Statistical capacity and corrupt bureaucracies

Manuel Oechslin¹ · Elias Steiner¹

Accepted: 26 March 2021 / Published online: 5 June 2021

© The Author(s) 2021

Abstract

In many developing countries, economic statistics (such as the growth rate of GDP) are imprecise, making it difficult to evaluate economic reforms and learn “what works”. Improving economic statistics has thus become a priority of international organizations. In this paper, we isolate an insidious mechanism—a type of observer effect—by which a push for better statistics can make matters worse. Precise statistics require the collection of data from a large number of firms. If firms suspect that detailed information, when spreading through the bureaucracy, is misused to collect bribes, they have weaker incentives to invest. As a result, the effects of reforms are muted, making it even harder to discover “what works”. To suppress this mechanism, efforts to improve economic statistics should be comprehensive and also include institutional aspects.

Keywords Economic statistics · Experimentation · Informativeness · Corruption · Observer effect

JEL Classification D83 · D73 · E01

Responsible Editor: Axel Dreher

Author contributions to conceptualization: M.O. (50%), E.S. (50%); to theoretical modeling: M.O. (60%), E.S. (40%); to statistical analysis and simulations: E.S. (100%); to writing: M.O. (70%), E.S. (30%).

The order of authors is chosen alphabetically.

✉ Manuel Oechslin
manuel.oechslin@unilu.ch

Elias Steiner
elias.steiner@unilu.ch

¹ University of Lucerne, Department of Economics, Frohburgstrasse 3, 6002, Lucerne, Switzerland

The census has nothing to do with taxation, with army or jury service [...], nor can any person be harmed in any way by furnishing the information required.—
President W.H. Taft, 1910 Census Proclamation

1 Introduction

In many developing countries, economic statistics such as the growth rate of GDP, the inflation rate, or the unemployment rate are highly unreliable. For example, in a widely noticed book, Jerven (2013) documented that the quality of African GDP numbers is extremely “poor”. At the same time, the World Bank’s then chief economist for Africa referred to the deficient state of African economic statistics as a “statistical tragedy” (Devarajan 2013). It is therefore no surprise that improving developing countries’ statistics has become a priority of international organizations, among them the World Bank, the IMF, and the OECD. These organizations pursue their objective through initiatives such as the “Partnership in Statistics for Development in the 21st Century” (PARIS21), a group concerned with technical issues and the funding of data collection and processing in poor countries. More recently, the push for better statistics in developing countries has gained additional momentum through the rise of digitalization and big data. Under the umbrella of the UN Global Working Group for Big Data in Official Statistics (GWG Big Data), both developing and advanced countries exchange experiences concerning the use of big data to improve economic statistics.¹

In many ways, improvements in the precision and availability of economic statistics would be highly welcome. In particular, considering that concepts such as “growth diagnostics” (e.g., Rodrik 2010) and “experimentation at scale” (e.g., Muralidharan and Niehaus 2017) have gained ground, accurate statistics are of increasing importance in the context of development policy. Growth diagnostics, for instance, is based on the notion that—when it comes to incremental economic reforms—which reforms work and which do not is highly context-specific, i.e., depends on a country’s economic and institutional status quo. Therefore, as Rodrik (2010, p. 41) puts it, growth diagnostics “emphasizes experimentation as a strategy for discovery of what works, along with monitoring and evaluation.”

A condition for meaningful monitoring and evaluation is the availability of accurate statistics. If the numbers are poor, evaluation may become impossible or may lead to erroneous conclusions about “what works” (Manski 2015). This paper does not deny that good statistics have many benefits. Yet, focusing on GDP statistics, we isolate an insidious mechanism by which a push for better statistics can have harmful side effects in developing countries. This mechanism—a type of observer effect—reduces the benefits or may reverse them into net losses. Our analysis suggests that efforts to improve developing countries’ statistics should not have a narrow focus on technical statistical capacity (i.e., data gathering); such efforts should be comprehensive and include institutional aspects (e.g., data confidentiality), notably in places

¹ When devising the Sustainable Development Goals (SDGs), the UN even called for a “data revolution” that would improve official (economic) statistics by harnessing digitalization and big data.

where the bureaucracy has an extractive nature. It is key that there be a symmetry between technical statistical capacity and the quality of the institutional setting.

