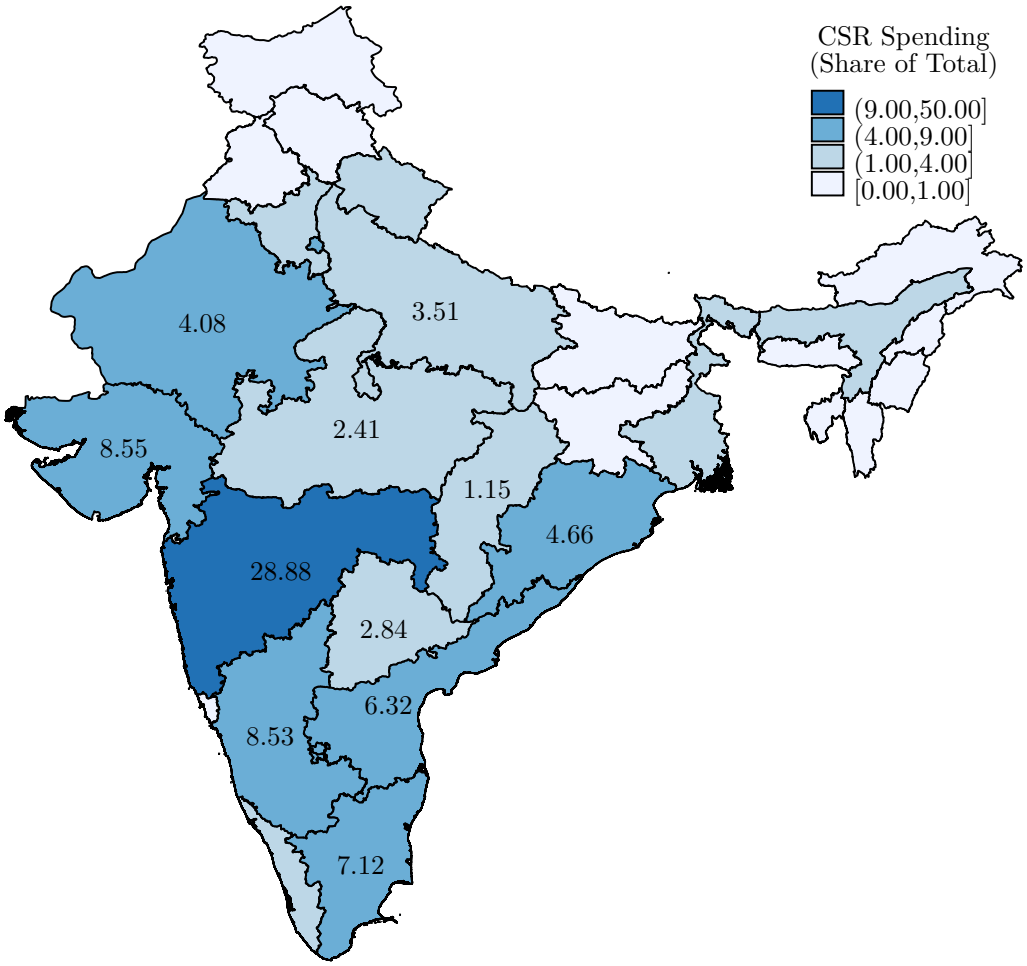
Notes: This figure depicts CSR spending over time. In Figure 1(a), CSR spending is aggregated over all firms and denominated in 2015 billion INR. Figure 1(b) depicts the CSR spending of a given firm in a given year over average profits in the past three years. The blue line (solid) depicts the mean over firms that are liable under the policy and the black line (dashed) depicts the mean over firms that are not liable.

Figure 2: Allocation Across Topics by Public Goods Providers



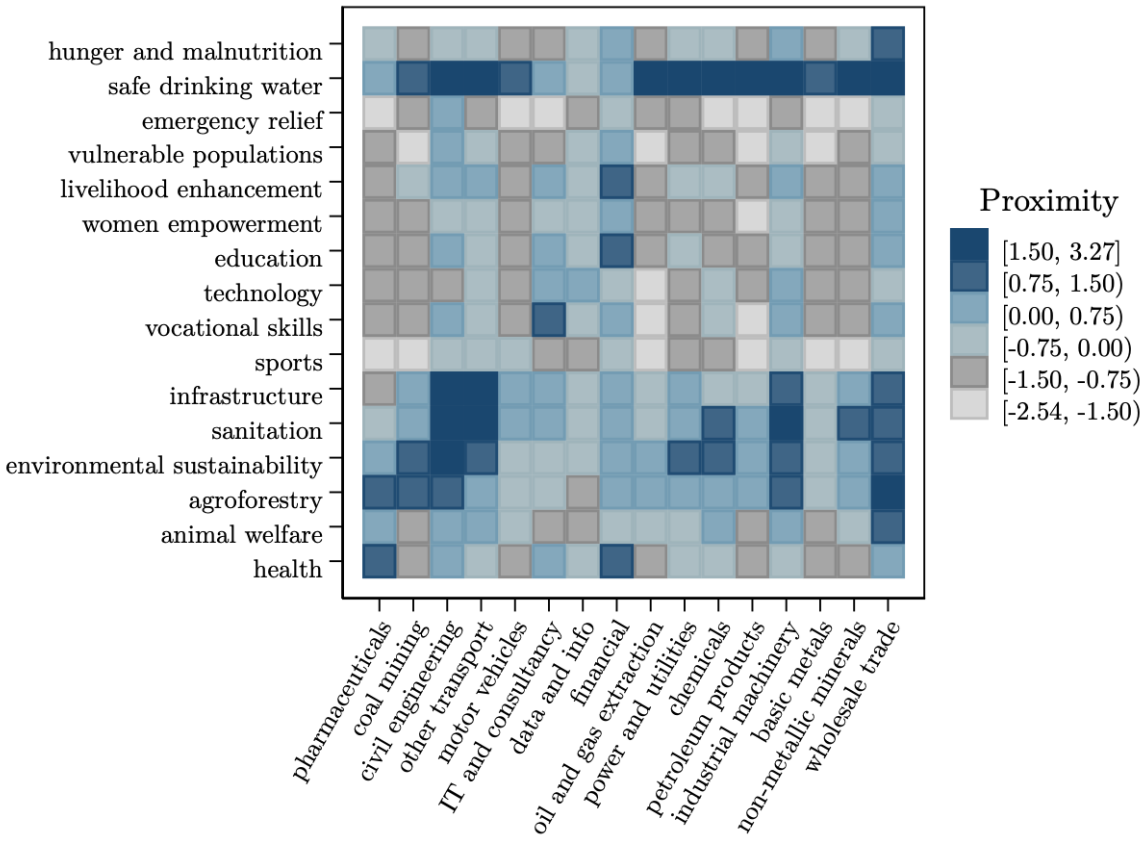
Notes: This figure depicts the share of each topic in terms of total CSR spending, total government spending, and number of NGOs. See Appendix C.2 for the mapping of CSR topics to government spending and the number of NGOs. NGO data does not include emergency relief.

Figure 3: Allocation of CSR Spending Across States



Notes: This figure depicts CSR spending shares by state.

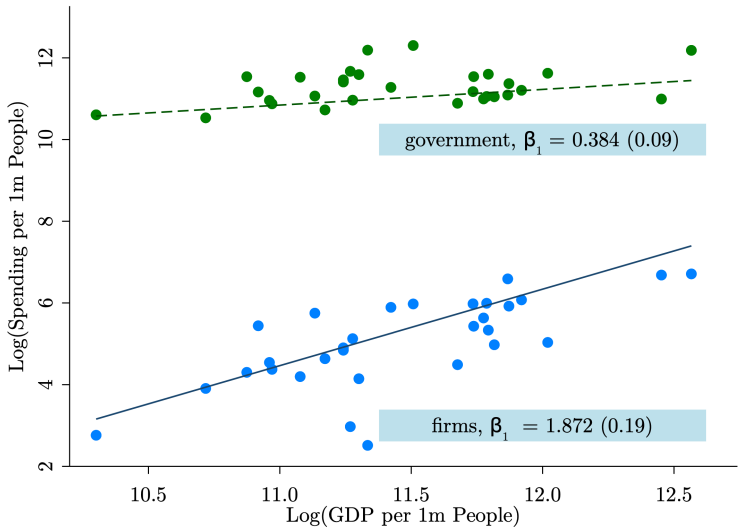
Figure 4: Proximity Across Topics and Largest Industries



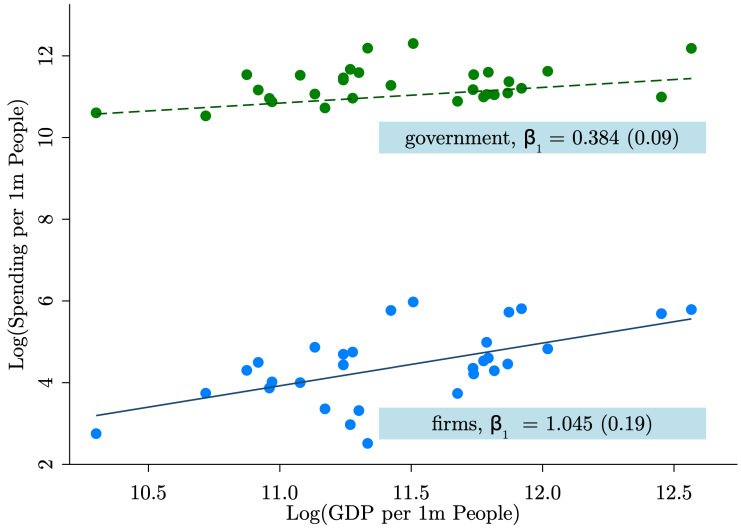
Notes: This figure depicts all 16 topics and the largest 16 industries by total CSR spending. The unit of observation is at the industry×topic level. $\text{Proximity}_{i(f)d}$ is the textual measure of closeness between an industry and a topic defined in section 5.1.

Figure 5: CSR Spending, Government Spending, and State-Level GDP

(a) CSR Spending



(b) CSR Spending Outside Headquarter States



Notes: This figure depicts the relationship between state-level GDP as well as firm (CSR) and government spending. The scattered dots indicate state-level observations, blue for firm spending and green for government spending. The lines indicate fitted linear approximations, blue for firm spending (solid) and green for government spending (dashed). The unit of observation is at the state level. The dependent variables are the log of spending (in millions, denominated in 2015 INR) per one million people by firms (CSR) and the government, aggregated from 2015 to 2019. The independent variable is the log of state-level GDP (in millions, denominated in 2015 INR) per one million people in 2013. Observations are weighted by the 2011 population. In Figure 5(a), spending in the headquarter state of the firms is included. In Figure 5(b), this spending is excluded.

Table 1: Spending Share and Most Frequent Project Types by Topic

Topic	Share	Three Most Frequent Project Types
Education	32%	Special education, Scholarships & support for meritorious students, School construction & infrastructure
Health	17%	Preventive health, Medical treatment and patient care, Medical infrastructure and equipment
Infrastructure	8%	Construction (roads, walls, street light), Construction (community centers), Rural infrastructure
Environmental sustainability	8%	Ecology/conservation projects, Tree plantation, Green energy
Vocational skills	6%	Vocational training, Skill acquisition, Staff training
Technology	5%	Education technology, Mobile science labs & support to labs, Computers & other equipment donations
Livelihood enhancement	5%	Sustainable livelihood and education, Disability inclusion, Economic development
Sanitation	5%	Sanitation infrastructure, Toilet construction, Cleanliness campaigns
Hunger and malnutrition	4%	Eradicating hunger, Midday meal scheme, Food distribution
Safe drinking water	2%	Water supply infrastructure (tanks, pumps, wells...), Water safety programs, Water purification
Vulnerable populations	2%	Hostels for old age, widows & orphans, Social welfare programs, Veterans support
Emergency relief	2%	Flood relief, Disaster relief, Contribution to Prime Minister Relief Fund
Sports	2%	Competition organization, Support to olympic/nationally-recognized sports, Support to sport clubs
Women empowerment	1%	Education and young women empowerment, Gender equality, Disadvantaged women and girls
Agroforestry	1%	Sustainable agriculture, Farmer training, Agronomy
Animal welfare	0%	Animal care, Animal shelters, Cow sheds

Notes: This table displays the share of total CSR spending and the most frequent project types by topic. The project types are the three largest clusters obtained by k-means clustering within each topic, as described in Appendix C.3.

Table 2: CSR Spending and Proximity

	CSR Share Unconditional$_{fd}$ (1)	Any CSR Spending$_{fd}$ (2)	CSR Share Conditional$_{fd}$ (3)
	Panel A: Not Weighted		
Proximity$_{i(f)d}$	0.010***	0.021***	0.021***
	(0.003)	(0.003)	(0.006)
Avg dep var	0.062	0.223	0.280
Firm FE	✓	✓	✓
Topic FE	✓	✓	✓
R-squared	0.24	0.33	0.36
Observations	105,168	105,168	21,684
	Panel B: Weighted by Total CSR Spending		
Proximity$_{i(f)d}$	0.018***	0.046***	0.025***
	(0.005)	(0.008)	(0.009)
Avg dep var	0.062	0.415	0.151
Firm FE	✓	✓	✓
Topic FE	✓	✓	✓
R-squared	0.27	0.37	0.33
Observations	105,168	105,168	21,684

Notes: This table describes the relationship between CSR spending and proximity, derived from equation (10). The unit of observation is at the firm×topic level. The dependent variables are the share of CSR spending of a firm (f) over topics (d), an indicator for any CSR spending by firm (f) in a given topic (d), and the share of CSR spending conditional on any spending. Proximity$_{i(f)d}$ is the textual measure of closeness between an industry and a topic defined in section 5.1. In Panel A, observations are unweighted. In Panel B, observations are weighted by the total CSR spending of each firm, winsorized at the 1st and 99th percentile. Standard errors are clustered at the industry×topic level. $***$, $**$ and $*$ indicate significance at the 1%, 5% and 10% levels.

Table 3: CSR Spending, State-Level Characteristics, and Firm Headquarters

	CSR Share Unconditional$_{fs}$		Any CSR Spending$_{fs}$		CSR Share Conditional$_{fs}$	
	(1)	(2)	(3)	(4)	(4)	(6)
Log(GDP per 1m People)$_s$	0.096***	0.014***	0.156***	0.058***	0.145***	-0.001
	(0.031)	(0.004)	(0.040)	(0.008)	(0.022)	(0.011)
1(Firm Headquarter State)$_{fs}$		0.601***		0.722***		0.333***
		(0.026)		(0.022)		(0.023)
Avg dep var	0.029	0.029	0.067	0.067	0.437	0.437
Firm FE	✓	✓	✓	✓	✓	✓
R-squared	0.08	0.52	0.18	0.44	0.42	0.57
Observations	196,415	196,415	196,415	196,415	9,736	9,736

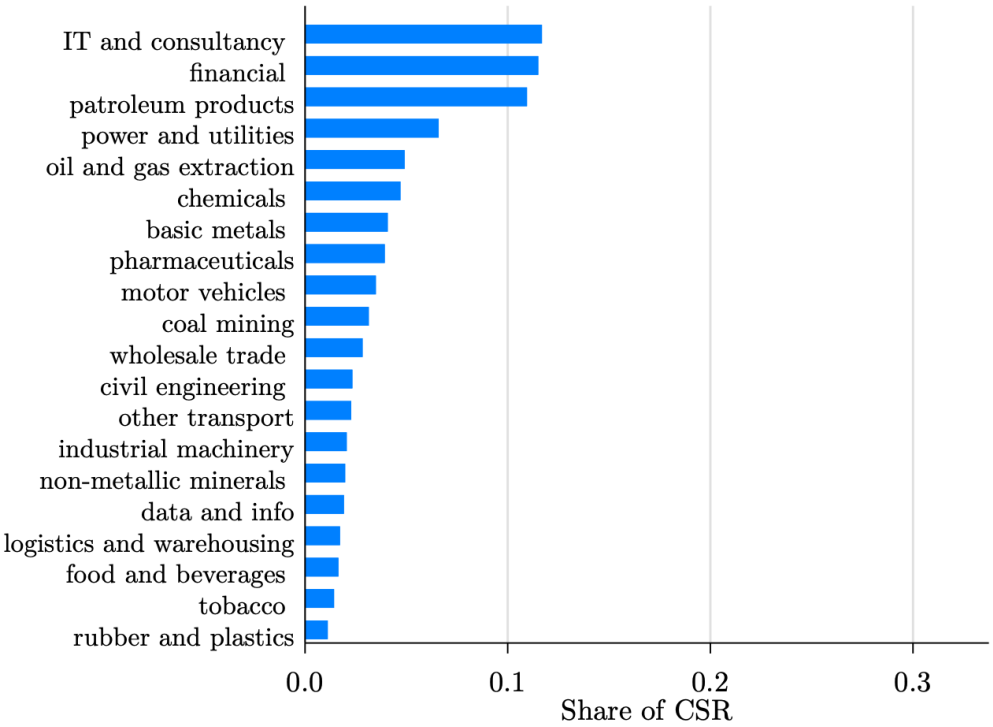
Notes: This table describes the relationship between state-level characteristics as well as firm headquarters and CSR spending, derived from equation (11). The unit of observation is at the firm×state level. The dependent variables are the share of CSR spending of a firm (f) over topics (d), an indicator for any CSR spending by firm (f) in a given topic (d), and the share of CSR spending conditional on any spending. The independent variables are the log of state-level GDP per 1 million people and an indicator that equals one if the firm is headquartered in the state as per government records. We control for the log of population in millions. Observations are weighted by the 2011 population. Standard errors are clustered at the state-level. $***$, $**$ and $*$ indicate significance at the 1%, 5% and 10% levels.

Appendices

(for Online Publication Only)

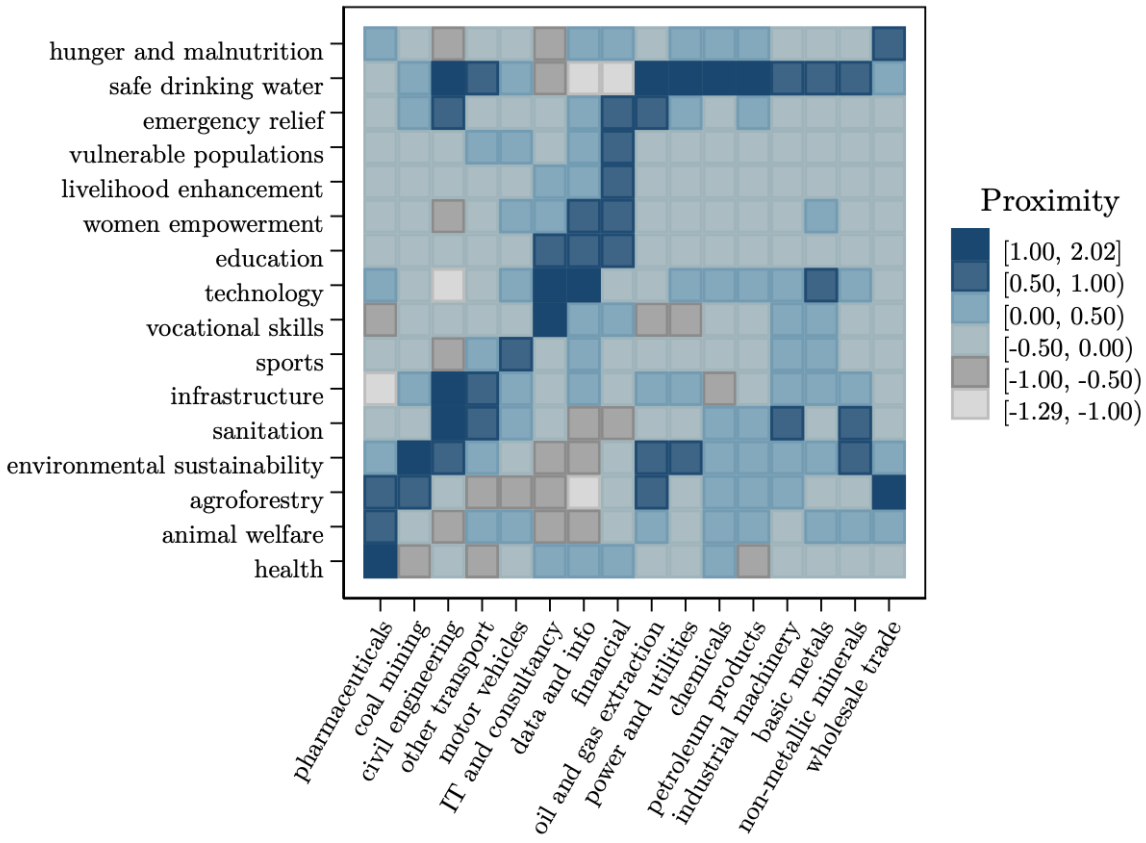
A Additional Tables and Figures

Figure A.1: CSR Spending Share by Industry



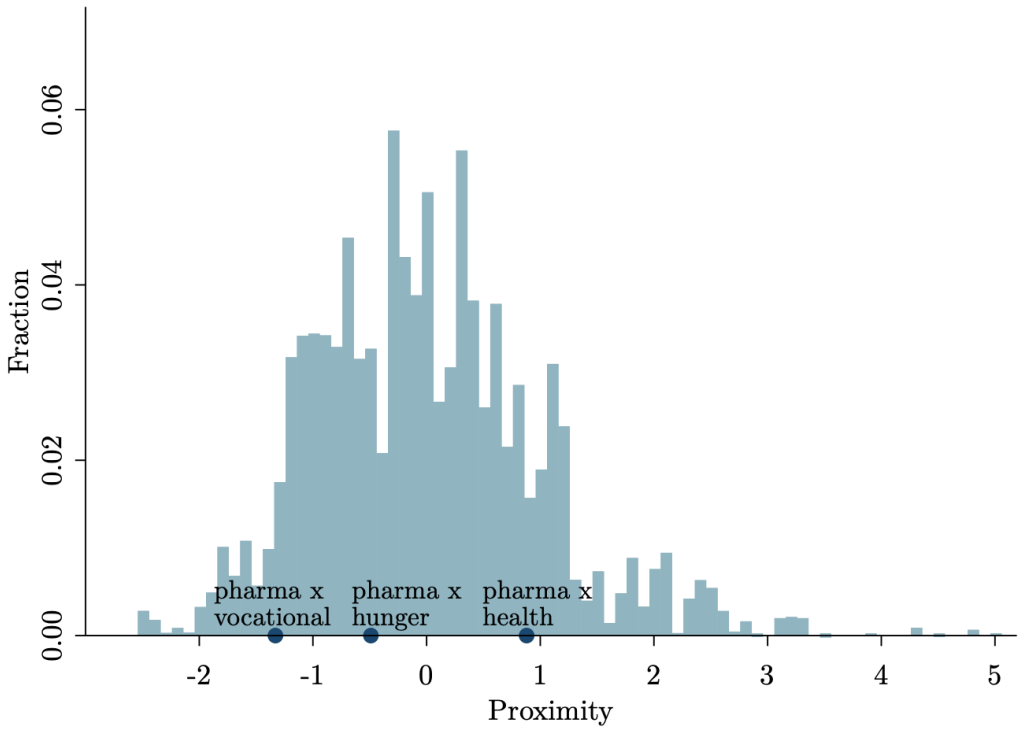
Notes: This figure depicts the share of each industry in total CSR expenditures for the 20 largest industries in the data.

Figure A.2: Proximity Across Topics and Largest Industries With Fixed Effects



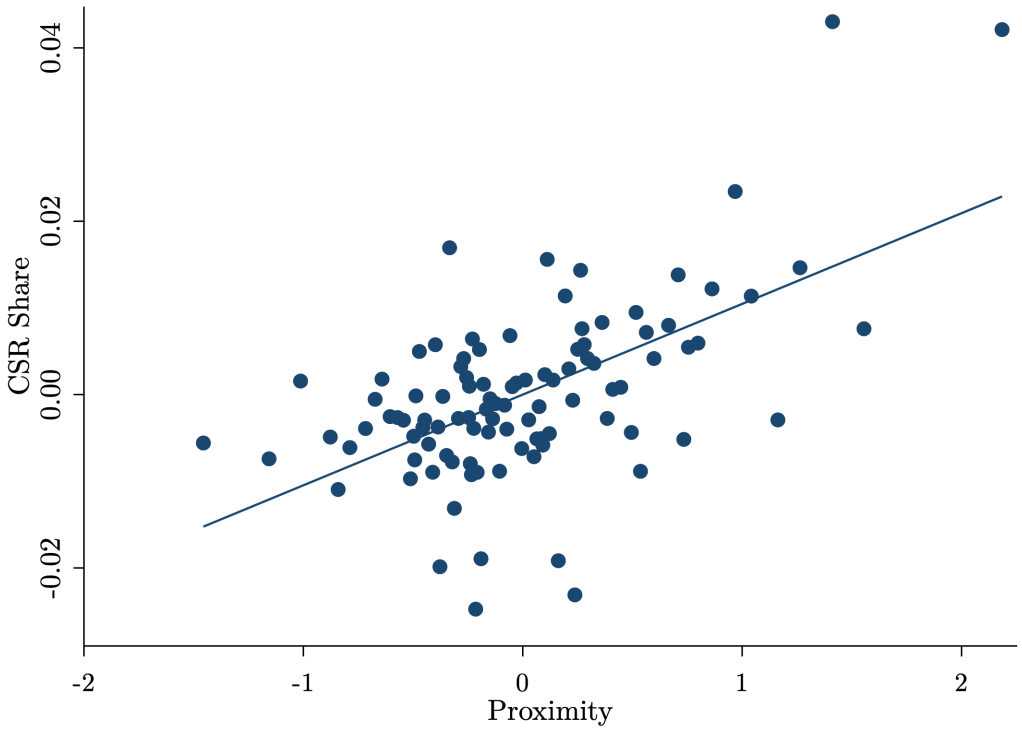
Notes: This figure depicts the proximity variable for all 16 topics and the largest 16 industries by total CSR spending. The unit of observation is at the industry×topic level. $\text{Proximity}_{i(f)d}$ is the textual measure of closeness between an industry and a topic defined in section 5.1. The proximity measure is residualised on firm and topic fixed effects.

Figure A.3: Distribution of Proximity



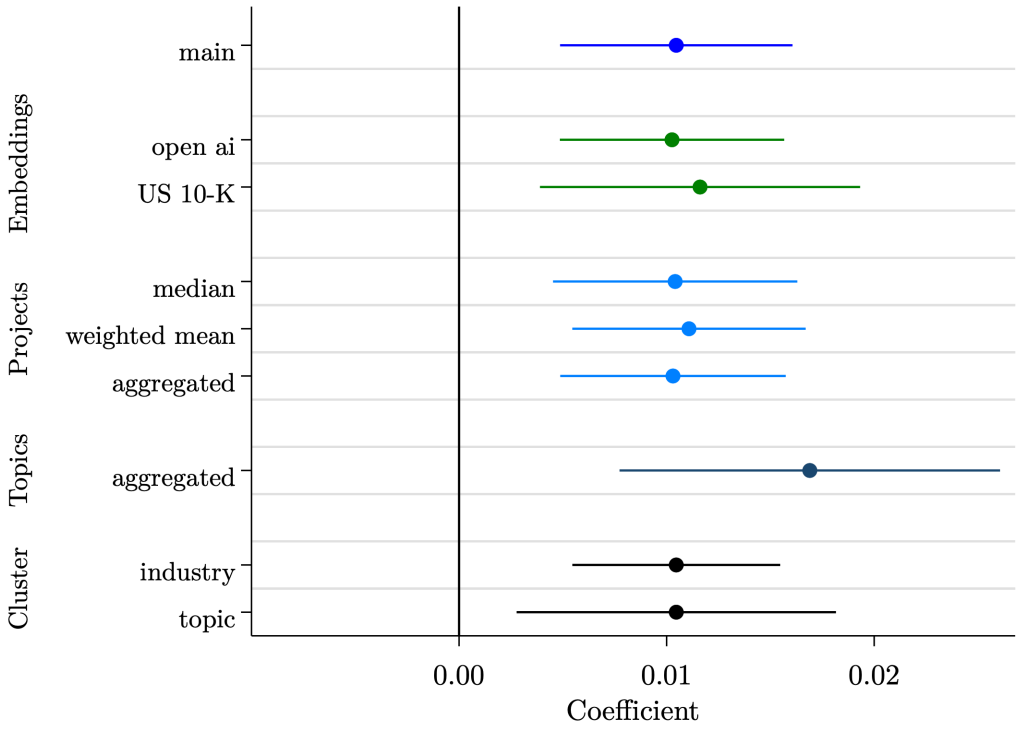
Notes: This figure depicts the distribution of the proximity variable. Proximity$_{i(f)d}$ is the textual measure of closeness between an industry and a topic defined in section 5.1. The unit of observation is at the industry×topic level. The figure shows three examples, corresponding to approximately one standard deviation below the mean (pharmaceuticals × vocational skills), the mean (pharmaceuticals × hunger and malnutrition), and one standard deviation above the mean (pharmaceuticals × health).

Figure A.4: CSR Spending and Proximity



Notes: This figure describes the relationship between CSR spending share and proximity. The unit of observation is at the firm×topic level, and the 105,168 observations are binned into 100 equal-sized bins. The variable on the y-axis is the unconditional CSR spending share for a firm (f) over topics (d). Proximity$_{i(f)d}$ is the textual measure of closeness between an industry and a topic defined in section 5.1. The variables are residualised on firm and topic fixed effects.

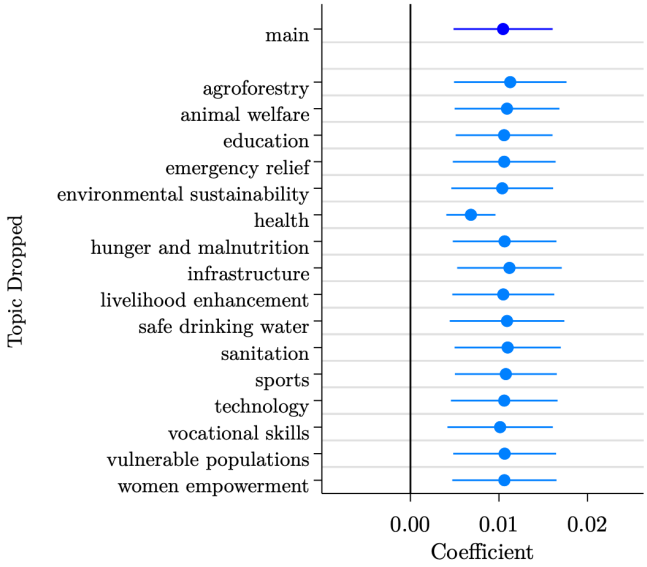
Figure A.5: Effect of Proximity on CSR Spending, Robustness I



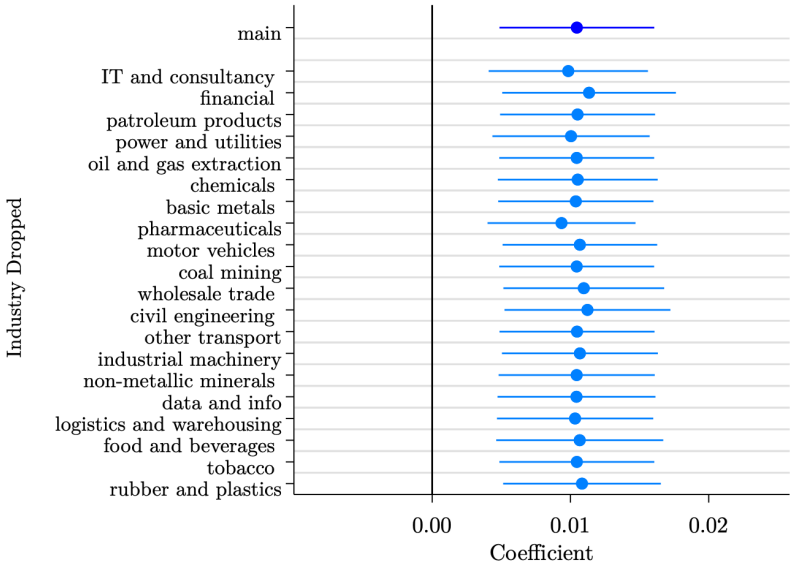
Notes: This figure describes the robustness of the relationship between CSR spending and proximity, derived from equation (10). The unit of observation is at the firm×topic level. The dependent variable is the share of CSR spending that firm (f) spends on topic (d). Proximity$_{i(f)d}$ is the textual measure of closeness between an industry and a topic defined in section 5.1. Row 1 describes the main specification from Table 2, Panel A, column 1. Rows 2 and 3 show results for different word embeddings, using either OpenAI's NLP model or text from 10-K filings in the US. Rows 4 to 6 demonstrate different methods of aggregating information from projects at the topic level. The baseline averages the cosine similarity at the project level across topics. Here we consider using the topic level median instead, weighting each project by the number of tokens, or aggregating at the topic level before constructing the cosine similarity. Row 7 groups together topics that are conceptually similar (see Table C.3) and rows 8 and 9 cluster standard errors at the industry or topic level. The figure shows 95% confidence intervals.

Figure A.6: Effect of Proximity on CSR Spending, Robustness II

(a) Dropping Each Topic in Turn

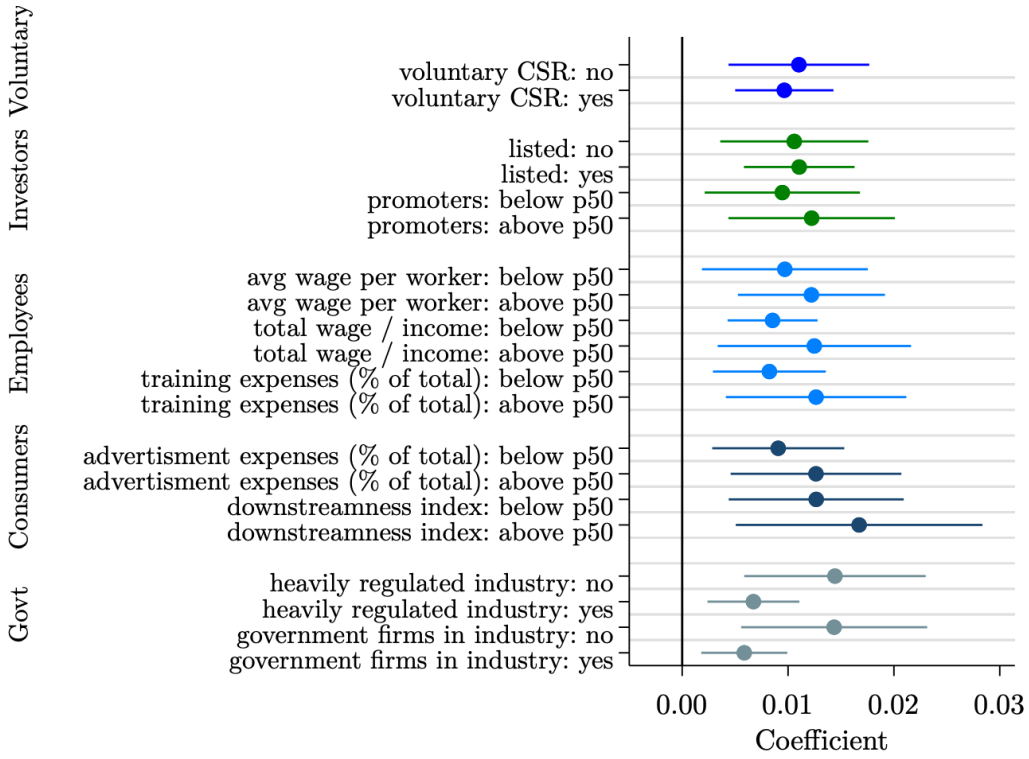


(b) Dropping Each of the Top 20 Industries in Turn



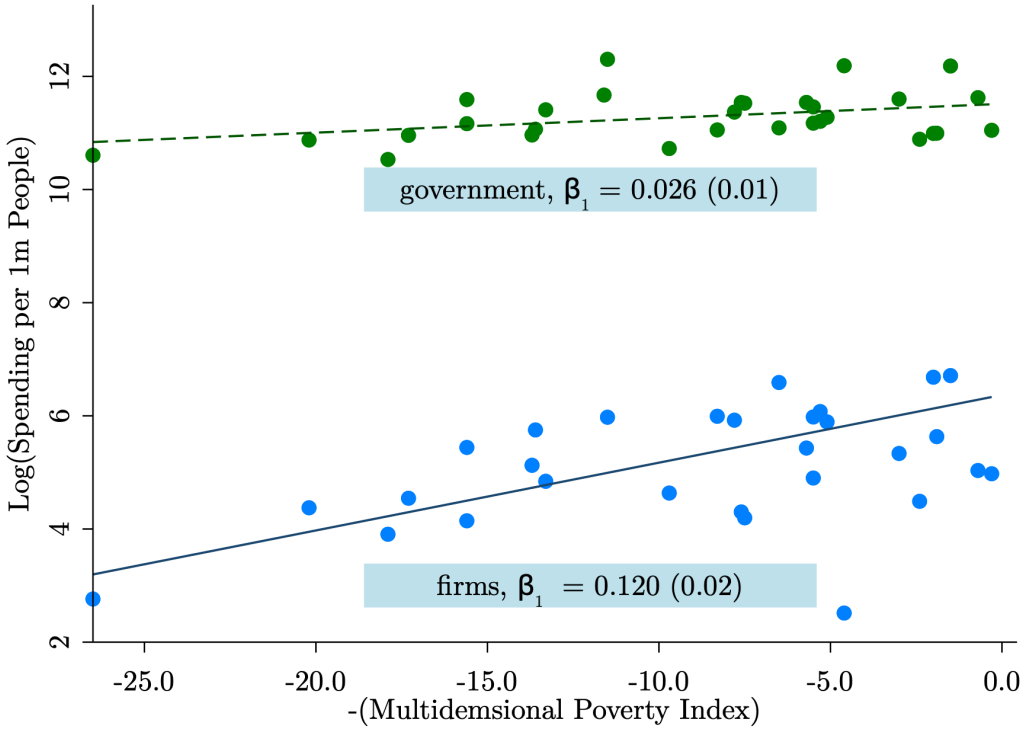
Notes: This figure describes the robustness of the relationship between CSR spending and proximity, derived from equation (10), dropping individual topics or industries. The unit of observation is at the firm×topic level. The dependent variable is the share of CSR spending that firm (f) spends on topic (d). $\text{Proximity}_{i(f)d}$ is the textual measure of closeness between an industry and a topic defined in section 5.1. The first row in each figure refers to the main specification from Table 2, Panel A, column 1. Figure A.6(a) describes robustness to dropping individual topics. Figure A.6(b) describes robustness to dropping individual industries. The figure shows 95% confidence intervals.

Figure A.7: Effect of Proximity on CSR Spending, Heterogeneity



Notes: This figure describes the heterogeneity of the relationship between CSR spending and proximity, derived from equation (10). The unit of observation is at the firm×topic level. The dependent variable is the share of CSR spending that firm (f) spends on topic (d). Proximity$_{i(f)d}$ is the textual measure of closeness between an industry and a topic defined in section 5.1. In the first group, the sample is split by firms that spend voluntarily and those who do not. Firms are defined as spending voluntarily if they spend more than 1% of their profits on CSR in 2014 before the policy or if they spend more than 2.5% of their profits on average in the years 2015 to 2019 after the policy. In the second group, the sample is split by how exposed firms are to investors, measured by whether the firms are listed on stock exchanges and whether the equity share of promoters is below or above median. Promoters in the Indian context are investors who own a significant stake in the company and play a key role in its management and decision-making. In the third group, the sample is split by how exposed firms are to employees, measured by the average wage per worker, the total wage bill over total income, and employee training expenses over total expenses. In the fourth group, the sample is split by how exposed firms are to consumers, as measured by advertisement expenses over total expenses, and a downstreamness index of the industry (obtained from Antràs et al. (2012)). In the fifth group, the sample is split by how exposed firms are to the government, as measured by whether the industry is heavily regulated (based on Awasthi et al. (2019)) and whether government firms are present in the Prowess sample for that industry. If not otherwise specified, heterogeneity variables are obtained from the 2013 Prowess accounting data. The figure shows 95% confidence intervals.

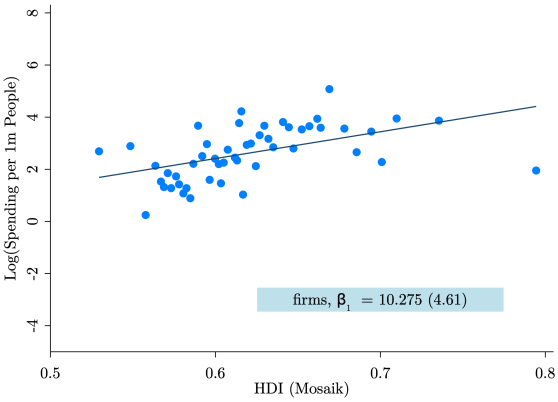
Figure A.8: Effect of State-Level Poverty on CSR and Government Spending



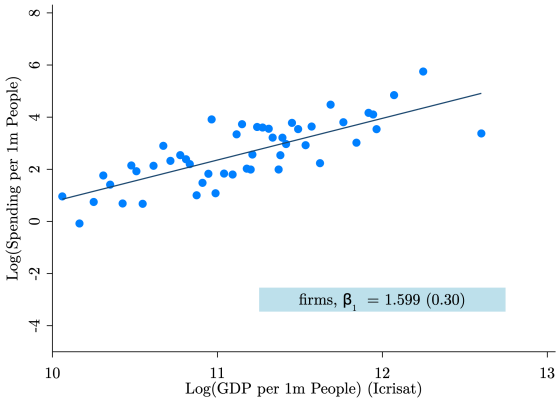
Notes: This figure depicts the relationship between a state-level multidimensional poverty index and firm (CSR) and government spending. The scattered dots indicate state-level observations, blue for firm spending and green for government spending. The lines indicate fitted linear approximations, blue for firm spending (solid) and green for government spending (dashed). The unit of observation is at the state level. The dependent variables are the log of spending (in millions, denominated by 2015 INR) per one million people by firms (CSR) and the government, aggregated from 2015 to 2019. The independent variable is the multidimensional poverty index in 2015/2016, multiplied by minus one (Government of India, 2024). Observations are weighted by the 2011 population.