Our argument rests on three observations. First, strengthening technical statistical capacity to improve GDP statistics necessarily means collecting more data. It includes a move to a regular economic census schedule and the enlargement of the firm surveys that underlie GDP estimates between the censuses (e.g., Berry et al., 2018, Jerven 2013, p. 26). Second, although there often are official guarantees of confidentiality for census and survey data, a large number of reports on the handling of government-collected data suggest a grave risk of confidentiality breaches that may allow detailed firm data to spread widely within the bureaucracy and beyond. As we will discuss in Section 2, the reasons for this include widespread IT security holes and the expansive sharing of collected data (e.g., Brookings Institution 2018).² Third, control of corruption is weaker in developing countries (e.g., Olken and Pande 2012), and corrupt officials use information on firm characteristics to “bribe discriminate” (Svensson 2003), with the consequence that larger firms pay higher bribes (e.g., Bai et al. 2019).

Connecting these observations, a push to strengthen technical statistical capacity must arouse fear of higher bribery costs among firms: with larger surveys, each firm faces a higher chance of being sampled and, given the possibility of confidentiality breaches, a higher chance of being confronted with bribe demands from somewhere within the bureaucracy. The expectation that the official confidentiality (or no-harm) assurances may not hold weakens firms’ incentives to invest. But if firms hold back for fear of getting (even more) entangled in bribe demands, their responses to any reform policy—and thus the policy’s overall effect on economic performance—will be muted. As a result, although the improved statistics reduce the noise in growth estimates, it may become more difficult for the government to discover what policies work. In other words, the informativeness of policy experiments may fall rather than rise. In this case, a push to improve technical statistical capacity slows down learning. While this particular mechanism is novel, individual elements are well documented. A growing body of empirical results suggests that corruption has a strong negative impact on investment (e.g., Beekman et al. 2014, Paunov 2016, Zakharov 2019). There is also evidence of corruption biasing the effects of reforms towards zero (e.g., Banerjee et al. 2019).

To examine the relationship between technical statistical capacity and societal learning about reforms, this paper proposes a theoretical two-period framework that features *ex ante* fundamental uncertainty about the effects of alternative reform options (as in Binswanger and Oechslin 2015, 2020) and *ex post* measurement uncertainty in the economy’s key statistic, the output estimate. Measurement uncertainty stems from the fact that the statistical office has to base its output estimate on a random sample of firms. Within the framework, the size of this firm sample can be interpreted as a measure of the economy’s technical statistical capacity. An improvement in technical statistical capacity reduces measurement uncertainty and hence improves the accuracy of output estimates. A further key element of the framework is that it treats firms’ investment decisions as endogenous and allows for the possibility

²Drawing on examples from Kenya and India, Section 2 also documents that frequent media reports on leaks of government-collected data make firms and the general population aware of these problems.

of firms being subjected to bribe collection by bureaucrats. Specifically, we assume that firms sampled by the statistical office face a positive chance of being confronted with (possibly additional) bribe demands that amount to a fixed proportion of their current revenue.

In this framework, what are the consequences of an exogenous improvement in technical statistical capacity? Holding constant firms' investments, a fall in measurement uncertainty permits a more reliable assessment of whether an implemented reform boosts output or whether the government should pursue adjustments to make the reform work; as a result, the learning process speeds up. However, firms' investments do not stay constant when technical statistical capacity improves: a larger sample implies that each individual firm faces a higher probability of being sampled and hence a higher risk of being subjected to bribe collection; as a result, firms scale back their investments,³ a response that has direct negative consequences for economic performance. But even more importantly, with smaller investments, economic reforms have a smaller effect on output—which, in turn, makes it more difficult for the government to determine whether an implemented reform works or needs adjustment.

So an improvement in technical statistical capacity has two opposing effects on the speed of the societal learning process and hence economic performance. A key implication of our framework is that—if control of corruption is sufficiently weak—there is a hump-shaped relationship between technical statistical capacity and economic performance: increasing the firm sample helps initially but reduces expected output beyond some critical threshold. In other words, although sampling is costless and the government is interested in learning about reforms, the optimal sample size is strictly smaller than the total number of firms. Corruption, by impairing the government's ability to identify the consequences of its reform decisions, retards economic growth by slowing down the learning process about “what works”.

Over the past few years, a growing literature on the quality of economic statistics in developing countries has emerged. Often, the quality is shown to be low, pointing to a large potential for improvements in the timeliness and precision of economic indicators (e.g., Devarajan 2013, Jerven 2013, Kiregyera 2015, Sandefur and Glassman 2015, Kerner et al. 2017). Many papers consider the link between the quality of statistics and policy making. Rodrik (2010) stresses the importance of high quality data for evidence-based development policy. Manski (2015) worries that imprecise estimates may lead to bad policy decisions if policy makers fail to account for measurement error. Binswanger and Oechslin (2015) argue that better statistics—by making evaluations of past policy changes more reliable—could reduce disagreements and promote economic reforms. More in line with the present paper, Binswanger and Oechslin (2020) identify adverse effects of better statistics in electoral democracies. Even though the present paper also rests on a model of policy learning, it differs significantly from the former. Here, we relate learning about policies to corruption and explicitly model firms' investment choices. This connects

³In this paper, we also sketch a modified model in which firms can switch to informal operations. In the modified setup, a higher risk of being confronted with bribe demands merely leads to a reallocation of investment towards informal operations. Yet the key insights of the baseline analysis are robust to this modification.

us with the large literature on corruption, in particular with research on how corruption constrains investment and, more generally, the growth aspirations of firms (e.g., Fisman and Svensson 2007, Estrin et al. 2013, Freund et al. 2016, Paunov 2016, Zakharov 2019, Colonnelli and Prem 2020) and with work on how corruption reduces the socially optimal level of government intervention (e.g., Immordino and Pagano 2010).

The rest of this paper is organized as follows. The next section presents motivating evidence. In Section 3, we describe the theoretical setup. Section 4 solves for the equilibrium, assuming a given level of technical statistical capacity. Section 5 derives the optimal capacity level and discusses the harmful role of corruption. Section 6 offers a modified version of the model that allows for informality and misreporting. Finally, Section 7 concludes.

2 Motivating evidence

2.1 Economic statistics and economic performance

We start by presenting motivating evidence on the relationship between the quality of economic statistics and economic performance. We also consider the moderating role of corruption. To capture the quality of economic statistics, we use the World Bank's Statistical Capacity Index (SCI). The SCI is available for 153 developing and emerging economies at yearly frequency. We have recoded the index so that it ranges from 0 to 1, where 1 indicates maximum statistical capacity (the original range is 0 to 100). The index measures the extent to which a country's statistical system adheres to international technical standards deemed essential for the quality of economic data. We use the growth rate of real GDP p.c. (PPP, constant 2011 \$) to capture economic performance and the World Bank's Control of Corruption Index (CCI) as a measure for (the absence of) corruption. The CCI is a corruption perception index that is concerned with the exercise of public power for private gain and is constructed from a broad range of sources.⁴ Our dataset includes 146 countries and covers the period from 2005 to 2016. The figures in this section rely on observations averaged over periods of three years (2005-07, 2008-10, 2011-2013, 2014-16). The full sample includes 556 observations.

Figure 1 shows a partial residual plot that illustrates the correlation between the residual growth rate of real GDP p.c. and statistical capacity.⁵ We see a significant positive relationship: an increase in the SCI of one standard deviation (0.16) is associated with a rise in real GDP p.c. growth of 0.6 percentage points. Figure 1 is based on the full sample and does not account for cross-country differences in corruption. The role of corruption is highlighted in Fig. 2, which again shows partial residual plots. Each subfigure considers two disjunct subsamples of the full sample. One of

⁴ The data were retrieved on July 4, 2018, from the [World Bank Open Data](#) portal. We dropped as outliers all observations that belong to either the bottom or top 0.5% in the distribution of GDP p.c. growth rates.

⁵ Besides SCI, the underlying OLS regression includes as explanatory variables the log of real GDP p.c. (to control for convergence) and period fixed effects (to control for period-specific shocks common to all countries).