Figure A.9: CSR Spending and District-Level Measures

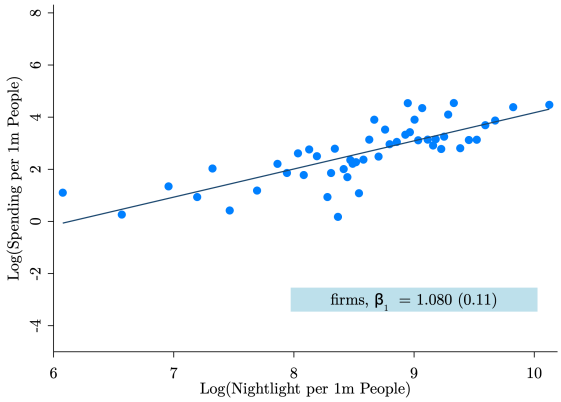
(a) Human Development Index (Mosaik)



(b) District GDP (Icrisat)

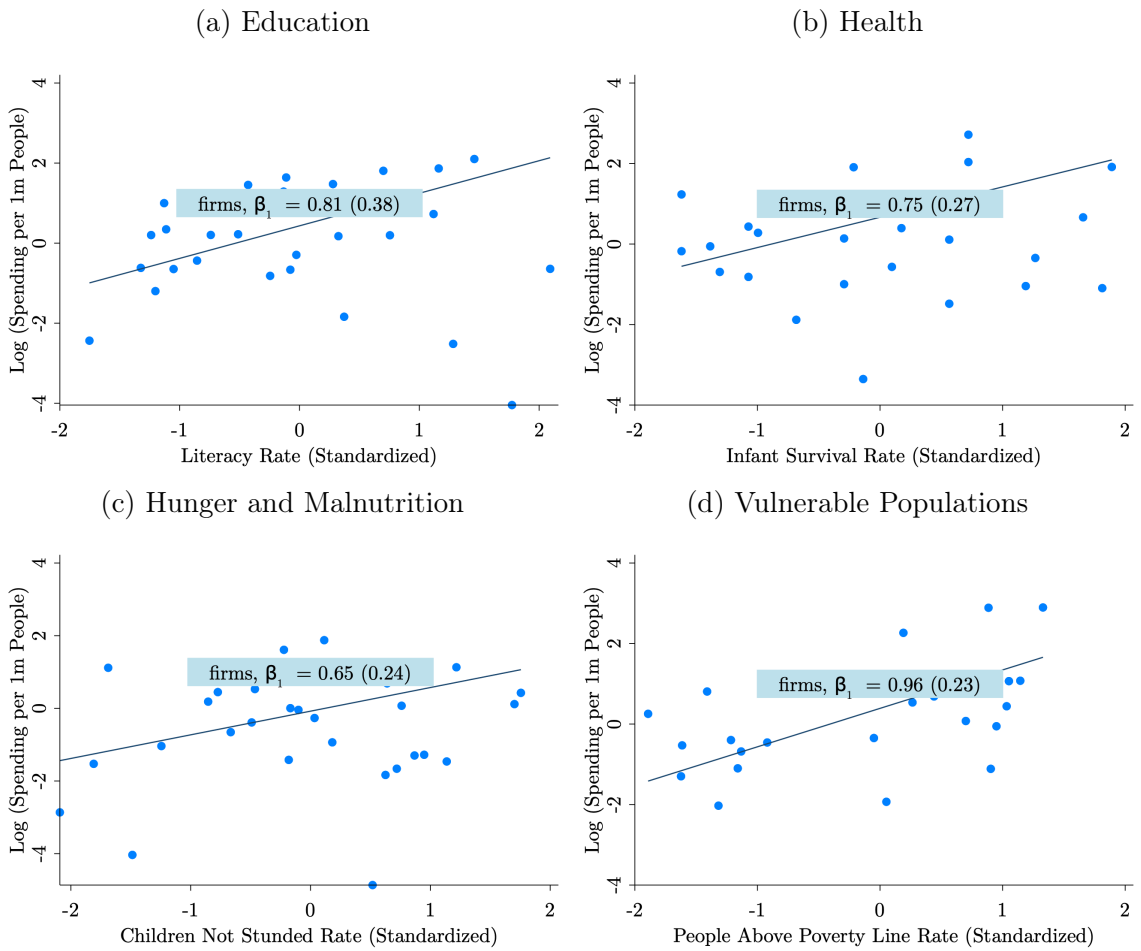


(c) Nightlight (SHRUG)



Notes: This figure depicts the relationship between CSR spending and district-level proxies for local development. The unit of observation is at the district level, and observations are binned into 50 equal-sized bins. The lines indicate fitted linear approximations. The dependent variables are the log of spending (in millions, denominated by 2015 INR) per one million people by firms (CSR), aggregated from 2015 to 2019. The independent variables are a human development index in Figure A.9(a) (Mosaik, 2023), a GDP measure in Figure A.9(b) (ICRISAT, 2020), and nightlight as a proxy for local economic conditions in Figure A.9(c) (Henderson et al., 2011; Asher et al., 2021; SHRUG, 2024). Observations are weighted by the 2011 population.

Figure A.10: CSR Topic Spending and Development Indicators



Notes: This figure depicts the relationship between CSR spending on given topics and the respective state-level development indicators. The scattered dots indicate state-level observations. The lines indicate fitted linear approximations. The unit of observation is at the state level. The dependent variable is the log of CSR spending (in millions, denominated by 2015 INR) per one million people on a given topic, aggregated from 2015 to 2019. In Figure A.10(a), the independent variable is the literacy rate in 2011 (Government of India, 2021b). In Figure A.10(b), the independent variable is the survival rate for infants in 2013 (Government of India, 2021a). In Figure A.10(c), the independent variable is the rate of children not stunted in 2016 (Government of India, 2020). In Figure A.10(d), the independent variable is the rate of people living above the poverty line in 2012 (Government of India, 2014). Observations are weighted by the 2011 population.

Table A.1: Summary Statistics

	Mean	SD	Median
Firm-level			
Income (m INR)	15,137	96,366	3,008
Voluntary CSR (yes/no)	0.258	0.438	0.000
Firm-topic-level			
CSR share unconditional (%)	0.062	0.189	0.000
Any CSR spending (yes/no)	0.223	0.416	0.000
CSR share conditional (%)	0.280	0.314	0.142
Firm-state-level			
CSR share unconditional (%)	0.029	0.151	0.000
Any CSR spending (yes/no)	0.067	0.250	0.000
CSR share conditional (%)	0.437	0.401	0.290
Observations			
Unique firms (nr)	6,573		
Unique firm×topics (nr)	105,168		
Unique firm×states (nr)	196,415		

Notes: This table describes the merged MCA and Prowess data (2015 to 2019). Income is calculated as an annual average over the time period, in real terms, denominated in 2015 INR. Firms are defined as spending voluntarily if they spend more than 1% of their profits on CSR in 2014 before the policy or if they spend more than 2.5% of their profits on average in the years 2015 to 2019 after the policy. CSR share unconditional is the share of CSR expenditure a firm spends on a topic or in a state. Any CSR spending is an indicator that equals one if the firm spends on a topic or state. CSR share conditional is the CSR share conditional on any spending on a topic or state. CSR spending is aggregated over the time period. Variables are not winsorized.

Table A.2: Firm CSR Share Proximity and Firm Product Share Proximity

	Firm CSR Share Proximity$_{ff'}$	
	(1)	(2)
Firm Product Share Proximity$_{ff'}$	0.04***	0.04**
	(0.01)	(0.02)
Firm Clustering		✓
R-squared	0.00	0.00
Observations	158,203	158,203

Notes: This table describes the relationship between the firm CSR share proximity and the firm product share proximity across all pairs of firms. The unit of observation is at the firm-level. Firm Product Share Proximity$_{ff'}$ is the cosine similarity of the vectors of product shares of firms f and f'. Let $\boldsymbol{\theta_f}$ be the vector of product sales shares $\theta_{fj} = (\text{Sales}_{fj}/\text{Sales}_f)$, the share of product j in the sales of firm f. Then,

$$\text{Firm Product Share Proximity}_{ff'} = \cos(\boldsymbol{\theta_f}, \boldsymbol{\theta_{f'}}) \tag{A.1}$$

Firm CSR Share Proximity$_{ff'}$ is the cosine similarity of the vectors of CSR shares of firms f and f'. Let $\boldsymbol{\psi_f}$ be the vector of CSR shares $\psi_{fd} = (\text{CSR}_{fd}/\text{CSR}_f)$, the share of topic d in the CSR of firm f. Then,

$$\text{Firm CSR Share Proximity}_{ff'} = \cos(\boldsymbol{\psi_f}, \boldsymbol{\psi_{f'}}) \tag{A.2}$$

We keep pairs of firms (f, f') such that $f < f'$ to avoid duplicates. Standard errors are not clustered in column 1 and clustered at the firm-level in column 2. $***$, $**$ and $*$ indicate significance at the 1%, 5% and 10% levels.

Table A.3: Firm Semantic Proximity Within and Across Industries

	Firm Semantic Proximity$_{ff'}$		
	2-digit	3-digit	4-digit
	(1)	(2)	(3)
Same Industry$_{ff'}$	0.118***	0.146***	0.155***
	(0.000)	(0.000)	(0.001)
Constant	0.679***	0.681***	0.681***
	(0.000)	(0.000)	(0.000)
R-squared	0.03	0.02	0.02
Observations	5,808,936	5,808,936	5,808,936

Notes: This table describes the average textual proximity of firm level descriptions between pairs of firms in the same industry, relative to firms in different industries. The unit of observation is at the firm-level. Firm Semantic Proximity$_{ff'}$ is defined as the cosine similarity of the embeddings characterizing firms f and f', obtained from the text in SEC 10-K filings. Same Industry$_{ff'}$ is a dummy equal to 1 if firms f and f' belong to the same SIC industry, at the 2, 3, and 4-digit levels in columns 2 to 4, respectively. We keep pairs of firms (f, f') such that $f < f'$ to avoid duplicates. $***$, $**$ and $*$ indicate significance at the 1%, 5% and 10% levels.

Table A.4: Firms' Production Across Industries and Industry Semantic Proximity

	Product Sales Share$_{f\ i(prod)}$		Any Product Sales Value$_{f\ i(prod)}$	
	(1)	(2)	(3)	(4)
Industry Semantic Proximity$_{i(f)i(prod)}$	0.037**	0.055**	0.100***	0.131***
	(0.015)	(0.022)	(0.020)	(0.027)
Avg dep var	0.045	0.045	0.055	0.055
Industry(Product) FE		✓		✓
R-squared	0.00	0.02	0.01	0.03
Observations	38,409	38,409	38,409	38,409

Notes: This table describes whether firm f in industry $i(f)$ is more likely to produce in industry $i(prod)$ if industries $i(f)$ and $i(prod)$ have similar embeddings. Industry Semantic Proximity$_{i(f)i(prod)}$ is the proximity between industry i of the firm and industry i of the product as defined by the cosine similarity of their respective embedding vectors. In columns 1 and 2, the dependent variable is firm's sales value of products in a given industry $i(prod)$, calculated over the entire sales value of the firm f. In columns 3 and 4, the dependent variable is an indicator that equals one if firm f produces in $i(prod)$. Columns 2 and 4 include fixed effects for the industry of the products. The unit of observation is at the firm×industry(product)-level. We exclude observations for which the firm industry equals the product industry. Standard errors are clustered on the industry level of the firm. $***$, $**$ and $*$ indicate significance at the 1%, 5% and 10% levels.

Table A.5: Industries' Input-Output Features and Industry Semantic Proximity

	IO Proximity$_{ii'}$		
	(1)	(2)	(3)
Industry Semantic Proximity$_{ii'}$	0.069***	0.434***	0.595***
	(0.007)	(0.056)	(0.058)
Benchmark	Leontief	Input	Output
R-squared	0.08	0.06	0.10
Observations	990	990	990

Notes: This table describes the relationship of textual proximity between pairs of industries and other benchmarks that capture proximity across industries. For each pair of industry ii', Industry Semantic Proximity$_{ii'}$ is the proximity between industry i and i' as defined by the cosine similarity of their respective embedding vectors. IO Proximity$_{ii'}$ are metrics for proximity between industries based on the input-output matrix for India (obtained from the OECD). Leontief$_{ii'}$ is the ii' entry of the Leontief inverse of the input-output matrix. Input$_{ii'}$ is the cosine similarity of the input shares of industries i and i'. Output$_{ii'}$ is the cosine similarity of the output shares of industries i and i'. The unit of observation is at the industry-level. We keep pairs of industries (i, i') such that $i < i'$ to avoid duplicates. $***$, $**$ and $*$ indicate significance at the 1%, 5% and 10% levels.

Table A.6: CSR Spending and Proximity on the Industry×Topic Level

	CSR Share Unconditional$_{id}$ (1)	Log(CSR Share Unconditional)$_{id}$ (2)
Proximity$_{id}$	0.020***	0.340***
	(0.006)	(0.063)
Avg dep var	0.072	.
Industry FE	✓	✓
Topic FE	✓	✓
R-squared	0.66	0.53
Observations	983	983

Notes: This table describes the relationship between CSR spending and proximity, derived from equation (10), on the industry×topic level. The dependent variable is the share of CSR spending of an industry (i) over topics (d). Proximity$_{id}$ is the textual measure of closeness between an industry and a topic defined in section 5.1. In column 1, the outcome is in levels. In column 2, the outcome is log transformed. Standard errors are clustered at the industry×topic level. $***$, $**$ and $*$ indicate significance at the 1%, 5% and 10% levels.

Table A.7: Comparison Explanatory Power of Proximity and Government Spending

	CSR Share Unconditional$_{fd}$			
	(1)	(2)	(3)	(4)
Proximity$_{i(f)d}$	0.017***	0.012**	0.018*	0.022***
	(0.005)	(0.005)	(0.010)	(0.007)
Government Share$_d$		0.093***		0.072***
		(0.004)		(0.006)
Weight	None	None	CSR spending	CSR spending
Var. explained proximity	0.167		0.180	
Var. explained gov share		0.929		0.718
Firm FE	✓	✓	✓	✓
Topic FE	✓		✓	
R-squared	0.21	0.16	0.25	0.17
Observations	65,730	65,730	65,730	65,730

Notes: This table compares the explanatory power of proximity and government spending shares on the allocation of CSR. Topics are aggregated at a level consistent between CSR topics and government expenditures, as indicated in Table C.3. Variation explained is the standard deviation of the proximity variable (columns 1 and 3) or government variable (columns 2 and 4) multiplied by the estimated coefficient of the respective variable, divided by the mean of the outcome variable. The unit of observation is at the firm×topic level. The dependent variable is the share of CSR spending of a firm (f) over topics (d). Proximity$_{i(f)d}$ is the textual measure of closeness between an industry and a topic defined in section 5.1. Government Share$_d$ is the share of government spending on a given topic d. Standard errors are clustered at the industry×topic level. $***$, $**$ and $*$ indicate significance at the 1%, 5% and 10% levels.

Table A.8: Indirect Project Implementation and Proximity

	Indirect Project Implementation$_{pfd}$	
	Excl Text Data	Incl Text Data
	(1)	(2)
Proximity$_{i(f)d}$	-0.007**	-0.009**
	(0.004)	(0.004)
Avg dep var	0.514	0.550
Firm FE	✓	✓
Topic FE	✓	✓
R-squared	0.62	0.57
Observations	82,201	82,201

Notes: This table describes the relationship between indirect project implementation and proximity, based on equation (A.4). The unit of observation is at the firm×project level. The dependent variable is an indicator that is one if the project is implemented indirectly. In column 1, we define a project to be implemented indirectly if firms report this to be the case based on an MCA question. In column 2, we additionally count projects as indirect if their textual description contains a token or bigram indicative of third-party implementation or an NGO (see Appendix C.1.5). Proximity$_{i(f)d}$ is the textual measure of closeness between an industry and a topic defined in section 5.1. Observations are weighted by the inverse of the number of projects by topic to give equal weight to all topics, as in the main proximity regressions. Standard errors are clustered at the topic×industry level. $***$, $**$ and $*$ indicate significance at the 1%, 5% and 10% levels.

$$\text{Indirect Project Implementation}_{pfd} = \beta_0 + \text{Proximity}_{i(f)d} + \alpha_f + \alpha_d + \varepsilon_{pfd} \tag{A.4}$$

B Results for Alternative Samples

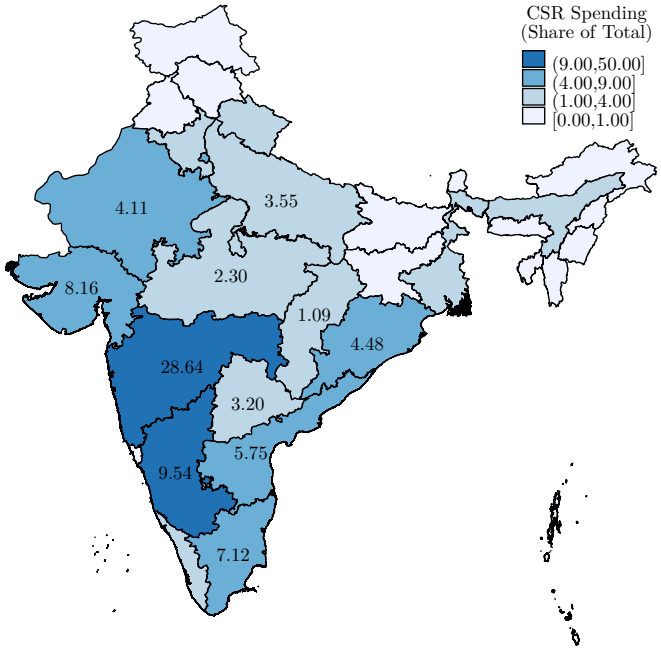
This appendix reproduces our key findings, when feasible, for two alternative samples. First, the sample of all 11,487 firms in the MCA data. By definition, we do not have industry information for firms that are only in the MCA data but are not in the accounting data. Thus, we cannot replicate our results in section 5 for this sample. Second, the voluntary CSR sample of 1,602 firms, defined in section 2.

Table B.1 presents the allocation of CSR expenditures across topics (see Table 1 for results on the main sample). We see that the allocation is very similar in these two samples compared to our main sample. Consequently, both our fact 1 (CSR spending is concentrated on health and education) and fact 2 (firms' allocation across topics correlates with the allocation of other public good providers) hold in all samples. Fact 3 (firms specialize in topics) is also similar in those samples. In the full sample of all firms in the MCA data, we observe an average Herfindahl-Hirschman Index of 0.64, and 38% of firms allocate more than 90% of their spending to only one topic. Subsetting on the firm $\times$ year observations where firms report multiple projects, 25% of firms allocate more than 90% of their spending to only one topic. In the sample of firms that spend voluntarily, we observe an average Herfindahl-Hirschman Index of 0.60, and 29% of firms allocate more than 90% of their spending to only one topic. Subsetting on the firm $\times$ year observations where firms report multiple projects, 19% of firms allocate more than 90% of their spending to only one topic. Finally, Figure B.1 demonstrates that fact 4 (CSR spending is highly concentrated geographically) also looks similar in all three samples.

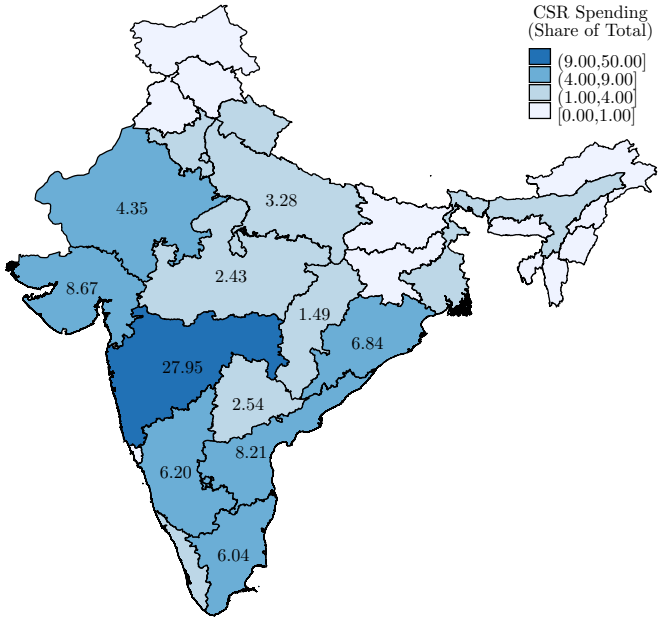
Table B.2 demonstrates that the results on allocative efficiency are similar if we restrict the sample to firms in the voluntary CSR sample. In terms of the relationship between state-level GDP and CSR spending, Figure B.2 draws a picture for all firms in the MCA data and firms that spend voluntarily that is similar to our main sample. Table B.3 provides similar evidence at the firm-level. To summarize, both the descriptive facts and findings that we demonstrate on our main sample hold for the sample of all firms in the MCA data and firms that spend voluntarily.