Fig. 1 Statistical capacity and growth, full sample, 2005–2016. *Note:* This figure shows a partial residual plot. The underlying linear OLS regression relates the average annual growth rate of real GDP p.c. to a constant, the SCI, the log of real GDP p.c., and period fixed effects. The value of the coefficient on the SCI (which is also the slope of the fitted line) is 3.913 (p -value: 0.000)

the subsamples contains all countries with an average CCI score belonging to the top 25% of the distribution (“low corruption”) and the other one all countries with an average CCI score belonging to the bottom quartile (“high corruption”). Examples of low-corruption countries are Uruguay and Chile (both high SCI), but also Botswana and Namibia (both low SCI). High-corruption countries are, among others, Libya and Equatorial Guinea (both low SCI), but also Russia and Ukraine (both high SCI). Subfigure (a) of Fig. 2 replicates the regression from Fig. 1 for the two subsamples. The plot suggests that the positive correlation found in Fig. 1 is driven by observations from low-corruption countries. While there is a positive relationship between the growth rate of real GDP p.c. and statistical capacity in the low-corruption subsample, no such relationship emerges among high-corruption countries.⁶ In Subfigure (b), the underlying regression additionally includes country fixed effects. The moderating role of corruption seems to be even stronger. However, in statistical terms, the difference becomes less significant.

The basic pattern shown in Fig. 2 is fairly robust to a number of modifications. It remains mostly unaffected when we use larger subsamples of low- and high-corruption countries,⁷ or when we split the full sample into disjunct subsamples

⁶One concern relating to the analysis of split samples is a possible lack of common support. But a first glance at Fig. 2 suggests that the two subsamples have a considerable common support in terms of SCI. A more formal procedure for visual inspection suggested by Hainmueller et al. (2019) confirms this impression.

⁷The 25%-threshold (top/bottom) is chosen for visual clarity. The general conclusion stated at the end of the preceding paragraph remains valid when we gradually move to larger thresholds, up to the median.

Fig. 2 Statistical capacity and growth, two subsamples, 2005-2016. *Note:* Each subfigure shows a partial residual plot for countries with an average CCI score belonging to the top (“low corruption”) or bottom (“high corruption”) 25% of the distribution (145 observations each). The underlying linear OLS regressions in Subfigure (a) relate the average annual growth rate of real GDP p.c. to a constant, the SCI, the log of real GDP p.c., and period fixed effects. The values of the coefficients on the SCI (which are also the slopes of the fitted lines) are -0.015 (p -value: 0.994) for the high-corruption subsample and 6.527 (p -value: 0.000) for the low-corruption subsample. The p -value for the difference in the two coefficients is 0.011. In Subfigure (b) country fixed effects are added to the underlying regression. The values of the coefficients on the SCI are -2.217 (p -value: 0.671) for the high-corruption subsample and 8.196 (p -value: 0.148) for the low-corruption subsample. The p -value for the difference in the two coefficients is 0.174

according to the World Bank’s Rule of Law Index (instead of the CCI). Moreover, given the concerns regarding GDP data quality, we were also using the (log) change in nighttime light intensity as a proxy for economic performance (Henderson et al. 2012). The source of the light data is Hodler and Raschky (2014), who aggregated the georeferenced raw data to ADM2 administrative levels (from where we aggregated it to the country level).⁸ The data are scaled from 0 to 63, with a larger number reflecting more intense nighttime lights. The data are available up to 2013, which leaves us with 399 observations from 134 countries. While the differences between low- and high-corruption countries are smaller and more sensitive to the threshold applied, Fig. 3 shows that the swap of GDP for light data does not change the basic pattern documented in Fig. 2.

Overall, we find a positive relationship between economic performance and statistical capacity among low-corruption countries, while no such relationship appears among high-corruption countries. This pattern is consistent with our model, which predicts improvements in technical statistical capacity to lift economic growth in low-corruption places but warns that such a positive relationship should not be expected in places where corruption is higher.

2.2 Data leaks and reactions

At a more general level, the mechanism we explore emphasizes that increased data gathering by the government makes people adjust their behavior in anticipation of

⁸The raw data come from the National Oceanic and Atmospheric Administration (NOAA) (2014). To aggregate it, Hodler and Raschky (2014) use the GADM database of Global Administrative Areas (2012).

Fig. 3 Statistical capacity and nighttime light intensity, two subsamples, 2005–2013. *Note:* This figure shows a partial residual plot for countries with an average CCI score belonging to the top (“low corruption”) or bottom (“high corruption”) 25% of the distribution (101 and 102 observations, respectively). The underlying linear OLS regressions relate the log difference of mean light intensity to a constant, the SCI, the log of mean light intensity, and period fixed effects. The values of the coefficients on the SCI (which are also the slopes of the fitted lines) are -0.015 (p -value: 0.771) for the high-corruption subsample and 0.046 (p -value: 0.215) for the low-corruption subsample

the possibility that the data, when leaked, is used to their disadvantage. In turn, these adjustments in behavior may well undermine the very purpose of data gathering. There is anecdotal evidence that broadly supports the relevance of this mechanism in developing countries. First of all, fear of data leaks is clearly well-founded. One reason is the prevalence of IT security holes. In a recent focus on Africa, the Brookings Institution (2018) warns that throughout the continent and across sectors, including government, the commitment to cybersecurity is weak. Again relating to Africa, Serianu (2017) reports that the government sector is among the most frequent victims of cybercriminal activity. This aggregate perspective meshes well with country-level accounts. As an example, consider Kenya, a country where reports about data leaks abound. For instance, in April 2016 servers of the Kenyan Ministry of Foreign Affairs were breached and one terabyte of (partly confidential) information stolen (Privacy International 2019b). This should not come as a surprise, considering that according to an estimate from 2017 more than 80% of public sector institutions in Kenya do not even have the means to detect network intruders. The frequent leaking of data held by public institutions does not go unnoticed and leads to reactions in the

Fig. 4 An enormous data leak. *Note:* This article appeared in the **Tribune** (Jan 4, 2018)

business community: in 2016, a newspaper headline ran "Hoteliers withhold data from KNBS [Kenya National Bureau of Statistics] citing fears of leaked business secrets".⁹

Besides IT security holes, the expansive use of data gathered by the government is another reason for why fear of leaks is often warranted. A particular ostensive illustration of this problem is India's Aadhaar biometric database. According to Privacy International (2019a) the breaches and leaks of personal data have been enormous. At some point in 2018, personal details of hundreds of millions of Indians apparently could be purchased online for as little as ₹ 500 (see Fig. 4). The leaks did not seem to be driven by weak IT security (Privacy International 2019a); they rather appeared to be a result of the fact that the data are shared across the public sector to an extent way beyond what was originally imagined. In this context, an article in the *Economist* (Dec 18, 2018) noted that "No one knows for certain how much Aadhaar-associated data have been shared with whom, (...)." It is obvious that the wide sharing of data, in combination with weak governance, increases the risk of someone making illicit use of them. The consequences for Aadhaar are severe: "A system designed to prevent fraud has given rise to a whole new economy of fraudulent activity (...)." As in Kenya, there is evidence that people respond to the danger of data breaches. For instance, reports suggest that HIV patients preferred to abandon essential treatment programs when they had their Aadhaar numbers linked to their patient identity cards (Privacy International 2019a).

The general pattern of anticipation and response that arises in the examples from Kenya and India can be found in many places. In the model below, it emerges in a generic setting that considers data gathering for the purpose of evaluating uncertain policies.

⁹ The number on network intruder detection comes from the *Business Daily* (Apr 27, 2017), a newspaper. The headline on the hoteliers' refusal to take part in data collection appeared in the *Star* (Nov 21, 2016).