Figure B.1: Allocation of Spending Across States (Full MCA Sample and Voluntary CSR Sample)

(a) Full MCA Sample



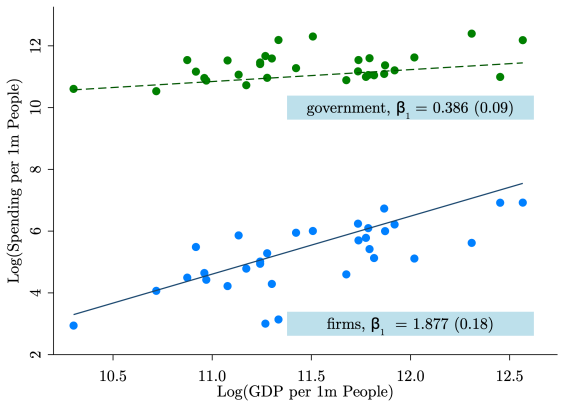
(b) Voluntary CSR Sample



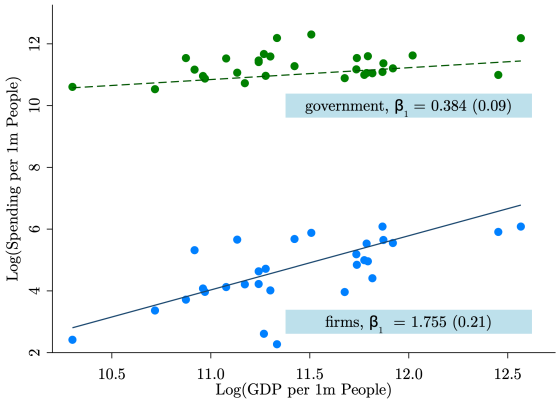
Notes: This figure depicts CSR spending shares by state, for all firms in the MCA data and for the voluntary CSR sample.

Figure B.2: CSR Spending, Government Spending, and State-Level GDP (Full MCA Sample and Voluntary CSR Sample)

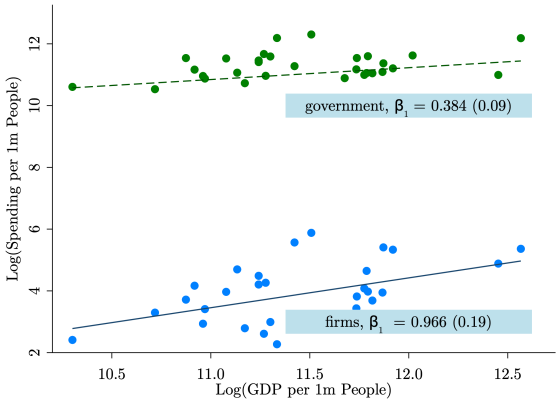
(a) CSR Spending (Full MCA Sample)



(b) CSR Spending (Voluntary CSR)



(c) CSR Outside HQ States (Voluntary CSR)



Notes: This figure presents the relationship between state-level GDP as well as firm (CSR) and government spending, for all firms in the MCA data and for the voluntary CSR sample. CSR spending outside of headquarter states is not provided for the full MCA sample since data on headquarters is obtained from Prowess. The scattered dots indicate state-level observations, blue for firm spending and green for government spending. The lines indicate fitted linear approximations, blue for firm spending (solid) and green for government spending (dashed). The unit of observation is at the state level. The dependent variables are the log of spending (in millions, denominated by 2015 INR) per one million people by firms (CSR) and the government, aggregated from 2015 to 2019. The independent variable is the log of state-level GDP (in millions, denominated by 2015 INR) per one million people in 2013. Observations are weighted by the 2011 population.

Table B.1: Spending Share by Topic (Full MCA Sample and Voluntary CSR Sample)

Topic	CSR Share Full MCA Sample	CSR Share Voluntary CSR Sample
Education	32%	30%
Health	18%	18%
Infrastructure	8%	8%
Environmental sustainability	8%	8%
Vocational skills	5%	6%
Technology	5%	6%
Livelihood enhancement	5%	4%
Sanitation	4%	5%
Hunger and malnutrition	4%	4%
Safe drinking water	2%	2%
Vulnerable populations	2%	2%
Emergency relief	2%	1%
Sports	2%	2%
Women empowerment	1%	1%
Agroforestry	1%	0%
Animal welfare	0%	0%

Notes: This table displays the share of total CSR spending by topic for all firms in the MCA data and for the voluntary CSR sample.

Table B.2: CSR Spending and Proximity (Voluntary CSR Sample)

	CSR Share Unconditional$_{fd}$ (1)	Any CSR Spending$_{fd}$ (2)	CSR Share Conditional$_{fd}$ (3)
	Panel A: Not Weighted		
Proximity$_{i(f)d}$	0.010***	0.021***	0.018***
	(0.002)	(0.004)	(0.006)
Avg dep var	0.062	0.261	0.239
Firm FE	✓	✓	✓
Topic FE	✓	✓	✓
R-squared	0.27	0.36	0.36
Observations	25,632	25,632	6,360
	Panel B: Weighted by Total CSR Spending		
Proximity$_{i(f)d}$	0.012**	0.047***	0.005
	(0.006)	(0.016)	(0.011)
Avg dep var	0.062	0.506	0.124
Firm FE	✓	✓	✓
Topic FE	✓	✓	✓
R-squared	0.29	0.38	0.34
Observations	25,632	25,632	6,360

Notes: This table describes the relationship between CSR spending and proximity, derived from equation (10), for the voluntary CSR sample. The unit of observation is at the firm×topic level. The dependent variables are the share of CSR spending of a firm (f) over topics (d), an indicator for any CSR spending by firm (f) in a given topic (d), and the share of CSR spending conditional on any spending. Proximity$_{i(f)d}$ is the textual measure of closeness between an industry and a topic defined in section 5.1. In Panel A, observations are unweighted. In Panel B, observations are weighted by the total CSR spending of each firm, winsorized at the 1st and 99th percentile. Standard errors are clustered at the industry×topic level. $***$, $**$ and $*$ indicate significance at the 1%, 5% and 10% levels.

Table B.3: CSR Spending, State-Level Characteristics, and Firm Headquarter (Full MCA Sample and Voluntary CSR Sample)

	CSR Share Unconditional$_{fs}$		Any CSR Spending$_{fs}$		CSR Share Conditional$_{fs}$	
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Full MCA Sample					
Log(GDP per 1m People)$_s$	0.097***		0.152***		0.137***	
	(0.030)		(0.040)		(0.020)	
Avg dep var	0.029		0.061		0.481	
Firm FE	✓		✓		✓	
R-squared	0.09		0.17		0.41	
Observations	333,240		333,240		13,991	
	Panel B: Voluntary Firms					
Log(GDP per 1m People)$_s$	0.087***	0.011**	0.143***	0.051***	0.102***	-0.026
	(0.024)	(0.005)	(0.031)	(0.010)	(0.020)	(0.017)
1(Firm Headquarter State)$_{fs}$		0.585***		0.709***		0.359***
		(0.034)		(0.027)		(0.024)
Avg dep var	0.030	0.030	0.077	0.077	0.386	0.386
Firm FE	✓	✓	✓	✓	✓	✓
R-squared	0.07	0.49	0.20	0.43	0.41	0.57
Observations	47,947	47,947	47,947	47,947	2,890	2,890

Notes: This table describes the relationship between state-level characteristics as well as firm headquarters and CSR spending, derived from equation (11), for all firms in the MCA data and firms in the voluntary sample. Headquarter information is not available for all firms in the MCA data since it is obtained from Prowess. The unit of observation is at the firm×state level. The dependent variables are the share of CSR spending of a firm (f) over topics (d), an indicator for any CSR spending by firm (f) in a given topic (d), and the share of CSR spending conditional on any spending. The independent variables are the log of state-level GDP per 1 million people and an indicator that equals one if the firm is headquartered in the state as per government records. We control for the log of population in millions. Observations are weighted by the 2011 population. Standard errors are clustered at the state-level. $***$, $**$ and $*$ indicate significance at the 1%, 5% and 10% levels.

C Data and Variable Construction

C.1 CSR Data

C.1.1 Sample Overview

The raw data contains 176,988 projects from 15,632 firms. To derive the main sample for our analysis, we take the following steps. First, we exclude projects where the text indicates an issue for classification across topics (see flags detailed in section C.1.2). This leaves 124,813 projects by 11,487 firms. Second, we match 61% of these firms and 91% of the CSR expenditure in the CSR data to the accounting data, resulting in 92,180 projects of 7,064 firms.[27] Third, we apply an additional set of cleaning steps, excluding: holding companies, firms without industry information, firms with zero CSR spending, and projects that cannot be assigned to a topic. This results in our final dataset of 86,334 projects by 6,573 firms. In addition, when we consider allocation across locations, we exclude four states with a population that is lower than one million, as well as the small state of Chandigarh, which does not have government spending data, and firms without headquarter locations. For our analysis at the firm×topic and firm×state levels, we fill the data, meaning that even if a firm does not spend on a topic or in a state, we include an observation with zero spending.

C.1.2 Cleaning of Project Descriptions Data

We execute the following steps to clean the CSR project descriptions:

1. Convert the text to lowercase, removing special characters and numbers
2. Tokenize the text (splitting strings into tokens), lemmatize, and stem the tokens
3. Translate Hindi tokens to English
4. Remove uninformative tokens:
    - Create a list of common uninformative token sequences: it includes common token sequences found in the project descriptions but unrelated to CSR projects, e.g., "CSR overheads", "project not found", "administration expenditures", "detailed in report", etc.
    - Remove tokens flagged as uninformative

[27] The accounting data is obtained from https://prowessdx.cmie.com.

Flags for uninformative descriptions. In addition, we flag observations with uninformative descriptions for the purpose of our textual analysis, defined as satisfying at least one of the following criteria:

1. No project description
2. "Word salads": project description lists the titles of many different topics
3. Project description where more than 60% of original bigrams correspond to uninformative token sequences defined above
4. Project description Word2Vec embedding is empty (e.g., if the only remaining token following the cleaning is a proper noun)
5. Project description equal (or highly similar) to the title of the topic or groupings of topics on the CSR portal: this avoids using project descriptions that are just a repetition of the title of the topic.

The raw data contains 176,988 projects. This procedure flags 68,992 observations as having uninformative descriptions. Observations flagged in 5 are not used in our textual corpus, but can be reliably classified across topics and hence enter our firm×topic CSR spending data.

C.1.3 Vectorization of the Textual Data

We employ the methodology introduced by Mikolov (2013) to transform all textual tokens into numerical representations suitable for analysis.

This method involves encoding words as numerical vectors, known as word embeddings, which capture semantic meaning based on the context in which they appear. There are two primary ways to generate these embeddings: (1) training them on a custom corpus containing domain-specific documents or (2) utilizing pre-trained embeddings derived from a large, general-purpose corpus that includes the words of interest. We opt for the latter approach because (a) we lacked access to a sufficiently large and diverse collection of documents to train reliable embeddings, and (b) existing research supports the effectiveness of pre-trained word embeddings (Rios and Lwowski, 2020). Specifically, we use the 300-dimensional embeddings from the Word2vec model of Mikolov (2013), which were trained on Google News data encompassing approximately 3 million words.[28]

[28]The model can be downloaded at https://code.google.com/archive/p/word2vec/.

After obtaining word embeddings for each token, we aggregate them into a single vector representation for each project description using a weighted average:

$$\mathbf{v}_p = \frac{1}{N_p} \sum_{j=1}^{N_p} w_{jp} * \mathbf{v}_{jp} \tag{C.1}$$

where $\mathbf{v}$ is the embedding vector associated to project p, $\mathbf{v}_{jp}$ is the embedding vector associated to token j, w_{jp} is the weight of token j, and N_p is the number of tokens in project p.

The weights w_{jp} are determined using the Term Frequency-Inverse Document Frequency (TF-IDF) method. TF-IDF assigns importance to each word based on how frequently it appears in a specific document (Term Frequency) while reducing the weight of commonly occurring words across all documents (Inverse Document Frequency). TF-IDF is widely used in natural language processing and information retrieval. To avoid TF-IDF systematically down-weighting terms frequently appearing in the largest topics (e.g., "school" in the topic education), we construct the TF-IDF weights in a corpus that has an equal number of projects for each topic. To construct this corpus, we use all observations in the topic with the largest number of observations ($N_{max} = 39,213$) and sample N_{max} observations with replacement in all the topics with a number of observations smaller than N_{max}.

Unless specified, $\mathbf{v}_p$ is normalized to have norm 1.

C.1.4 Classification Across Topics

The initial data contains 28 different topics. This initial classification has three issues: (i) some topics have only a handful of observations so that the project descriptions would be insufficient to reliably estimate proximity with the firms' industries; (ii) some topics have a large overlap in terms of the vocabulary they use; (iii) some observations are clearly misclassified (e.g., a project with description 'school construction' being classified as sanitation). We proceed as follows.

Step 1: Manually classify the 28 original topics into 19 aggregated topics. We reclassify all topics with less than 2,000 observations unless there is no sufficiently close topic. This mapping is detailed in Table C.1, which also shows the number of observations by initial topic. Figure C.1 reports the average pairwise similarity

between the initial topics.

Step 2: Automatic correction for misclassifications. The intuition for the procedure is that we detect an observation as being misclassified if it is significantly more similar to the average description in other topics than in its own topic. We use the following algorithm:

1. Compute average embedding of projects $\mathbf{v}_{pd(p)}$ (not normalized) in topic d:

$$\mathbf{v}_d = \frac{\sum_{p=1}^{N_d} \mathbf{v}_{pd(p)}}{N_d} \tag{C.2}$$

2. For each project p in topic d, compute:
   - The similarity between p and its own topic $d(p)$: $\text{OwnProximity}_p = \cos(\mathbf{v}_p, \mathbf{v}_{d(p)})$
   - The largest similarity between p and any topic d': $\text{Max1Proximity}_p = \max_{d'} \cos(\mathbf{v}_p, \mathbf{v}_{d'})$
   - The topic d' with the maximum similarity: $\text{IsMax}_p = \arg\max_{d'} \cos(\mathbf{v}_p, \mathbf{v}_{d'})$
   - The second-largest similarity between p and any topic d': $\text{Max2Proximity}_p = \max_{d' \neq \text{IsMax}_p} \cos(\mathbf{v}_p, \mathbf{v}_{d'})$
3. If $\text{Max1Proximity}_p > \lambda^{miscl} \times \text{OwnProximity}_p$, we say observation p is misclassified
   - If $\text{Max1Proximity}_p > \lambda^{recl} \times \text{Max2Proximity}_p$, we reclassify observation p to topic IsMax_p
   - Otherwise, we say observation p cannot be classified
4. Discard observations that cannot be classified
5. Repeat until the topic assigned in iteration n is the same as the topic assigned in iteration $n + 1$.

In our implementation, we use $\lambda^{miscl} = \lambda^{recl} = 1.2$. We impose a number of additional rules:

1. Observations in agroforestry contain projects related to farming as well as projects related to environmental sustainability. Because of the small number of observations in agroforestry, the algorithm tends to reclassify environmental

projects in agroforestry. We manually assign the projects containing the tokens ['tree plantation', 'protection flora fauna', 'maintenance flora fauna', 'biodiversity protection', 'environmental sustainability'] to environmental sustainability.

2. Rural development contains a mix of infrastructure projects and of projects corresponding to the other topics but implemented in rural areas. We reclassify observations in this topic using $\lambda^{miscl} = \lambda^{recl} = 1.05$ so that observations related to other topics are reclassified "aggressively", and observations remaining in rural development mostly consist of infrastructure projects.
3. 'Slum area development' and 'Other central government funds' consist of highly heterogeneous projects. We force the reclassification of projects in these topics into the closest topic.

Step 3: Assigning topics to observations with missing value. Having obtained a clean definition of each topic, we attempt to classify observations with a missing topic. Using a methodology similar to the one described above, we assign an observation p with missing topic to topic d if p has significantly higher proximity to d than all other topics d'.

Looking at the word clouds for each topic, the procedure performs well. A caveat of our methodology is that all the projects with tokens in Hindi that we could not translate (proper nouns, spelling mistakes) get bundled in the topic "Art and culture" which has the initial largest share of such tokens. In the absence of solution to this caveat, we remove this topic from our baseline analysis. Since this topic is small, this does not materially affect our results.

C.1.5 Definition of Third-Party Implementation

To define third-party implementation, we initially rely on the original MCA variable. This variable is self-reported by the firm and likely contains measurement error. In particular, the mandate does not detail what extent of outsourcing implies that firms should declare their CSR projects as implemented via a third-party. We find for example that the probability of indirect implementation is only weakly correlated with the likelihood that the project description contains an NGO name. We consider a version of the variable where we relabel projects as indirectly implemented if the project description contains text possibly referring to indirect implementation or an NGO name as a robustness check.

- Check for tokens and bigrams indicative of third-party implementation: ['gift', 'offering', 'grant', 'endowment', 'input', 'participation', 'support', 'contribution', 'donation', 'partnership', 'ngo', 'benefaction', 'collaboration', 'alliance', 'association', 'joint venture', 'cooperation', 'affiliation', 'organization', 'agency', 'implemented by', 'bureau', 'department', 'authority', 'office', 'establishment', 'advisory', 'counsel', 'guidance', 'expert advice', 'consulting', 'foundation', 'in collaboration', 'professional services', 'fund', 'executed by', 'carried out by', 'enacted by', 'put into effect by', 'non-profit organization', 'operated through', 'voluntary organization', 'nonprofit group', 'together', 'jointly', 'in partnership', 'working together', 'donations to']
- Check for mention of NGOs using a list of the 23 NGOs most frequently present in our dataset

C.1.6 Description of the Final Dataset

Figures C.2 to C.4 show word clouds of the project descriptions by topic.

C.2 Other Data Sources

C.2.1 Government Expenditures

We obtain government expenditures data from the Reserve Bank of India for 2010–2021, retrieved on 8th of December 2022.[29] We use state-level expenditures, which includes expenditures from central government transfers to state and covers most expenditures in India that correspond to the topics in the CSR data except some food subsidies implemented directly by the central government. We use the 'revised' expenditure variables and consider all regular and capital expenditures labeled 'development expenditures'. We exclude from our analysis expenditures that have no equivalent in the CSR data: energy and transport, tax collection expenditures, interest payments, 'organs of state' (this includes police and judiciary), grants to lower levels of governments (these represent 1% of state expenditures), and other (which includes items like tourism expenditures). The mapping between government expenditures and CSR topics is described in Table C.3.

[29] https://www.rbi.org.in/Scripts/AnnualPublications.aspx?head=State%20Finances%20:%20A%20Study%20of%20Budgets

C.2.2 NGO Activity

To compare CSR spending with NGO activity by topic, we exploit a report prepared by the state of Haryana that collects data on the 150 highest-capability NGOs operating in the state and classifies them by sustainable development goal (which we map to our topics).[30] The mapping between sustainable development goals and CSR topics is described in Table C.3.

C.2.3 Indian Industry Descriptions

The industry descriptions come from the National Industry Classification report from 2008. We extract the descriptions from its 'Detailed Structure' and 'Explanatory Notes' sections. We manually clean the text for typos and mentions of products/services that should be excluded from each industry. Table C.4 shows an example of the full text for one industry i. We obtain the word embeddings associated with the description of each industry i. We clean the text as described in steps 1–4 of section C.1.2. We obtain word embeddings using the same Word2vec model, again applying TF-IDF. For each industry i, we obtain an embeddings vector $\mathbf{v}_i$.

C.2.4 US 10-K filings

We construct an alternative version of the proximity metric where the industry corpus uses firm-level descriptions of firm activity using the US SEC 10-K filings. This is the same data as the one used by Hoberg and Phillips (2016) to define product similarity between pairs of firms.

First, we construct a crosswalk from US SIC codes to Indian NIC 2008 2-digit industry codes by combining published concordances from the Forum for Research in Empirical International Trade with manual adjustments for ambiguous cases.

Second, we merge the SEC 10-K data with Compustat (by Central Index Key) to obtain firms' SIC codes. We use our mapping to obtain firm's industries as per the NIC 2-digit classification. We extract only section 1 ("Business") of each 10-K filing. The text is sometimes very long, and the end of the business section often contains text not related to the business description (often it contains sections on the competition, regulation, etc.). To keep the text most closely related to the firm's

[30] https://sdgcc.in/wp-content/uploads/2020/10/SDG-NGO-LINKAGES.pdf

activity, we retain the first half of the text. We then apply the same text cleaning pipeline as for the Indian NIC textual descriptions.