3 The model

3.1 Output, revenue, and economic policy

We consider a two-period economy with $N > 0$ firms. Each firm $i \in \{1, \dots, N\}$ is equipped with a technology to produce a homogeneous good whose price is normalized to 1. The technology is uniform across firms and represented by the production function

$$y_{it} = y(A_t, x_{it}) = A_t x_{it}^\alpha, \quad (1)$$

where $t \in \{1, 2\}$ denotes time and $0 < \alpha < 1$. A_t is a productivity parameter and x_{it} refers to a firm-specific investment. For simplicity, the per-unit cost of investment is normalized to 1, too. Following Barro (1990) and the subsequent literature, we assume that economic policy plays a productive role in output generation; it offers complementary inputs to private production (such as reliable power supply) and so potentially creates a positive link between government, investment, and economic performance. However, unlike much of the literature, our setup includes a status-quo policy and assumes that any deviation from the status quo has an uncertain effect on production. For concreteness, suppose that economic policy, P_t , enters production function (1) via the productivity parameter. Following Binswanger and Oechslin (2020), we assume $P_t \in \{-1, 0, 1\}$, where 0 is the status-quo policy and -1 and 1 refer to two alternative reform policies. Policy affects productivity according to

$$A_t = 1 + \sqrt{\gamma} P_t S, \quad (2)$$

where $0 < \gamma < 1$ reflects the economic significance of the reform and $S \in \{-1, 1\}$ captures the unobserved and invariable “state of the world” that materializes prior to the start of the economy.¹⁰ In Eq. (2), we use the square root of γ to simplify the formal presentation of the main results below. Together, Eqs. (1) and (2) imply that a reform policy is beneficial (harmful) if its sign is the same as (is different from) the sign of the state. S takes each of its two possible values with probability $1/2$. This specific value is chosen for analytical convenience and not important for our argument. What matters is that there is some uncertainty as to whether a particular reform alternative has a positive or a negative effect on productivity. For the policy maker, the only way to gain information is “policy experimentation”: implement one of the alternatives, monitor the result—and then adjust if necessary. We note that in terms of key results our analysis would be unchanged if there were no status-quo policy. The difference would be that in such a simplified setup the government would be forced to try an uncertain policy while in the present framework this will be an endogenous decision.

While producing the homogeneous good is the core activity, each firm additionally runs its individual side business. There are many possibilities: wholesale trade in a

¹⁰For example, suppose that A is an increasing function of the reliability of power supply. The status quo ($P = 0$), consisting of state-owned power plants and privately-owned distribution networks, produces mediocre results—possibly because of frictions between the two parts. In this setting, plausible alternative reforms would be the privatization of power generation ($P = 1$) or the nationalization of power distribution ($P = -1$).

different good, property dealing, services such as repairs, etc. The side business is volatile and enters the analysis as an exogenous net contribution ζ_{it} to the total firm revenue z_{it} :

$$z_{it} = y_{it} + \zeta_{it}. \quad (3)$$

To capture the volatile nature of individual side businesses, we assume that ζ_{it} is a continuous i.i.d. random variable with support on $[0, \infty)$, mean $\chi > 0$, and variance σ . In combination with the informational frictions introduced in Section 3.2 below, the plausible consequence of firms running side businesses will be that the impact of any reform cannot be immediately inferred from the data of a single or a handful of firms. Hence the case for systematic data gathering. To keep the analysis tractable, the modeling of side businesses is parsimonious. Among other things, a side business is unaffected by policy and its contribution to total firm revenue is independent of the scale of the firm's core activity. The latter assumption, a measure of independence from core activities, is important for some of our findings.¹¹

It is worthwhile to note that there is no need to impose a specific distribution function for the ζ_{it} s. But it is helpful to make two “technical” assumptions in this regard. First, to secure that the model is scale invariant regarding the number of firms, we assume

$$\sigma = \theta N, \quad (4)$$

where $\theta > 0$. The parameter θ will govern the volatility of the average side business contribution. Second, for a reason that will become clear below, we assume that the distribution of the ζ_{it} s has a light left tail. Finally, we will use an asterisk (*) to mark values reflecting optimal firm choices: x_{it}^* denotes firm i 's optimal investment level, while y_{it}^* and z_{it}^* refer to, respectively, the resulting output of the homogeneous good and the resulting total firm revenue.