When constructing the proximity metric, we want the text corpus to be balanced across industries. This ensures that the industry embeddings are as representative for each industry, and avoids TF-IDF from discounting words related to specific, overrepresented industries. Hence, we select up to 400 firm×year observations per NIC industry using random subsampling or oversampling for balance.[31]

C.3 Construction of CSR Project Clusters

For descriptive purposes, we implement k-means clustering to partition CSR projects into thematically coherent groups within each development topic. The k-means algorithm iteratively assigns observations to clusters by minimizing the within cluster sum of distances to centroids, then updates the centroids based on cluster membership until convergence (MacQueen, 1967; Lloyd, 1982). We employ the cosine distance as our distance measure: $d(p, p') = 1 - \cos(\mathbf{v_p}, \mathbf{v_{p'}})$.

For each of the 16 development topics, we perform cosine k-means clustering in a systematic grid search with $k \in \{3, \ldots, 7\}$ clusters, selecting the optimal number of clusters based on the highest silhouette score. The silhouette score of each project p assigned to cluster C_p is defined as follows:

$$s(p) = \frac{b(p) - a(p)}{\max\{a(p), b(p)\}}$$

where: $a(p) = \frac{1}{|C_p|-1} \sum_{p' \in C_p, p' \neq p} d(p, p')$ is the average distance between point p and all other points in the same cluster C_p; $b(p) = \min_{C \neq C_p} \left(\frac{1}{|C|} \sum_{p' \in C} d(p, p') \right)$ is the minimum average squared distance between point p and all points in another cluster $C \neq C_p$. The overall silhouette score for the clustering is then computed as the average over all points: $S = \frac{1}{N} \sum_{i=p}^{N} s(p)$. The silhouette score measures both cluster cohesion (how close observations are to their own cluster centroid) and separation (how distant they are from neighboring clusters), providing a robust metric

[31] We keep all observations if there are fewer than or equal to 400 per industry; if there are more, we use proportional subsampling within firms or, when there are more than 400 unique firms, randomly select one observation per firm (prioritizing those with non-missing employment and revenue data) until reaching the target; for industries with fewer than 400 observations, we use random oversampling with replacement.

for cluster quality that ranges from -1 to 1, with higher values indicating better-defined clusters (Rousseeuw, 1987). This optimization process ensures that each development topic achieves its most natural thematic subdivision.

Table C.2 summarizes the clustering results across development topics, where each row corresponds to a cluster within a specific social topic. The first column lists the social topic. The second column provides a cluster name. Column 3 lists the number of observations in each cluster. Column 4 lists the silhouette score of each cluster. Finally, columns 5 to 9 report the top five bigrams in project descriptions. The cluster name is based on manual inspection of the top five bigrams and trigrams and the project descriptions closest to the cluster centroid.

C.4 Construction of the Proximity Variable

This procedure requires that the CSR project description is sufficiently informative and that we could reliably classify the observation in a CSR topic. We therefore restrict the sample to observations not flagged as uninformative (see section C.1.2) and with a valid topic assignment (see section C.1.4). We thus work with 92,979 observations.

Define $\mathbf{v}_d$ as the average embedding of projects $\mathbf{v}_{pd(p)}$ in topic d:

$$\mathbf{v}_d = \frac{1}{N_d} \sum_{p \in \mathcal{P}_d} \mathbf{v}_{pd(p)} \tag{C.3}$$

Define

$$\text{Proximity}_{id} = \cos(\mathbf{v}_d, \mathbf{v}_i) \tag{C.4}$$

We also construct Proximity_{id} by taking the median or the mean across projects weighted by token count. The correlation coefficient between these different variants exceeds 0.98.

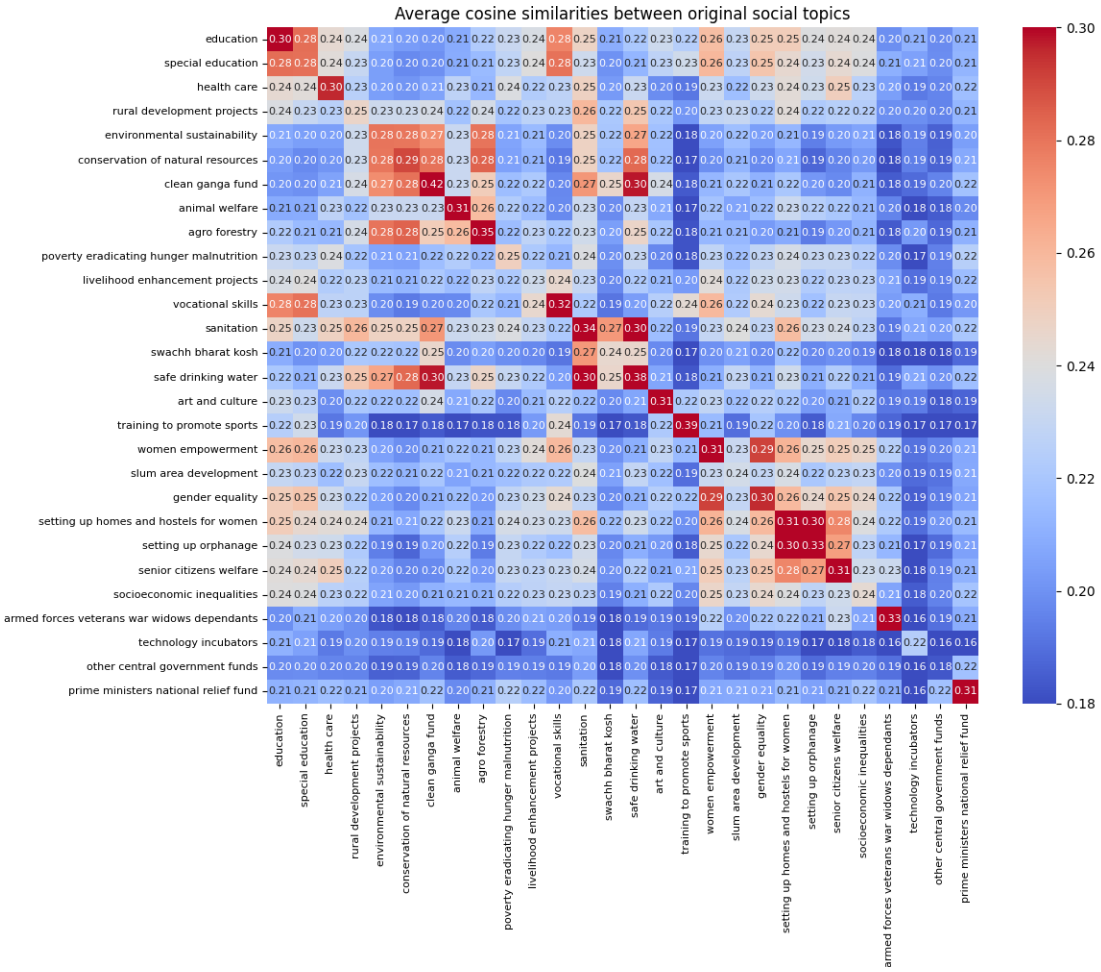
Analysis of the closest (furthest) CSR topics and industries. To interpret the semantic proximity between development topics and industries, we conduct an analysis of the token-level drivers underlying the proximity scores.

For each industry i, we identify the development topic d with the highest and lowest proximity, as defined in equation (9). Similarly, for each topic d, we identify the industry i with the highest and lowest proximity.

To understand the textual basis for these relationships, we extract the most representative tokens for each industry and topic. For each industry i and topic d, we select the 20 tokens in the text that are closest to the group centroid, and call these 'core tokens'. For each industry$\times$topic pair, we compute the cosine similarity between all possible pairs of core tokens, and report the 10 core token pairs with the highest similarity.

Table C.5 presents, for each industry i among the top 20 industries, the closest social topic d, and the closest token pairs associated to this pair (i, d) as defined above. Table C.6 repeats this exercise, but shows the furthest social topic d for each industry i. Tables C.7 and C.8 present the same results starting from social topics d and selecting the closest and further industry i, respectively.

Figure C.1: Average Cosine Similarity Between Initial Topics



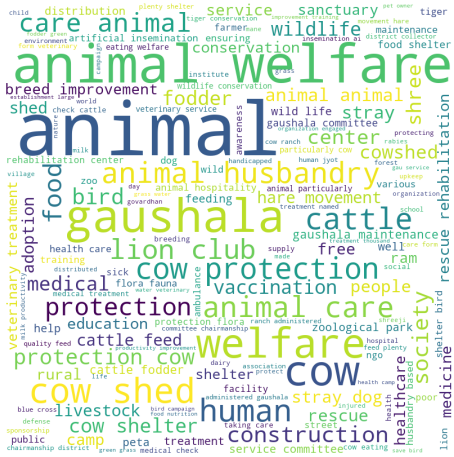
Notes: This figure depicts the average cosine similarity of projects for each pair of topics. The diagonal elements report the average cosine similarity of projects within a topic.

Figure C.2: Project Descriptions by Topics: Word Clouds

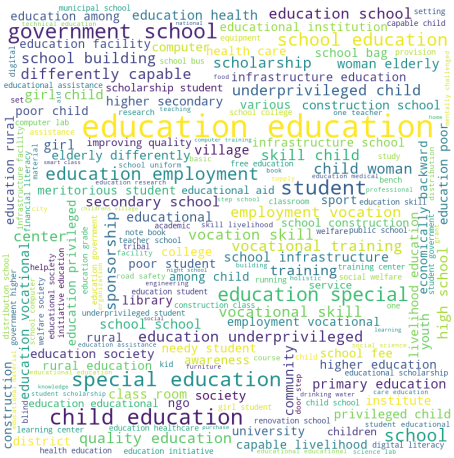
(a) Agroforestry



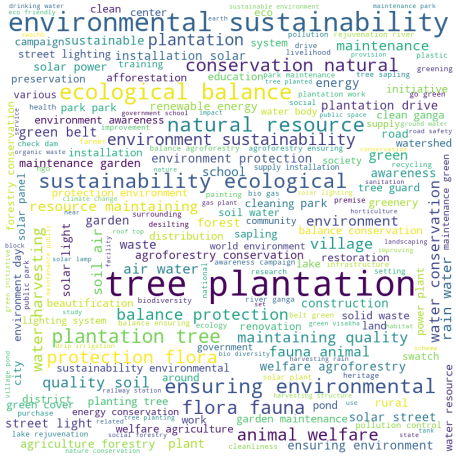
(b) Animal Welfare



(c) Education



(d) Environmental Sustainability



(e) Health



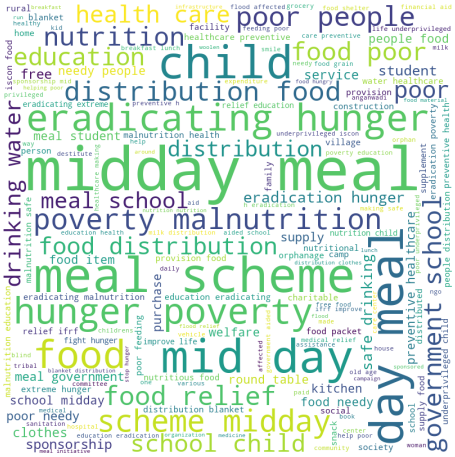
(f) Livelihood Enhancement



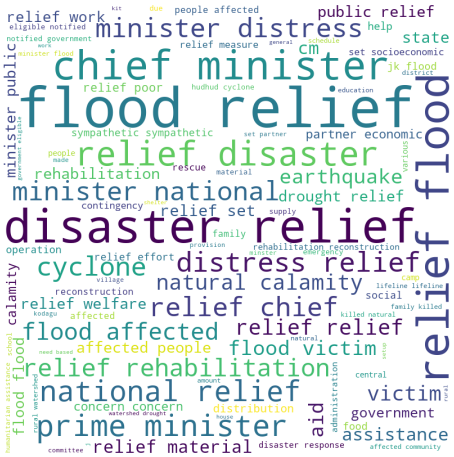
Notes: This figure shows word clouds of CSR project descriptions by topic.

Figure C.3: Project Descriptions by Topics: Word Clouds

(a) Hunger and Malnutrition



(b) Emergency Relief



(c) Infrastructure



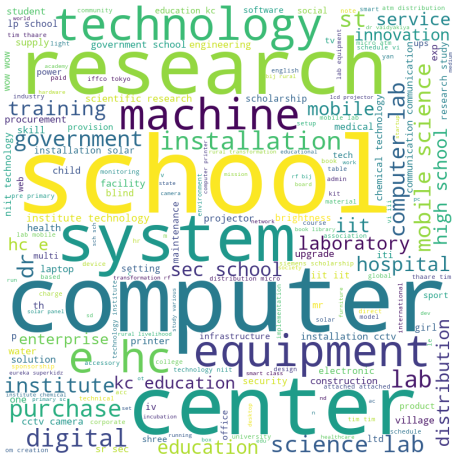
(d) Safe Drinking Water



(e) Sanitation



(f) Technology



Notes: This figure shows word clouds of CSR project descriptions by topic.

Figure C.4: Project Descriptions by Topics: Word Clouds

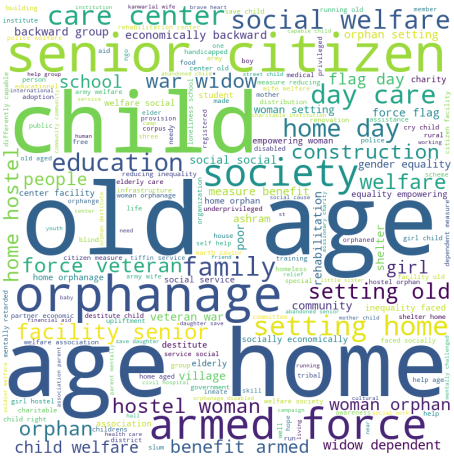
(a) Sports



(b) Vocational Skills



(c) Vulnerable Populations



(d) Women Empowerment



Notes: This figure shows word clouds of CSR project descriptions by topic.

Table C.1: Definition of Topics

Initial	Obs #	Final
Education	36,962	Education
Special education	1,761	Education
Health care	20,787	Health
Rural development projects	7,785	Infrastructure
Environmental sustainability	5,878	Environmental sustainability
Conservation of natural resources	728	Environmental sustainability
Clean Ganga Fund	47	Environmental sustainability
Animal welfare	1,334	Animal welfare
Agro forestry	175	Agroforestry
Poverty eradicating hunger malnutrition	4,716	Hunger and malnutrition
Livelihood enhancement projects	3,772	Livelihood enhancement
Vocational skills	3,189	Vocational skills
Sanitation	2,881	Sanitation
Swachh bharat kosh	701	Sanitation
Safe drinking water	2,000	Safe drinking water
Art and culture	1,903	Removed
Training to promote sports	2,076	Sports
Women empowerment	2,176	Women empowerment
Slum area development	304	Reallocated to closest topic
Gender equality	461	Vulnerable populations
Setting up homes and hostels for women	472	Vulnerable populations
Armed forces veterans war widows dependents	465	Vulnerable populations
Senior citizens welfare	911	Vulnerable populations
Setting up orphanage	472	Vulnerable populations
Socioeconomic inequalities	1,653	Vulnerable populations
Technology incubators	221	Technology
Other central government funds	1,421	Reallocated to closest topic
Prime Ministers National Relief Fund	631	Emergency relief

Notes: This table reports the mapping between initial topics and the 16 topics in the main sample. It also records the number of observations by initial topic.

Table C.2: Optimal K-Means Cosine Clusters

Topic	Cluster Theme	N of Projects	Silhouette Score	Bigram #1	Bigram #2	Bigram #3	Bigram #4	Bigram #5
Education	Special education	11698	0.091	child_education	special_education	education_special	education_employment	education_education
Education	Scholarships & support for meritorious students	8102	-0.069	government_school	student_education	meritorious_student	school_education	scholarship_student
Education	School construction & infrastructure	4423	0.075	government_school	school_building	school_infrastructure	school_construction	high_school
Health	Preventive health	5454	0.215	health_care	preventive_health	care_health	care_preventive	preventive_healthcare
Health	Medical treatment and patient care	4802	-0.113	cancer_patient	medical_treatment	health_care	palliative_care	poor_patient
Health	Medical infrastructure and equipment	3171	-0.026	medical_aid	hospital_medical	medical_equipment	medical_facility	medical_camp
Health	Mobile health camps	1364	0.104	health_camp	medical_camp	camp_health	health_checkup	blood_camp
Infrastructure	Construction (roads, walls, street light)	1904	0.016	road_construction	street_light	construction_road	boundary_wall	village_construction
Infrastructure	Construction (community centers)	1443	0.052	community_hall	construction_community	gram_panchayat	community_center	village_construction
Infrastructure	Rural infrastructure	966	0.057	rural_infrastructure	rural_dev	dev_rural	rural_transformation	infrastructure_rural
Environmental sustainability	Ecology/conservation projects	1914	0.185	environmental_sustainability	ecological_balance	sustainability_ecological	ensuring_environmental	natural_resource
Environmental sustainability	Tree plantation	1711	0.112	tree_plantation	plantation_tree	plantation_drive	maintenance_garden	planting_tree
Environmental sustainability	Green energy	869	0.038	solar_street	installation_solar	street_light	green_belt	solar_power
Environmental sustainability	Water harvesting & conservation	730	0.013	water_harvesting	rain_water	water_conservation	clean_ganga	harvesting_rain
Vocational skills	Vocational training	1761	0.117	vocational_training	vocational_skill	employment_vocational	skill_vocational	skill_training
Vocational skills	Skill acquisition	798	0.035	skill_training	education_skill	skill_center	skill_education	skill_skill
Vocational skills	Staff training	385	-0.093	driver_training	training_center	staff_salary	skill_training	salary_staff
Technology	Education technology	1443	-0.019	e_hc	kc_education	sec_school	hc_e	high_school
Technology	Mobile science labs & support to labs	609	0.135	mobile_science	science_lab	product_initiative	initiative_marketing	marketing_product
Technology	Computers, cameras and other equipment donations	596	-0.023	installation_cctv	cctv_camera	computer_lab	eureka_superkidz	installation_solar
Livelihood enhancement	Sustainable livelihood and education	1264	0.105	sustainable_livelihood	education_livelihood	livelihood_education	livelihood_livelihood	social_welfare
Livelihood enhancement	Disability inclusion	414	0.019	differently_capable	capable_livelihood	capable_child	person_disability	mentally_retarded
Livelihood enhancement	Economic development	405	-0.018	world_vision	financial_inclusion	quality_life	social_economic	partner_economic
Livelihood enhancement	Donations of vehicles	156	-0.054	tractor_traveller	selected_organization	product_selected	vehicle_product	traveller_vehicle
Sanitation	Sanitation infrastructure	1666	-0.042	sanitation_facility	drinking_water	school_sanitation	sanitation_hygiene	sanitary_napkin
Sanitation	Toilet construction	1454	0.237	construction_toilet	toilet_construction	toilet_block	school_construction	toilet_school
Sanitation	Cleanliness campaigns	254	0.150	swachh_campaign	campaign_swachh	swach_campaign	cleanliness_campaign	construction_toilet
Hunger and malnutrition	Eradicating hunger	957	0.073	eradicating_hunger	hunger_poverty	poverty_malnutrition	health_care	eradication_hunger
Hunger and malnutrition	Midday meal scheme	890	0.284	midday_meal	meal_scheme	day_meal	mid_day	scheme_midday
Hunger and malnutrition	Food distribution	858	0.049	distribution_food	food_poor	food_distribution	poor_people	food_relief
Hunger and malnutrition	Money gifts	182	-0.081	round_table	joy_giving	giving_week	table_round	specified_money
Safe drinking water	Water supply infrastructure (tanks, pumps, wells. . .)	1287	-0.136	drinking_water	hand_pump	water_supply	water_tank	tube_well
Safe drinking water	Water safety programs	939	0.212	drinking_water	safe_drinking	water_facility	water_safe	water_supply
Safe drinking water	Water purification (notably reverse osmosis equ...	644	0.053	reverse_osmosis	osmosis_plant	drinking_water	osmosis_water	water_purifier
Vulnerable populations	Hostels for old age, widows & orphans	2164	0.005	old_age	age_home	setting_home	home_hostel	hostel_woman
Vulnerable populations	Social welfare programs	748	0.159	social_welfare	welfare_social	social_service	child_welfare	backward_group
Vulnerable populations	Veterans support	644	-0.039	armed_force	war_widow	force_veteran	benefit_armed	flag_day
Vulnerable populations	Senior citizen care	321	0.109	senior_citizen	facility_senior	old_age	care_center	age_home
Emergency relief	Flood relief	618	0.338	flood_relief	relief_flood	flood_affected	flood_victim	relief_rehabilitation
Emergency relief	Disaster relief	511	0.077	disaster_relief	relief_disaster	natural_calamity	relief_rehabilitation	relief_relief
Emergency relief	Contribution to Prime Minister Relief Fund	316	0.397	chief_minister	relief_chief	distress_relief	minister_distress	prime_minister
Emergency relief	Cyclone relief	234	-0.081	cyclone_relief	cm_relief	relief_cyclone	disaster_relief	relief_material
Emergency relief	Contribution to eligible projects notified by g...	98	-0.123	notified_government	eligible_notified	government_eligible	national_relief	prime_minister
Sports	Sport competition organization	897	-0.159	sport_game	game_sport	sport_tournament	cricket_tournament	rural_sport
Sports	Support to olympic / nationally-recognized sports	505	-0.061	olympic_sport	rural_sport	recognized_sport	sport_olympic	nationally_recognized
Sports	Support to sport clubs	295	0.053	rotary_club	club_rotary	cricket_club	cricket_association	club_nigdi
Sports	Rural sport	233	0.335	rural_sport	sport_rural	sport_sport	sport_art	art_culture
Sports	Sponsorship	224	-0.126	sponsorship_sport	sport_event	sport_sponsorship	event_sponsorship	event_sport
Sports	Support to sport training, incl. rural	115	0.462	recognized_sport	nationally_recognized	sport_nationally	sport_training	rural_sport
Sports	Sport equipment	64	0.032	sport_equipment	sport_material	sport_item	equipment_sport	distribution_sport
Women empowerment	Education and young women empowerment	874	0.142	woman_empowerment	empowerment_woman	empowerment_education	education_woman	youth_empowerment
Women empowerment	Gender equality	509	0.070	empowering_woman	gender_equality	equality_empowering	woman_empowering	woman_gender
Women empowerment	Disadvantaged women and girls	258	0.044	girl_child	asare_woman	woman_asare	woman_child	girl_woman
Women empowerment	Textile industry training for women	153	0.045	sewing_machine	center_woman	training_center	woman_empowerment	tailoring_center
Women empowerment	Women associations, councils, and civic partici...	71	-0.095	state_council	council_woman	woman_state	mahila_dakshata	service_sedan
Agroforestry	Sustainable agriculture	232	0.029	organic_farming	agriculture_farm	national_agriculture	water_resource	improved_agriculture
Agroforestry	Farmer training	172	0.146	training_farmer	farmer_training	farmer_farmer	farmer_agriculture	farmer_education
Agroforestry	Agronomy	149	-0.073	guar_seed	seed_distribution	agronomic_practice	distribution_guar	extension_service
Animal welfare	Animal care	444	0.176	animal_welfare	animal_husbandry	animal_care	welfare_animal	care_animal
Animal welfare	Animal shelters	383	-0.148	lion_club	animal_welfare	hare_movement	stray_dog	welfare_animal
Animal welfare	Cow sheds	316	0.096	cow_shed	cow_protection	protection_cow	shed_cow	welfare_cow

Notes: This table summarizes the CSR project clusters obtained by k-means clustering. Details are provided in the text.

Table C.3: Mapping of CSR Topics to Government Spending and NGO Activity

Aggregated Topics	**CSR Topics**	**Government Spending Categories**	**NGO Activity**
Education	Education + sports	Education, sports, art & culture	Quality education
Vulnerable populations	Vulnerable populations + women empowerment	Social security & welfare	No poverty + gender equality + reduced inequality
Environmental sustainability	Environmental sustainability + agroforestry + animal welfare	Soil & water conservation + forestry & wild life + irrigation	Responsible consumption & production + climate action + life below water + life on land + (1/3) affordable & clean energy
Health	Health	Medical & public health + family welfare	Good health & well-being
Water & sanitation	Safe drinking water + sanitation	Water supply & sanitation	Clean water & sanitation
Industry & technology	Technology	Science, technology & environment	Industry, innovation & infrastructure + (1/3) affordable & clean energy
Infrastructure	Infrastructure	Rural development	(1/3) Affordable & clean energy
Vocational skills	Vocational skills + livelihood enhancement	Labor & labor welfare	Decent work & economic growth
Hunger & malnutrition	Hunger & malnutrition	Nutrition	Zero hunger
Emergency relief	Emergency relief	Relief on account of natural calamities	NA

Notes: This table reports the mapping of CSR topics to government spending and NGO activity.

Table C.4: Example of Text from NIC Handbook (Division 16)

Panel A: Text from Explanatory Notes

16 Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials. This division includes the manufacture of wood products, such as lumber, plywood, veneers, wood containers, wood flooring, wood trusses, and prefabricated wood buildings. The production processes include sawing, planning, shaping, laminating, and assembling of wood products starting from logs that are cut into bolts, or lumber that may then be cut further, or shaped by lathes or other shaping tools. The lumber or other transformed wood shapes may also be subsequently planed or smoothed, and assembled into finished products, such as wood containers. With the exception of sawmilling, this division is subdivided mainly based on the specific products manufactured.

161 Sawmilling and planning of wood.

162 Manufacture of products of wood, cork, straw and plaiting materials. This group includes the manufacture of products of wood, cork, straw or plaiting materials, including basic shapes as well as assembled products.

Panel B: Text from Further Industry Breakdown

Saw milling and planing of wood

Saw milling and planing of wood

Sawing and planing of wood

Manufacture of unassembled wooden flooring including parquet flooring

Manufacture of wooden railway sleepers

Activities related to saw milling and planing of wood n.e.c.

Manufacture of products of wood, cork, straw and plaiting materials

Manufacture of veneer sheets; manufacture of plywood, laminboard, particle board and other panels and board

Manufacture of ply wood and veneer sheets

Manufacture of particle board and fibreboard including densified wood

Manufacture of flush doors and other boards or panels

Manufacture of other plywood products n.e.c.

Manufacture of builders' carpentry and joinery

Manufacture of structural wooden goods [intended to be used primarily in the construction industry such as beams, rafters, roof struts, glue-laminated and metal connected, prefabricated wooden roof trusses, doors, windows, shutters and their frames, whether or not containing metal fittings, stairs, railings, wooden beadings and mouldings, shingles and shakes etc.]

Manufacture of prefabricated buildings, or elements thereof, predominantly of wood

Manufacture of builders' carpentry and joinery n.e.c.

Manufacture of wooden containers

Manufacture of wooden boxes, barrels, vats, tubs, packing cases etc.

Manufacture of plywood chests

Manufacture of market basketry, grain storage bins and similar products made of bamboo or reed

Manufacture of other wooden containers and products entirely or mainly of cane, rattan, bamboo, willow, fibre, leaves and grass n.e.c.

Manufacture of other products of wood; manufacture of articles of cork, straw and plaiting materials

Manufacture of wooden industrial goods

Manufacture of cork and cork products

Manufacture of wooden agricultural implements

Manufacture of various articles made of bamboo, cane and grass

Table C.5: Closest Development Topic by Industry

Industry Name	Topic	Closest 10 Token Pairs
IT and consultancy (62)	Vocational skills	[skill,expertise: 0.433], [training,testing: 0.388], [preservice,communication: 0.337], [preservice,technical: 0.333], [learning,communication: 0.331], [vocational,technical: 0.323], [training,technical: 0.322], [aptitude,expertise: 0.316], [carpentry,designing: 0.311], [skilled,expertise: 0.308]
Financial (64)	Livelihood enhancement	[welfare,pension: 0.409], [employment,income: 0.383], [employment,investment: 0.350], [farming,business: 0.342], [disadvantaged,income: 0.340], [sustainable,investment: 0.335], [employment,business: 0.335], [community,business: 0.325], [welfare,income: 0.317], [agriculture,banking: 0.314]
Petroleum products (19)	Safe drinking water	[electricity,gas: 0.508], [electricity,fuel: 0.455], [electricity,coal: 0.448], [electricity,propane: 0.403], [electricity,kerosene: 0.399], [groundwater,mineral: 0.384], [water,gas: 0.381], [sullage,liquid: 0.376], [reservoir,gas: 0.371], [water,liquid: 0.365]
Power and utilities (35)	Safe drinking water	[water,water: 1.000], [electricity,electricity: 1.000], [sewage,water: 0.662], [groundwater,water: 0.659], [potable,water: 0.625], [electricity,electric: 0.614], [electricity,energy: 0.603], [lake,water: 0.590], [electricity,power: 0.581], [river,water: 0.577]
Oil and gas extraction (6)	Safe drinking water	[electricity,gas: 0.508], [filtration,separator: 0.496], [electricity,coal: 0.448], [water,sand: 0.446], [bunding,separator: 0.423], [filtration,gasification: 0.413], [reservoir,hydrocarbon: 0.407], [bunding,sand: 0.402], [effluent,gasification: 0.400], [reservoir,drilling: 0.399]
Chemicals (20)	Safe drinking water	[checkdams,agarbatti: 0.458], [filtration,coating: 0.420], [filtration,deodorizing: 0.416], [potability,nitric: 0.399], [filtration,reagent: 0.395], [filtration,detergent: 0.391], [filtration,polyurethane: 0.387], [filtration,chemical: 0.382], [filtration,spray: 0.379], [sullage,agarbatti: 0.378]
Basic metals (24)	Safe drinking water	[groundwater,manganese: 0.448], [potability,manganese: 0.414], [effluent,manganese: 0.366], [potable,manganese: 0.362], [water,manganese: 0.330], [effluent,smelting: 0.313], [electricity,smelting: 0.309], [filtration,smelting: 0.308], [filtration,manganese: 0.308], [sewage,manganese: 0.305]
Pharmaceuticals (21)	Health	[healthcare,pharmaceutical: 0.540], [medical,unani: 0.424], [medical,pharmaceutical: 0.409], [medical,allopathic: 0.407], [doctor,homoeopathic: 0.400], [outpatient,allopathic: 0.396], [helath,homoeopathic: 0.372], [treatment,antibiotic: 0.367], [physician,allopathic: 0.354], [doctor,allopathic: 0.352]
Motor vehicles (29)	Safe drinking water	[electricity,electrical: 0.573], [sewer,electrical: 0.372], [sewage,electrical: 0.364], [sewerage,electrical: 0.361], [bunding,radiator: 0.340], [electricity,alternator: 0.340], [filtration,exhaust: 0.338], [irrigation,tractor: 0.338], [irrigation,electrical: 0.336], [drainage,electrical: 0.333]
Coal mining (5)	Safe drinking water	[bunding,belowground: 0.507], [electricity,fuel: 0.455], [electricity,coal: 0.448], [filtration,cleaning: 0.412], [checkdams,belowground: 0.410], [groundwater,mineral: 0.384], [sullage,belowground: 0.382], [sewerage,quarrying: 0.373], [checkdams,quarrying: 0.372], [drainage,grading: 0.369]
Wholesale trade (46)	Agroforestry	[agricultural,agricultural: 1.000], [dairy,dairy: 1.000], [agriculture,agricultural: 0.873], [farming,agricultural: 0.718], [ag,agricultural: 0.711], [livestock,agricultural: 0.645], [farm,agricultural: 0.638], [dairying,dairy: 0.631], [horticulture,agricultural: 0.625], [agrarian,agricultural: 0.619]
Civil engineering (42)	Infrastructure	[construction,construction: 1.000], [highway,highway: 1.000], [road,road: 1.000], [sewerage,sewerage: 1.000], [sewerage,sewage: 0.786], [sewerage,sewer: 0.725], [highway,road: 0.697], [drainage,sewer: 0.654], [drainage,sewerage: 0.631], [drainage,sewage: 0.613]
Other transport (30)	Safe drinking water	[lake,pontoon: 0.477], [river,pontoon: 0.458], [lake,boat: 0.445], [river,boat: 0.442], [lake,watercraft: 0.398], [creek,pontoon: 0.391], [river,barge: 0.382], [sullage,transport: 0.372], [sullage,pontoon: 0.368], [sewerage,transport: 0.367]
Industrial machinery (28)	Safe drinking water	[electricity,electrical: 0.573], [filtration,compressor: 0.467], [electricity,turbine: 0.424], [filtration,coating: 0.420], [filtration,electroplating: 0.417], [filtration,hydraulic: 0.409], [electricity,furnace: 0.398], [bunding,conveyor: 0.393], [filtration,turbine: 0.391], [bunding,hydraulic: 0.390]
Non-metallic minerals (23)	Safe drinking water	[bunding,rockwool: 0.493], [checkdams,rockwool: 0.484], [sullage,rockwool: 0.468], [bunding,aluminous: 0.452], [irrigation,rockwool: 0.417], [filtration,nonwoven: 0.402], [checkdams,aluminous: 0.393], [filtration,rockwool: 0.391], [water,rockwool: 0.388], [potable,calcined: 0.382]
Data and info (63)	Technology	[psn,internet: 0.355], [psn,web: 0.346], [edu,internet: 0.344], [brail,library: 0.334], [edu,web: 0.319], [edu,portal: 0.302], [pwd,database: 0.300], [pwd,searchable: 0.298], [pwd,portal: 0.293], [pwd,web: 0.293]
Logistics and warehousing (52)	Safe drinking water	[sullage,lighterage: 0.500], [sewerage,infrastructure: 0.450], [drainage,maintenance: 0.408], [checkdams,lighterage: 0.392], [sewerage,railway: 0.389], [bunding,lighterage: 0.380], [sullage,transport: 0.372], [sewerage,transport: 0.367], [drainage,infrastructure: 0.364], [electricity,infrastructure: 0.359]
Food and beverages (10)	Agroforestry	[dairy,milk: 0.732], [dairy,meat: 0.549], [dairy,vegetable: 0.480], [pisciculture,oilcake: 0.472], [farmer,vegetable: 0.468], [livestock,meat: 0.466], [grower,potato: 0.463], [dairy,potato: 0.463], [agricultural,oilcake: 0.452], [farmer,potato: 0.450]
Tobacco (12)	Agroforestry	[agricultural,agricultural: 1.000], [agriculture,agricultural: 0.873], [farming,agricultural: 0.718], [ag,agricultural: 0.711], [livestock,agricultural: 0.645], [farm,agricultural: 0.638], [horticulture,agricultural: 0.625], [agrarian,agricultural: 0.619], [dairy,agricultural: 0.603], [irrigation,agricultural: 0.570]
Rubber and plastics (22)	Safe drinking water	[water,hose: 0.447], [reservoir,pipe: 0.425], [sewer,pipe: 0.415], [bunding,conveyor: 0.393], [sewage,pipe: 0.391], [sullage,carboy: 0.381], [filtration,conveyor: 0.378], [creek,tub: 0.372], [water,tub: 0.370], [bunding,hose: 0.362]

Notes: This table presents, for each industry, the closest social topic based on the semantic proximity score (see equation (9)). Results are restricted to the top 20 industries. For each pair, we report the 10 most similar token pairs (from the 20 tokens closest to each group's centroid in the word embedding space), along with their cosine similarity scores.

Table C.6: Furthest Development Topic by Industry

Industry Name	Topic	Closest 10 Token Pairs
IT and consultancy (62)	Emergency relief	[assistance,expertise: 0.334], [assistance,service: 0.300], [emergency,communication: 0.254], [reconstruction,maintenance: 0.250], [assistance,technical: 0.249], [reconstruction,installation: 0.243], [assistance,communication: 0.221], [emergency,maintenance: 0.214], [assistance,facility: 0.198], [emergency,service: 0.197]
Financial (64)	Technology	[ict,intermediation: 0.222], [motorola,intermediation: 0.205], [edu,intermediation: 0.202], [tecnology,intermediation: 0.202], [dx,intermediation: 0.193], [pmc,intermediation: 0.188], [pwd,intermediation: 0.183], [psn,mortgage: 0.183], [engg,banking: 0.182], [pwd,corporation: 0.179]
Petroleum products (19)	Sports	[kabbadi,kerosene: 0.216], [hockey,bitumen: 0.185], [cycling,refining: 0.176], [football,refinery: 0.152], [cycling,bitumen: 0.149], [gymnastics,refining: 0.147], [judo,butane: 0.146], [kabbadi,briquette: 0.145], [hockey,gas: 0.145], [judo,kerosene: 0.143]
Power and utilities (35)	Sports	[cycling,diesel: 0.190], [cycling,energy: 0.179], [athletics,energy: 0.167], [sport,diesel: 0.164], [sport,generation: 0.155], [cycling,thermal: 0.153], [cycling,solar: 0.150], [tennis,water: 0.146], [hockey,gas: 0.145], [triathlon,diesel: 0.140]
Oil and gas extractions (6)	Sports	[hockey,shale: 0.183], [cycling,liquefaction: 0.172], [tennis,sand: 0.168], [cycling,mining: 0.167], [boxing,mining: 0.160], [triathlon,sand: 0.160], [football,shale: 0.158], [rugby,shale: 0.156], [cycling,pyrolysis: 0.154], [sporting,mining: 0.153]
Chemicals (20)	Emergency relief	[catastrophe,chemical: 0.204], [disaster,chemical: 0.178], [diaster,chemical: 0.149], [aid,dye: 0.146], [flood,chemical: 0.139], [emergency,chemical: 0.137], [quake,chemical: 0.129], [disater,chemical: 0.122], [catastrophe,reagent: 0.118], [reconstruction,chemical: 0.117]
Basic metals (24)	Emergency relief	[cyclone,alumina: 0.229], [cyclone,tin: 0.209], [diaster,alumina: 0.203], [catastrophe,alumina: 0.184], [flooding,alumina: 0.183], [flooding,manganese: 0.179], [reconstruction,alumina: 0.173], [reconstruction,metallurgic: 0.166], [earthquake,copper: 0.161], [reconstruction,steel: 0.160]
Pharmaceuticals (21)	Emergency relief	[catastrophe,chemical: 0.204], [emergency,allopathic: 0.190], [disaster,chemical: 0.178], [relief,homoeopathic: 0.163], [humanitarian,pharmaceutical: 0.160], [emergency,unani: 0.153], [cyclone,unani: 0.152], [diaster,chemical: 0.149], [diaster,suger: 0.145], [tsunami,unani: 0.144]
Motor vehicles (29)	Emergency relief	[emergency,electrical: 0.240], [earthquake,electrical: 0.224], [flooding,electrical: 0.209], [cyclone,lorry: 0.201], [emergency,brake: 0.195], [emergency,passenger: 0.184], [devastation,trailer: 0.182], [quake,electrical: 0.179], [tsunami,airbags: 0.177], [emergency,ignition: 0.167]
Coal miningg (5)	Sports	[cycling,transportation: 0.305], [athletics,transportation: 0.209], [rugby,opencast: 0.205], [athlete,hard: 0.188], [sport,hard: 0.186], [tournament,operation: 0.179], [cycling,agglomeration: 0.171], [cycling,mining: 0.167], [boxing,mining: 0.160], [football,hard: 0.160]
Wholesale trade (46)	Emergency relief	[aid,food: 0.331], [flood,agricultural: 0.255], [assistance,food: 0.254], [humanitarian,agricultural: 0.245], [humanitarian,food: 0.241], [displaced,food: 0.240], [emergency,electrical: 0.240], [relief,food: 0.238], [earthquake,electrical: 0.224], [flooding,agricultural: 0.222]
Civil engineering (42)	Technology	[engg,engineering: 0.367], [cnc,engineering: 0.285], [tecnology,engineering: 0.277], [cnc,repair: 0.263], [pwd,railway: 0.243], [cnc,industrial: 0.235], [ict,industrial: 0.225], [engg,industrial: 0.221], [cnc,maintenance: 0.215], [hyundai,railway: 0.211]
Other transport (30)	Emergency relief	[rescue,helicopter: 0.437], [emergency,helicopter: 0.334], [rescue,hovercraft: 0.301], [humanitarian,cargo: 0.269], [rescue,boat: 0.266], [rescue,crane: 0.263], [cyclone,vessel: 0.234], [rescue,pontoon: 0.231], [diaster,barge: 0.231], [emergency,transport: 0.229]
Industrial machinery (28)	Emergency relief	[emergency,electrical: 0.240], [emergency,equipment: 0.238], [assistance,equipment: 0.232], [earthquake,electrical: 0.224], [flooding,electrical: 0.209], [reconstruction,machinery: 0.198], [emergency,hydraulic: 0.195], [aid,equipment: 0.190], [quake,electrical: 0.179], [rescue,hydraulic: 0.177]
Non-metallic minerals (23)	Emergency relief	[flooding,rockwool: 0.198], [reconstruction,cement: 0.192], [flood,rockwool: 0.190], [cyclone,rockwool: 0.166], [earthquake,tile: 0.165], [quake,earthenware: 0.162], [cyclone,nonwoven: 0.158], [earthquake,earthenware: 0.158], [catastrophe,graphite: 0.154], [quake,tile: 0.150]
Data and info (63)	Emergency relief	[reconstruction,infrastructure: 0.383], [rescue,search: 0.367], [assistance,service: 0.300], [assistance,infrastructure: 0.240], [flooding,infrastructure: 0.235], [disaster,infrastructure: 0.233], [flood,infrastructure: 0.223], [assistance,agency: 0.216], [diaster,infrastructure: 0.212], [disaster,agency: 0.209]
Logistics and warehousing (52)	Sports	[cycling,transportation: 0.305], [cycling,transport: 0.284], [cycling,railway: 0.250], [paralympic,railway: 0.210], [athletics,transportation: 0.209], [tournament,facility: 0.196], [paralympic,airfield: 0.181], [kabbadi,railway: 0.179], [badminton,berthing: 0.177], [cycling,freight: 0.170]
Food and beverages (10)	Sports	[kabbadi,papads: 0.369], [kabbadi,sweetmeat: 0.364], [kabbadi,curd: 0.322], [kabbadi,oilcake: 0.287], [kabbadi,rice: 0.249], [badminton,tapioca: 0.248], [kabbadi,khandsari: 0.242], [badminton,papads: 0.217], [kabbadi,iodized: 0.211], [kabbadi,tapioca: 0.207]
Tobacco (12)	Emergency relief	[flood,agricultural: 0.255], [humanitarian,agricultural: 0.245], [flooding,agricultural: 0.222], [aid,agricultural: 0.215], [assistance,agricultural: 0.181], [reconstruction,agricultural: 0.178], [emergency,agricultural: 0.171], [cyclone,bidi: 0.162], [assistance,suitable: 0.162], [displaced,agricultural: 0.153]
Rubber and plastics (22)	Emergency relief	[emergency,hose: 0.183], [rescue,hose: 0.179], [disater,carboy: 0.170], [relief,rubber: 0.136], [catastrophe,material: 0.136], [diaster,container: 0.135], [assistance,material: 0.134], [aid,material: 0.127], [flood,material: 0.125], [emergency,tube: 0.122]

Notes: This table presents, for each industry, the furthest social topic based on the semantic proximity score (see equation (9)). Results are restricted to the top 20 industries. For each pair, we report the 10 most similar token pairs (from the 20 tokens closest to each group's centroid in the word embedding space), along with their cosine similarity scores.