3.2 Informational constraints and output estimation

The government runs a statistical office that is tasked with the timely collection of firm-level data with the sole purpose of producing economic statistics at the end of each period.¹² Yet there are practical problems placing constraints on the level of detail with which firm-level data can be collected: within the limited time frame for completing statistical questionnaires, firms can only just identify total firm revenue; they do not have more detailed information on individual components yet. In brief, firm i observes z_{it}^* but does not individually observe y_{it}^* or ζ_{it} before period t ends. So firms can only report data on total firm revenue. Naturally, with time, the informational constraints ease and firms become able to separately identify the components

¹¹If the side businesses in- or deflated in perfect proportionality to the scale of the core activities ($z_{it} = y_{it} \cdot \zeta_{it}$), data gathering would still depress investment (and thus the effect of policy), but an interior level of data gathering could no longer be a maximizer of the informativeness of policy experimentation. The reason is that, with multiplicative shocks, a fall in investment mechanically reduces the side-business induced volatility of z_{it} .

¹²This is consistent with Principle 6 of the United Nation's [Fundamental Principles of Official Statistics](#), which says that “Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.”

of z_{it}^* . Specifically, firm i learns (y_{i1}^*, ζ_{i1}) in period 2 just before it is set to choose its level of investment. Provided that $P_1 \neq 0$, learning (y_{i1}^*, ζ_{i1}) permits the identification of A_1 and hence state S . Assuming that firms (possibly) identify S before they decide on their second-period investment simplifies the analysis without changing its substance.

Firms provide the statistical office with accurate revenue data (e.g., because misreporting, if detected, carries a prohibitive fine). The office uses the data to compute estimates of average total firm revenue, which are then published at the end of each period and become valuable in the case of reforms. All activities of the statistical office are costless. The office's estimates are based on a random sample of $n \leq N$ firms, where n is determined before the start of the economy. The ratio $p \equiv n/N$ is an obvious measure of *technical statistical capacity*. With this, the office's estimate of average total firm revenue in period t can be written as

$$Z_t^p \equiv \frac{1}{pN} \sum_{i=1}^{pN} z_{it}^* = \left(\frac{1}{pN} \sum_{i=1}^{pN} y_{it}^* \right) + \zeta_t^p, \quad (5)$$

where

$$\zeta_t^p \equiv \frac{1}{pN} \sum_{i=1}^{pN} \zeta_{it}. \quad (6)$$

Assuming that pN is sufficiently large, the Lindeberg-Lévy CLT implies that the distribution of $\sqrt{pN}(\zeta_t^p - \chi)$ is closely approximated by $N(0, \theta N)$. We therefore work with

$$\zeta_t^p \sim N(\chi, \theta/p). \quad (7)$$

If the government has implemented a reform policy in period 1 ($P_1 \neq 0$), it has to rely on Z_1^p as the only available source of information when it determines its second-period policy at the very beginning of that period.¹³ The government understands the firms' decision problem (and hence can infer their first-period investments, x_{i1}^*) as well as the parameters of the side businesses. Based on this, and with the help of Z_1^p , it uses Bayes' rule to compute the ex-post probability that the implemented first-period reform policy is the beneficial one:

$$r(Z_1^p) \equiv \Pr[P_1 = S | Z_1^p]. \quad (8)$$

In Section 6, we sketch a modified version of the model in which the data firms report to the statistical office is not necessarily fully accurate. Misreporting becomes an option as firms can shift parts of their core business into informal operations whose output cannot be detected.

¹³There is a chance that the realization of ζ_{i1} is so close to 0 so that $y_{i1}^*|_{P_1 \neq S} + \zeta_{i1} < y_{i1}^*|_{P_1 = S}$, in which case firm i —and the statistical office if i is sampled—can infer S in period 1. But since we assume that the distribution of ζ_{i1} has a light left tail, this chance is so small that it can be ignored in our analysis.

3.3 The bureaucracy and the government

While officially firm-level data collected by the statistical office may not be used for any purpose other than output estimation, there is a potential for data misuse: in both periods, there is a probability $\pi > 0$ of a confidentiality breach that puts the data into the hands of a “corrupt” government official. Following Svensson (2003) and Fisman and Svensson (2007), we may think of an official outside the statistical office whose power to collect bribes derives from his discretion in the application and enforcement of complex regulations. In practice, there is more than one way by which the official can get hold of confidential firm data (Section 2.2). Just like any hacker from outside the government bureaucracy, the official may take advantage of insufficient cyber defenses. Alternatively, the official may obtain access to the firm data due to an illicit practice of expansive data sharing within the government bureaucracy.