Table C.7: Closest Industry by Development Topic

Topic	Industry Name	Closest 10 Token Pairs
Environmental sustainability	Forestry and logging (2)	[forestry,forestry: 1.000], [forest,forest: 1.000], [forestry,silviculture: 0.690], [forestry,forest: 0.676], [forestry,timber: 0.647], [reforestation,replanting: 0.646], [forest,timber: 0.608], [agriculture,forestry: 0.604], [agricultural,forestry: 0.586], [afforestation,replanting: 0.545]
Education	Education (85)	[education,education: 1.000], [educational,educational: 1.000], [school,school: 1.000], [vocational,vocational: 1.000], [literacy,literacy: 1.000], [academic,academic: 1.000], [classroom,classroom: 1.000], [student,student: 1.000], [eduction,education: 0.798], [elementary,school: 0.787]
Infrastructure	Civil engineering (42)	[construction,construction: 1.000], [highway,highway: 1.000], [road,road: 1.000], [sewerage,sewerage: 1.000], [sewerage,sewage: 0.786], [sewerage,sewer: 0.725], [highway,road: 0.697], [drainage,sewer: 0.654], [drainage,sewerage: 0.631], [drainage,sewage: 0.613]
Hunger and malnutrition	Food and beverage services (56)	[food,food: 1.000], [meal,meal: 1.000], [eat,meal: 0.559], [nutritious,food: 0.547], [food,meal: 0.543], [nutrition,food: 0.520], [food,beverage: 0.510], [eat,drink: 0.507], [nutritious,meal: 0.504], [eat,food: 0.504]
Vocational skills	Education (85)	[vocational,vocational: 1.000], [training,training: 1.000], [education,education: 1.000], [educational,educational: 1.000], [education,educational: 0.678], [vocational,education: 0.611], [mentoring,tutoring: 0.594], [education,literacy: 0.594], [teaching,education: 0.571], [teaching,classroom: 0.563]
Technology	Computer and electronics (26)	[tecnology,technology: 0.466], [lenovo,modem: 0.453], [lcd,modem: 0.438], [tft,optical: 0.432], [pda,device: 0.424], [lenovo,computer: 0.421], [lcd,diode: 0.413], [lcd,optical: 0.404], [pda,modem: 0.397], [tecnology,electromedical: 0.392]
Health	Health care (86)	[health,health: 1.000], [medical,medical: 1.000], [hospital,hospital: 1.000], [patient,patient: 1.000], [dental,dental: 1.000], [clinic,clinic: 1.000], [outpatient,outpatient: 1.000], [treatment,treatment: 1.000], [doctor,doctor: 1.000], [outpatient,inpatient: 0.820]
Sanitation	Specialized construction (43)	[construction,construction: 1.000], [cleaning,cleaning: 1.000], [construction,excavation: 0.571], [construction,demolition: 0.564], [construction,building: 0.533], [cleaning,sanding: 0.479], [sewer,plumbing: 0.454], [drainage,plumbing: 0.448], [construction,installation: 0.436], [cleanliness,cleaning: 0.434]
Sports	Sports and recreation (93)	[sport,sport: 1.000], [sporting,sporting: 1.000], [athlete,athlete: 1.000], [sportsman,sportsman: 1.000], [athlete,sportsman: 0.583], [athletics,athlete: 0.565], [sport,racing: 0.562], [sport,sporting: 0.560], [athletics,sport: 0.552], [boxing,sport: 0.533]
Vulnerable populations	Public administration and defence (84)	[education,education: 1.000], [welfare,health: 0.506], [education,health: 0.473], [welfare,social: 0.473], [welfare,education: 0.468], [education,agriculture: 0.433], [family,community: 0.426], [education,social: 0.365], [welfare,agriculture: 0.361], [welfare,government: 0.357]
Safe drinking water	Water collection, treatment and supply (36)	[water,water: 1.000], [irrigation,irrigation: 1.000], [river,canal: 0.722], [sewage,water: 0.662], [groundwater,water: 0.659], [creek,canal: 0.627], [potable,water: 0.625], [lake,canal: 0.610], [drainage,irrigation: 0.600], [lake,water: 0.590]
Women empowerment	Education (85)	[education,education: 1.000], [literacy,literacy: 1.000], [education,educational: 0.678], [education,vocational: 0.611], [education,literacy: 0.594], [literacy,educational: 0.534], [education,academic: 0.519], [education,school: 0.506], [education,diploma: 0.481], [education,instruction: 0.474]
Animal welfare	Crops and animals (1)	[livestock,agricultural: 0.645], [livestock,farming: 0.545], [husbandry,farming: 0.494], [cattle,wheat: 0.479], [cattle,agricultural: 0.462], [livestock,wheat: 0.438], [cattle,farming: 0.422], [livestock,crop: 0.419], [pig,potato: 0.417], [husbandry,agricultural: 0.415]
Emergency relief	Public administration and defence (84)	[reconstruction,infrastructure: 0.383], [humanitarian,military: 0.334], [assistance,service: 0.300], [humanitarian,social: 0.299], [assistance,education: 0.293], [aid,education: 0.291], [aid,government: 0.284], [humanitarian,economic: 0.262], [humanitarian,security: 0.256], [emergency,health: 0.256]
Livelihood enhancement	Public administration and defence (84)	[education,education: 1.000], [community,community: 1.000], [social,social: 1.000], [agriculture,agriculture: 1.000], [farming,agriculture: 0.711], [educational,education: 0.678], [vocational,education: 0.611], [literacy,education: 0.594], [welfare,health: 0.506], [rural,agriculture: 0.486]
Agroforestry	Crops and animals (1)	[farming,farming: 1.000], [agricultural,agricultural: 1.000], [agriculture,agricultural: 0.873], [farm,farming: 0.731], [farming,agricultural: 0.718], [ag,agricultural: 0.711], [agriculture,farming: 0.711], [farmer,farming: 0.687], [dairying,farming: 0.675], [livestock,agricultural: 0.645]

Notes: This table presents, for each topic, the closest industry based on the semantic proximity score (see equation (9)). For each pair, we report the 10 most similar token pairs (from the 20 tokens closest to each group's centroid in the word embedding space), along with their cosine similarity scores.

Table C.8: Furthest Industry by Development Topic

Topic	Industry Name	Closest 10 Token Pairs
Environmental sustainability	Postal and courier (53)	[wetland,parcel: 0.312], [agriculture,transport: 0.308], [forestry,transport: 0.284], [forestry,local: 0.264], [sustainable,transport: 0.263], [environmental,local: 0.260], [agricultural,transport: 0.253], [conservation,local: 0.250], [forestry,postal: 0.238], [water,transport: 0.227]
Education	Metal ores mining (7)	[lifeskills,beneficiating: 0.300], [nonformal,beneficiating: 0.247], [vocational,beneficiation: 0.226], [curriculum,beneficiating: 0.226], [vocational,beneficiating: 0.226], [eduction,beneficiating: 0.222], [lifeskills,beneficiation: 0.210], [eduction,beneficiation: 0.198], [elementary,molybdenum: 0.194], [education,beneficiating: 0.191]
Infrastructure	Publishing (58)	[infrastructure,software: 0.293], [sarai,encyclopaedia: 0.248], [chaupal,encyclopaedia: 0.237], [locality,periodical: 0.229], [rural,internet: 0.222], [sarai,periodical: 0.219], [building,publishing: 0.219], [kutcha,encyclopaedia: 0.207], [locality,encyclopaedia: 0.206], [infrastructure,directory: 0.206]
Hunger and malnutrition	Metal ores mining (7)	[nutritional,mineral: 0.276], [food,mineral: 0.263], [anganwadis,bauxite: 0.249], [malnourished,manganese: 0.237], [anganwadi,bauxite: 0.235], [malnutrition,manganese: 0.234], [anganwadis,beneficiation: 0.222], [nutritious,mineral: 0.222], [nutrition,mineral: 0.222], [sanitation,bauxite: 0.213]
Vocational skills	Metal ores mining (7)	[skilling,beneficiation: 0.402], [upskilling,beneficiation: 0.362], [enterpreneurship,beneficiating: 0.338], [reskilling,beneficiation: 0.328], [reskilling,beneficiating: 0.312], [upskilling,beneficiating: 0.303], [enterpreneurship,beneficiation: 0.302], [employability,beneficiation: 0.284], [vocationally,beneficiating: 0.283], [skilling,beneficiating: 0.281]
Technology	Metal ores mining (7)	[amu,ilmenite: 0.323], [cnc,beneficiating: 0.315], [tecnology,beneficiating: 0.314], [amu,niobium: 0.289], [ict,beneficiating: 0.283], [amu,tungsten: 0.282], [amu,cobalt: 0.278], [cnc,tungsten: 0.261], [tft,ilmenite: 0.251], [tft,tantalum: 0.250]
Health	Metal ores mining (7)	[dental,manganese: 0.201], [education,beneficiating: 0.191], [healthcare,mining: 0.189], [health,manganese: 0.185], [education,beneficiation: 0.185], [pediatric,antimony: 0.183], [treatment,beneficiation: 0.179], [obstetric,ore: 0.179], [treatment,cobalt: 0.178], [outpatient,molybdenum: 0.176]
Sanitation	Metal ores mining (7)	[construction,mining: 0.350], [water,mineral: 0.350], [water,manganese: 0.330], [sewage,manganese: 0.305], [sewerage,manganese: 0.291], [drainage,manganese: 0.275], [sewerage,bauxite: 0.273], [anganwadis,bauxite: 0.249], [construction,beneficiating: 0.245], [water,uranium: 0.243]
Sports	Metal ores mining (7)	[paralympic,gold: 0.262], [judo,gold: 0.261], [kabbadi,gold: 0.232], [gymnastics,gold: 0.229], [badminton,gold: 0.229], [sporting,gold: 0.195], [football,nickel: 0.179], [rugby,beneficiation: 0.177], [hockey,gold: 0.170], [boxing,gold: 0.169]
Vulnerable populations	Metal ores mining (7)	[destitute,bauxite: 0.199], [education,beneficiating: 0.191], [education,beneficiation: 0.185], [orphange,beneficiating: 0.163], [welfare,beneficiation: 0.151], [orphange,bauxite: 0.149], [destitute,beneficiation: 0.145], [orphaned,beneficiation: 0.139], [orphange,rutile: 0.138], [child,mineral: 0.133]
Safe drinking water	Publishing (58)	[checkdams,atlas: 0.217], [potability,content: 0.211], [electricity,internet: 0.202], [checkdams,periodical: 0.195], [checkdams,encyclopaedia: 0.190], [lake,atlas: 0.168], [potability,software: 0.167], [drinking,content: 0.163], [electricity,directory: 0.163], [electricity,newspaper: 0.159]
Women empowerment	Metal ores mining (7)	[empowerment,beneficiation: 0.445], [upliftment,beneficiation: 0.439], [upliftment,beneficiating: 0.347], [empowerment,beneficiating: 0.322], [upliftment,bauxite: 0.281], [upliftment,rutile: 0.275], [disadvantaged,beneficiation: 0.264], [entrepreneurship,beneficiation: 0.254], [equality,beneficiation: 0.245], [upliftment,ilmenite: 0.240]
Animal welfare	Metal ores mining (7)	[gaushalas,beneficiating: 0.308], [livestock,mineral: 0.268], [gaushala,beneficiation: 0.261], [cattle,ore: 0.260], [gaushalas,beneficiation: 0.252], [cattle,uranium: 0.249], [husbandry,beneficiating: 0.239], [gaushalas,bauxite: 0.236], [husbandry,beneficiation: 0.236], [gaushala,beneficiating: 0.227]
Emergency relief	Basic metals (24)	[cyclone,alumina: 0.229], [cyclone,tin: 0.209], [diaster,alumina: 0.203], [catastrophe,alumina: 0.184], [flooding,alumina: 0.183], [flooding,manganese: 0.179], [reconstruction,alumina: 0.173], [reconstruction,metallurgic: 0.166], [earthquake,copper: 0.161], [reconstruction,steel: 0.160]
Livelihood enhancement	Metal ores mining (7)	[upliftment,beneficiation: 0.439], [farming,mining: 0.384], [upliftment,beneficiating: 0.347], [agriculture,mining: 0.331], [agriculture,mineral: 0.296], [agriculture,beneficiating: 0.286], [employability,beneficiation: 0.284], [upliftment,bauxite: 0.281], [upliftment,rutile: 0.275], [sustainable,beneficiation: 0.265]
Agroforestry	Postal and courier (53)	[agriculture,transport: 0.308], [agricultural,transport: 0.253], [orchard,parcel: 0.235], [irrigation,transport: 0.232], [pisciculture,local: 0.229], [farming,transport: 0.227], [pisciculture,various: 0.226], [orchard,pickup: 0.224], [agricultural,local: 0.220], [livestock,transport: 0.219]

Notes: This table presents, for each topic, the furthest industry based on the semantic proximity score (see equation (9)). For each pair, we report the 10 most similar token pairs (from the 20 tokens closest to each group's centroid in the word embedding space), along with their cosine similarity scores.

D Implementation of the CSR Mandate

D.1 Additional Details on the CSR Mandate

This appendix provides more detail on the implementation of the CSR mandate. The CSR mandate follows from Section 135 of the Companies Act, 2013. The mandate stipulates that firms above a certain size must form a CSR committee, formulate a CSR policy, and spend at least 2% of their average profits over the last three years on CSR.

Schedule VII of the Act lists the activities that qualify for CSR expenditures. This list of activities corresponds to the CSR topics listed in Table C.1. It is further clarified that CSR activities should be undertaken by the companies in "project/programme mode". One-off events such as awards, charitable contribution, sponsorships, etc. would not be qualified as part of CSR expenditure.

CSR activities do not include the activities undertaken in pursuance of normal course of business of a company. The CSR projects that benefit only the employees of the company do not satisfy the mandate. Contributions to a political party cannot be considered as CSR activity. Contribution in kind is not a CSR expenditure. Finally, sponsorship activities meant to derive marketing benefits do not qualify.

The amount spent on CSR cannot be claimed as business expenditure and hence is not tax-deductible. While no specific tax exemption has been extended to CSR, spending on several activities (e.g., rural development projects, contribution to Prime Minister National Relief Fund) already enjoy exemptions under different sections of Tax Act, 1961, subject to fulfillment of specified conditions.

Firms can implement their CSR activities directly, or via implementing agencies. Implementing agencies can be an entity established by the company itself, a government entity, or any other entity with a track record of undertaking similar activities. The mandate does not detail what extent of outsourcing implies that firms should declare their CSR projects as implemented via a third-party.

D.2 Change in CSR Expenditures After the Mandate was Implemented

To estimate the effect of the CSR mandate on CSR expenditures, we compare the evolution of CSR expenditures, as reported in the Prowess data, among liable and non-liable firms before and after the mandate's implementation. We define firms as liable under the act if they have profits above INR 50 million, income above INR 10 billion, or net worth above INR 5 billion in any of the three preceding financial years, as observed in the Prowess Data. All other firms present in this data constitute the non-liable group. We estimate the following difference-in-differences specification:

$$\left(\frac{\text{CSR}_{fy}}{\overline{\text{Profit}}^{3y}_{fy}}\right) = \beta \text{Post}_y \times \text{Treated}_{fy} + \gamma_y + \gamma_f + \gamma_g + \varepsilon_{fy} \tag{D.1}$$

where f indexes the firm and y the year, the outcome variable is CSR spending scaled by average profits in the preceding three years (y-3, y-2, and y-1). Treated_{fy} is equal to one if the firm is liable under the CSR regulation in year y, Post_y is a dummy equal to one every year from 2015 onwards, γ_y are year fixed effects and γ_f are firm fixed effects. γ_g are group fixed effects, which indicate the liability status of a given firm in a given year. Standard errors are clustered at the firm level.

We define CSR spending in Prowess as the sum of two variables: social and community expenses and donations. Social and community expenses are expenses incurred by firms for the benefit of society in general. Donations include donations for social causes, religious purposes, or political parties. Both social and community expenses, as well as donations, are reported in the schedules or notes to financial statements of the annual reports under the break-up of expenses or under welfare expenses. In 2015, in alignment with the introduction of the policy, Prowess began to collect additionally explicit CSR data. Since this variable was not available before, we do not utilize it to estimate the policy impact, which requires pre- and post-policy data.

Table D.1, Panel A, describes the results. In column 1, we observe that the share of CSR spending over average profits increases by 1.1% for liable firms relative to non-liable firms after the mandate is implemented. This effect remains stable if we replace the year fixed effects with year×industry×state fixed effects in column 2. Note that the effect is not 2% because non-liable firms also spend on CSR. The CSR spending

of non-liable firms before and after the policy is 0.8% on average, while that of liable firms rises from 0.7% to 2.0% on average (see also Figure 1(b) for raw trends).

The key identification assumption is parallel trends. This ensures that pre-existing trends between liable and non-liable firms do not influence the estimate. While this assumption is untestable, Figure D.1 documents parallel pre-trends in an event study analysis.

D.3 Are Existing Expenditures Incorrectly Relabeled as 'CSR'?

We next discuss the possibility that firms incorrectly labeled non-CSR expenditures as CSR after the mandate was implemented to reach the mandate's required level of expenditures. Note that such relabeling is not a concern for our study unless relabeled expenditures are systematically assigned to some topics or location in a way that biases our results. During our period, the government relied on provisions such as mandatory disclosures, board and CSR committee accountability with an independent director, and audit of accounts to ensure correct accounting of CSR expenditures. In 2019, the government also introduced fines for failing to meet the mandate's requirements.

Firms have two possible options to manipulate their level of spending. First, firms might manipulate their accounting variables to change their treatment status. We initially investigate which threshold is most binding: income, net worth, or profit (Figure D.2(a) to D.2(c)). We observe that for income and net worth, only 6% of firm-years have values higher than the respective threshold. In contrast, 32% of firm-years have values higher than the respective threshold for profits. This suggests that profit is the binding threshold for the majority of firms. Figure D.2(d) depicts the distribution of profits before the policy, between 2007 and 2014. Figure D.2(e) depicts it after the policy, between 2015 and 2019. Zooming in on the part of the distribution around the threshold, visual inspection shows only minor bunching below the threshold. This evidence suggests that only a small minority of firms manipulated their liability status. To address this dimension of manipulation, we further test a version of the difference-in-difference specification in equation (D.1) in columns 3 and 4 of Table D.1, but instrument the treatment status Treated_{fy} with a pre-policy variable Treated_{f}. The latter is an indicator equal to one if the firm is liable under the CSR regulation in the year 2014, that is, if either profits, income, or net worth

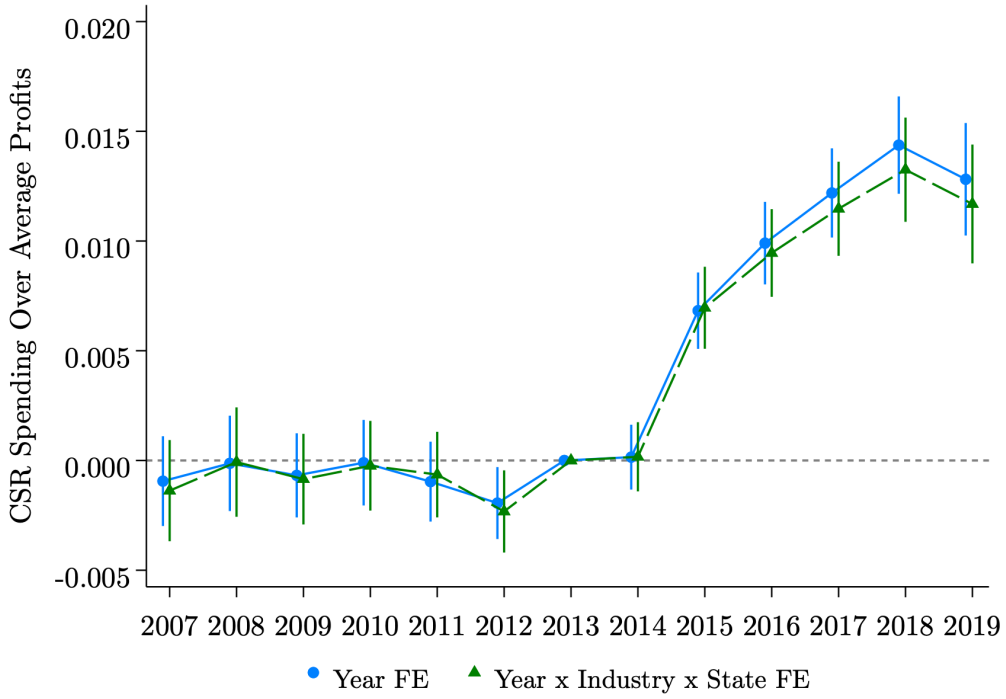
are above their respective thresholds in any of the three preceding financial years (2011–2013). Results are quantitatively similar under this specification.

Second, firms might manipulate by wrongly relabeling some of their expenditures as CSR to increase their total CSR spending. Section D.1 describes which expenses are excluded from CSR spending by the policy. To test whether firms are relabeling, we run equation (D.1), using as dependent variables expenditures reported in Prowess that firms could plausibly relabel. In Table D.2, the first column is our CSR variable, the sum of social expenses and donations; the second and third columns present results for each of these categories in turn: we see some substitution away from donations, which may not all have been expenditures that would count as CSR in the 2013 law. Our definition of CSR as the sum of social expenses and donations is immune to concerns related to reallocation between these two categories. We see little effect on expenditures on the (work) environment, employee welfare or training, social amenities, or advertisement. There appears to be a decrease in marketing expenses, which could indicate relabeling, but might also be consistent with firms spending less on this dimension because overall the policy has a negative effect on their firm outcomes. Overall, given the inherent difficulty of testing for relabeling, these findings should be viewed as suggestive.

D.4 Does CSR Crowd-Out Government Expenditures?

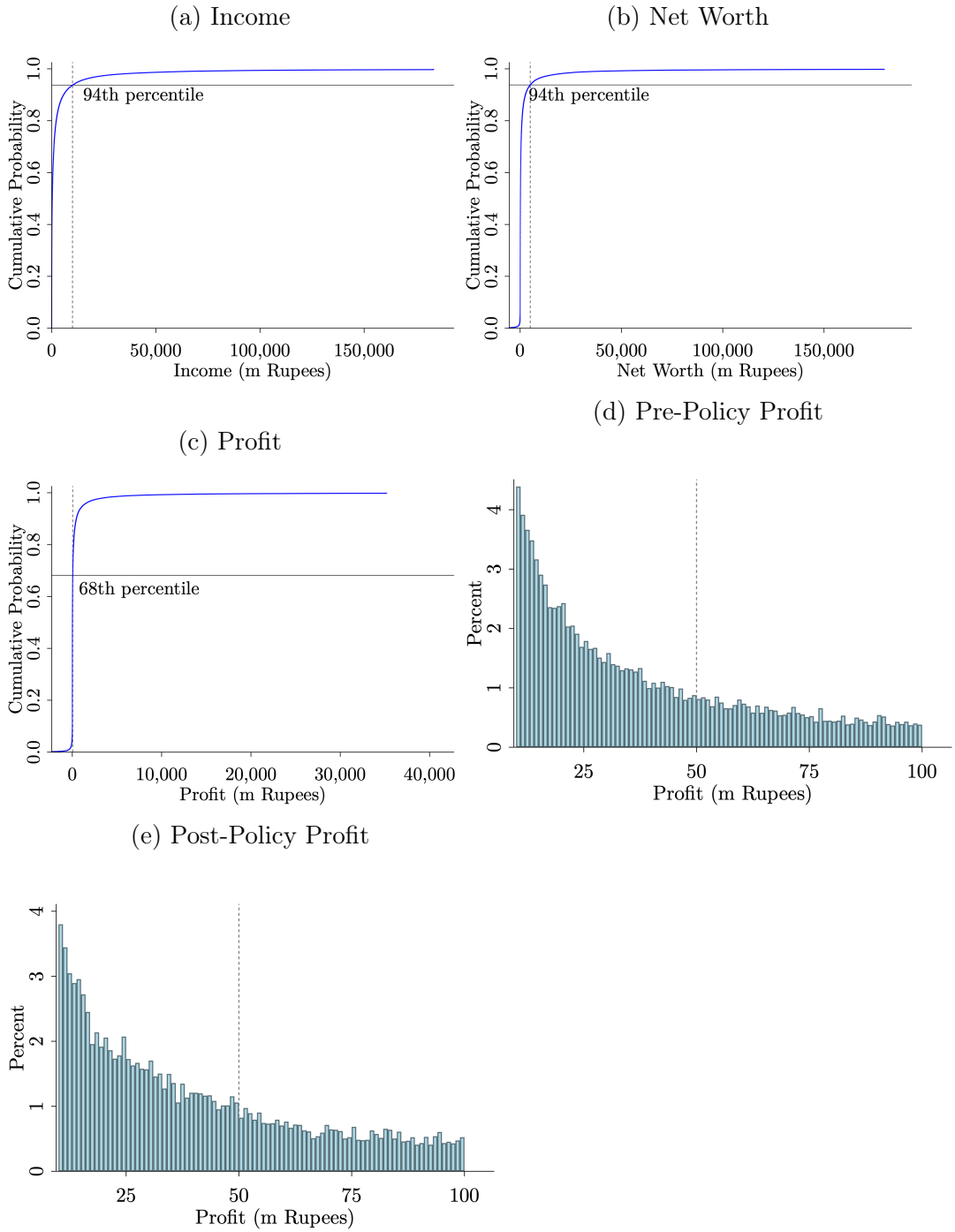
Finally, we investigate the impact of CSR spending on government spending in Table D.3. A simple OLS regression of the log of government spending on the log of CSR spending on the state×year level yields an insignificant coefficient of 0.012. To further explore this relationship, we employ an instrumental variables approach. The instrument is constructed as the share of aggregated state-level firm profit subject to the policy (measured prior to 2013), interacted with a post-policy indicator. While this instrument positively predicts CSR spending, the first-stage coefficient is not statistically significant. In the second stage, the estimated effect of CSR on government spending remains similar in magnitude to the OLS result and likewise insignificant. Given the limitations of the research design, we refrain from making any causal claims about the impact of CSR spending on government spending.

Figure D.1: Effect of the Policy on CSR Spending



Notes: This figure describes the effect of the policy on CSR spending, derived from equation (D.1). The unit of observation is at the firm×year level. The dependent variable is the CSR spending of a given firm (f) in a given year (y) over average profits in the past three years. The independent variables are the interactions of Treated_{fy} with year indicators. Treated_{fy} is an indicator equal to one if the firm is liable under the CSR regulation in year y, that is, if either profits, income, or net worth are above their respective thresholds in any of the three preceding financial years. The green dashed line replaces the year fixed effects in the regression with year×industry×state fixed effects. Variables are winsorized at the 99th percentile. Standard errors are clustered at the firm level. The figure shows 95% confidence intervals.

Figure D.2: Manipulation of Liability Status



Notes: This figure tests for manipulation of the liability status. The unit of observation is at the firm×year level. Figures D.2(a) to D.2(c) show the cumulative probability for the three size thresholds. Figure D.2(d) shows the profit distribution between 2007 and 2014. Figure D.2(e) shows the profit distribution between 2015 and 2019.

Table D.1: Effect of the Policy on CSR Spending

	$CSR_{fy}/\overline{Profit}^{3y}_{fy}$			
	DID	DID	DID-IV	DID-IV
	(1)	(2)	(3)	(4)
$Treated_{fy} \times Post_y$	0.011***	0.011***		
	(0.001)	(0.001)		
$Treated_f \times Post_y$			0.012***	0.012***
			(0.001)	(0.001)
Firm FE	✓	✓	✓	✓
Group FE	✓	✓	✓	✓
Year FE	✓		✓	
Year×Industry×State FE		✓		✓
F Statistic			231	174
R-squared	0.31	0.34	0.00	0.00
Observations	197,729	197,729	197,729	197,729

Notes: This table describes the effect of the policy on CSR spending, derived from equation (D.1). The unit of observation is at the firm×year level. In columns 1 and 2, the dependent variable is the CSR spending of a given firm (f) in a given year (y) over average profits in the past three years. The independent variable is the interaction of $Treated_{fy}$ with $Post_y$. $Treated_{fy}$ is an indicator equal to one if the firm is liable under the CSR regulation in year y, that is, if either profits, income, or net worth are above their respective thresholds in any of the three preceding financial years. $Post_y$ is a dummy equal to one every year from 2015 onwards. In columns 3 and 4, we instrument the time-varying liability variable $Treated_{fy}$ with $Treated_f$, which is an indicator equal to one if the firm is liable under the CSR regulation in the year 2014, that is, if either profits, income, or net worth, are above their respective thresholds in any of the three preceding financial years (2011–2013). Variables are winsorized at the 99th percentile. Standard errors are clustered at the firm level.

Table D.2: Manipulation by Relabeling

	CSR (Social + Donations) (1)	Social (2)	Donations (3)	Environment (4)	Employee Welfare (5)
$\text{Treated}_{fy} \times \text{Post}_y$	0.44**	0.49***	-0.05**	-0.03	0.51
	(0.19)	(0.18)	(0.03)	(0.03)	(0.44)
Avg dep var	0.18	0.07	0.11	0.02	0.53
Firm FE	✓	✓	✓	✓	✓
Group FE	✓	✓	✓	✓	✓
Year×Ind.×State FE	✓	✓	✓	✓	✓
R-squared	0.30	0.29	0.35	0.31	0.24
Observations	197,729	197,729	197,729	197,729	197,729

	Employee Training (6)	Social Amenities (7)	Advertising (8)	Marketing (9)
$\text{Treated}_{fy} \times \text{Post}_y$	0.01*	-0.00	0.05	-0.15***
	(0.01)	(0.00)	(0.03)	(0.05)
Avg dep var	0.03	0.00	0.60	1.14
Firm FE	✓	✓	✓	✓
Group FE	✓	✓	✓	✓
Year×Ind.×State FE	✓	✓	✓	✓
R-squared	0.39	0.18	0.42	0.55
Observations	197,729	197,729	197,729	197,729

Notes: This table tests for manipulation by relabeling regular business expenses, derived from equation (D.1). The unit of observation is at the firm×year level. All expenses are divided by total expenses and multiplied by 100. In column 1, the dependent variable is CSR spending, defined as the sum of social expenses and donations. Columns 2 and 3 split the components of this CSR variable. Columns 4 to 9 have as dependent variables expenses that the firm could have possibly relabeled. The independent variable is the interaction of Treated_{fy} with Post_y. Treated_{fy} is an indicator equal to one if the firm is liable under the CSR regulation in year y, that is, if either profits, income, or net worth are above their respective thresholds in any of the three preceding financial years. Post_y is a dummy equal to one every year from 2015 onwards. Variables are winsorized at the 99th percentile. Standard errors are clustered at the firm level.

Table D.3: CSR and Government Spending

	Log(Government Spending)$_{sy}$		Log(CSR)$_{sy}$
	OLS	Second Stage	First Stage
	(1)	(2)	(3)
Log(CSR)$_{sy}$	0.012	0.010	
	(0.009)	(0.028)	
Instrument (Treated Profit Share×Post)$_{sy}$			0.828
			(1.518)
State FE	✓	✓	✓
Year FE	✓	✓	✓
R-squared	1.00	0.60	0.96
Observations	326	326	326

Notes: This table tests for the effect of CSR spending on government spending. The unit of observation is the state×year level. Column 1 provides an OLS regression (equation (D.2)). Column 2 provides the second stage (equation (D.3)), and column 3 the first stage (equation (D.4)). In column 1 and 2, the dependent variable is the log of government spending in a given state and year. In column 3, the dependent variable is the log of CSR spending in a given state and year. The independent variable in column 1 is the log of CSR spending in a given state and year, and in column 2 the same variable instrumented by the share of treated profits in a given state times a Post variable that is one after the policy introduction. Controls include state GDP, the number of firms in Prowess, the number of treated firms in Prowess, the average income and profit of firms in Prowess, as well as state-level linear time trends and region×year fixed effects. Standard errors are clustered on the state level.

$$\text{Log(Government Spending)}_{sy} = \beta \text{Log(CSR)}_{sy} + \phi X_{sy} + \gamma_s + \gamma_y + \epsilon_{sy} \tag{D.2}$$

$$\text{Log(Government Spending)}_{sy} = \beta \widehat{\text{Log(CSR)}}_{sy} + \phi X_{sy} + \gamma_s + \gamma_y + \epsilon_{sy} \tag{D.3}$$

$$\text{Log(CSR)}_{sy} = \beta \text{Treated Profit Share} \times \text{Post}_{sy} + \phi X_{sy} + \gamma_s + \gamma_y + \epsilon_{sy} \tag{D.4}$$

E Model

E.1 Model Derivations

Production. Production occurs across projects, indexed by $p \in \mathcal{P}$. $\mathcal{P}$ can be partitioned into $\mathcal{P}^{pub}$ for CSR projects and $\mathcal{P}^{pri}$ for for-profit goods sold in competitive markets. Projects are produced by combining tasks $\tau \in \mathcal{T}$. The relative importance of tasks varies across projects as characterized by the vector $\boldsymbol{\phi}_\mathbf{p} = [\phi_{p\tau}]_{\tau\in\mathcal{T}}$ such that $\forall p,\ \sum_\tau \phi_{p\tau} = 1$. To perform tasks, firms hire workers and are endowed with a task-specific productivity vector $\mathbf{z_f} = [z_{f\tau}]_{\tau\in\mathcal{T}}$. Firm f's output in project p is:

$$y_{fp} = \left[\prod_{\tau\in\mathcal{T}} x_{fp\tau}^{\phi_{p\tau}}\right]^\rho$$

where $\rho < 1$ and $x_{fp\tau}$ is the amount of task τ used in production of project p. In addition, $x_{fp\tau} = \exp(z_{f\tau})\ell_{fp\tau}$, where $\ell_{fp\tau}$ is the labor assigned to task τ and project type p. The wage w is taken to be exogenous.

Project-specific productivity. Conditional on producing project p, the firm allocates labor ℓ_{fp} across tasks in a way that maximizes project-specific returns:

$$\pi_{fp} = \max_{\{\ell_{fp\tau}\}} \zeta_{fp} y_{f,p} - w\ell_{fp}$$

subject to: $\sum_{\tau\in\mathcal{T}} \ell_{fp\tau} = \ell_{fp}$. When p is a for-profit good, ζ_{fp} is the market price of good p. When p is a CSR project ζ_{fp} captures how much firm f values project type p, as above. The optimal labor allocation satisfies: $\frac{\ell_{fp\tau}}{\ell_{fp}} = \phi_{p\tau}$. Project-level profit maximization yields an expression for project output as a function of firm-project productivity α_{fp} :

$$y_{fp} = \exp(\alpha_{fp})\ell_{fp}^\rho, \tag{E.1}$$

with:

$$\alpha_{fp} = \rho||\boldsymbol{\phi}_\mathbf{p}||||\mathbf{z_f}||\cos(\boldsymbol{\phi}_\mathbf{p}, \mathbf{z_f}) + \kappa_p \tag{E.2}$$

where $\kappa_p = \rho\sum_\tau \phi_{p\tau}\log(\phi_{p\tau})$ is a product-level constant.

$\cos(\boldsymbol{\phi}_\mathbf{p}, \mathbf{z_f})$ is the cosine similarity between the vector of returns to tasks for that

project type $\boldsymbol{\phi_p}$ and the firms' vector of task-specific technologies $\mathbf{z_f}$. In the main text, we assume that $||\boldsymbol{\phi_p}|| = \phi, \forall p$ and $||\mathbf{z_f}|| = 1, \forall f$. Then,

$$\alpha_{fp} = \rho\phi \cos(\boldsymbol{\phi_p}, \mathbf{z_f}) + \kappa_p \tag{E.3}$$

Privately and socially optimal allocations. The firm allocates labor across projects to maximizes total returns:

$$\max_{\{\ell_{fp}\}} \sum_{p \in \mathcal{P}^{pri}} (\zeta_{fp} y_{fp} - w\ell_{fp}) + \sum_{p \in \mathcal{P}^{pub}} (\zeta_{fp} y_{fp} - w\ell_{fp})$$

where we now define the exogenous CSR expenditure requirement as $\sum_{p \in \mathcal{P}^{pub}} w\ell_{fp} = E$. We obtain that firms' privately optimal CSR shares follow the expression in equation (5), with α_{fp} defined in (7). The social planner maximizes the social welfare function given in (3) above. This yields the same socially optimal allocation across CSR project types p as above (equation (4)).

Implications for CSR productivities across industries. Our empirical tests exploit variation in firms' technologies at the level of their industry. We define an industry as a set of private goods $\mathcal{P}_i \subset \mathcal{P}^{pri}$ centered around a technological vector $\boldsymbol{\phi_i}$: $\forall p \in \mathcal{P}_i, \boldsymbol{\phi_p} = \boldsymbol{\phi_i} + \boldsymbol{\epsilon_p}$. We assume that $\boldsymbol{\epsilon_p}$ is mean-zero and i.i.d. That is, for all p, i, $\boldsymbol{\phi_i} \perp \boldsymbol{\epsilon_p}$.

In addition, we assume that firms belonging to industry i have a productivity vector centered around $\boldsymbol{\phi_i}$: $\mathbf{z_f} = \boldsymbol{\phi_i} + \boldsymbol{\epsilon_f}$. Likewise, we assume that $\boldsymbol{\epsilon_f}$ is mean-zero and i.i.d. That is, for all f, i, $\boldsymbol{\phi_i} \perp \boldsymbol{\epsilon_f}$.

This ensures that, on average, firms in industry i have $p \in \mathcal{P}_i$ as their main product, in line with how firms are classified across industries in standard datasets.

Proof. Assume that the valuation of private goods is equalized across goods: $\zeta_{fp} = \zeta_f$. Then, the sales share of product p, denoted s_{fp}, is strictly increasing in α_{fp}. Take f in industry i. $\mathbf{z_f} = \boldsymbol{\phi_i} + \boldsymbol{\varepsilon_f}$. Take product p in industry i and product p' in industry i'. We have:

$$\alpha_{fp} = \rho\ \boldsymbol{\phi_i} \cdot \boldsymbol{\phi_i} + \kappa_p$$
$$\alpha_{fp'} = \rho\ \boldsymbol{\phi_i} \cdot \boldsymbol{\phi_{i'}} + \kappa_{p'}$$

Assuming that $\kappa_p = \sum_\tau \phi_{p\tau} \log(\phi_{p\tau})$ is approximately constant across products, and using the Cauchy-Schwarz inequality, we obtain the desired results. □

For a firm f in industry i, we can then write:

$$\alpha_{fp} = \rho\phi \cos(\boldsymbol{\phi_p}, \boldsymbol{\phi_i}) + \kappa_p \tag{E.4}$$

E.2 Model-Based Tests of Measurement Approach

Test 1. Industry predicts semantic proximity across firms. Let $i(f)$ denote the industry of firm f. If $i(f) = i(f')$ and $i(f) \neq i(f")$, then:

$$\cos(\mathbf{z_f}, \mathbf{z_{f'}}) \geq \cos(\mathbf{z_f}, \mathbf{z_{f''}}) \tag{E.5}$$

Proof. From our definition of industries,

$$\cos(\mathbf{z_f}, \mathbf{z_{f'}}) = \mathbf{z_f} \cdot \mathbf{z_{f'}} = \boldsymbol{\phi_i} \cdot \boldsymbol{\phi_i}$$

Following the same steps, $\cos(\mathbf{z_f}, \mathbf{z_{f''}}) = \boldsymbol{\phi_i} \cdot \boldsymbol{\phi_{i''}}$ for $i \neq i''$. $\boldsymbol{\phi_i} \cdot \boldsymbol{\phi_i} \geq \boldsymbol{\phi_i} \cdot \boldsymbol{\phi_{i''}}$ by the Cauchy-Schwarz inequality. □

Test 2. Semantic proximity across industries predicts firms' sales shares across industries. Assume that the valuation of private goods is equalized across goods: $\zeta_{fp} = \zeta_f$. Take a product p in industry i. Let f' (f'') be a firm in industry i' (i''). Then,

$$\cos(\boldsymbol{\phi_i}, \boldsymbol{\phi_{i'}}) \geq \cos(\boldsymbol{\phi_i}, \boldsymbol{\phi_{i''}}) \Rightarrow s_{f'p} \geq s_{f''p} \tag{E.6}$$

That is, fixing product p in industry i, a firm in an industry closer to industry i will have a higher sales share on product p.

Proof. Using the assumption that $\zeta_{fp} = \zeta_f$, s_{fp} is strictly increasing in $\alpha_{fp} = \rho\phi \cos(\boldsymbol{\phi_p}, \mathbf{z_f}) + \kappa_p$ for any firm f and product p. For firm f' in industry i',

$$\alpha_{f'p} = \rho\phi \cos(\boldsymbol{\phi_i'}, \boldsymbol{\phi_i}) + \kappa_p \tag{E.7}$$

and similarly for firm f'' in industry i''. The conclusion immediately follows. □