In many instances, corruption is governed by norms (see, e.g., Malesky and Samphantharak 2008, della Porta and Vannucci 2012, World Bank 2015). While authorizing corruption, these norms sometimes include restrictions, a sense of what would be an official’s customary share and what would amount to excessive, “intolerable” extortion. For the model, assume a norm that tolerates bribe collection as long as it is not out of proportion with a target’s economic capacity. Such a norm may be rooted in the experience that the economic damage associated with bribe collection beyond a certain relative limit causes civil unrest, putting in danger the government bureaucracy as a whole. For concreteness, assume that the corrupt official is protected by the responsible superior (e.g., a high-ranking bureaucrat) as long as a collected bribe does not exceed a share $0 < \hat{\beta} < 1$ of the target firm’s total revenue; for bribes in excess of this limit, if credibly reported to the superior, the norm demands that the corrupt official face a severe penalty (i.e., one that is large relative to the bribe collected). Further assume that firms can, and also do, credibly report excessive bribe collection.

In the model, $\hat{\beta}$ is one of two parameters that determine the scope for corruption. The second parameter is the probability of a confidentiality breach, π . In what follows, we use $\beta \equiv \pi \hat{\beta} \in (0, 1)$ as an overall measure of corruption and refer to it as the bureaucracy’s *vulnerability to corruption*. With the help of β , the model is able to capture a broad spectrum of institutional realities. If β takes a relatively low value, we have a bureaucracy whose integrity is merely compromised. By contrast, a value close to 1 may capture a setting in which an illicit practice of data sharing, combined with a norm authorizing vast bribe collection, enables an extractive bureaucracy that is hardly bound to the rule of law at all.

Regardless of the concrete level of β , leaked firm data offer valuable information to the corrupt official. Eliminating the informational asymmetry between the official and the firms in the sample, the data permit bribe discrimination—i.e., tailored bribes that extract the maximum possible without any risk of sanctions. By contrast, collecting bribes from firms that are not part of the sample is risky. Without information on firm revenue, collecting a bribe entails the possibility of a penalty. The official considers collecting a uniform bribe from those firms, weighing the

penalty to be expected as a function of the bribe level. Since a uniform (“lump-sum”) bribe does not affect firms’ incentives, we assume without loss of generality that the outcome of this optimization process is a bribe of zero. In summary, in the case of a confidentiality breach in period t , the official collects $\hat{\beta} z_{kt}$ from all firms k that are part of that period’s sample and leaves the remaining firms alone.¹⁴ Without a confidentiality breach in t , the official abstains from collecting bribes in that period.

The government determines statistical capacity, p , and is in charge of economic policy, P_t , $t \in \{1, 2\}$ (mind the difference between lower- and upper-case letters). We consider a benevolent government that, however, has inherited a bureaucracy whose vulnerability to corruption is deep-rooted and thus cannot be reduced within the time horizon of the model (i.e., must be taken as exogenous). Naturally, this assumption entails that β is unaffected by both p and P_t . Put differently, a possible improvement in statistical capacity, or a deviation from the status-quo economic policy, would not help in addressing problems with data confidentiality or corruption.¹⁵ The government’s objective is to maximize the expected lifetime total revenue of the representative firm i and its objective function reads

$$V = E_1 \{ z_{i1}^* + z_{i2}^* \}, \quad (9)$$

where the expectation in Eq. (9) is formed at the beginning of period 1. So our analysis of optimal technical statistical capacity assumes a relatively “favorable” environment in which the government does not pursue special interests but aims at maximizing economic performance and in which running the statistical office is free of charge.

3.4 Time line

The timing of actions is as follows. Prior to the start of the economy, Nature determines the unobserved state of the world S and the government chooses statistical capacity p .

In the first period, the government sets P_1 ; observing the government’s decision, all firms i choose x_{i1} ; the statistical office draws the random firm sample, thereby complying with the government’s choice of p ; Nature determines $\{\zeta_{i1}\}_{i=1}^N$; the statistical office collects the data and—if there is a confidentiality breach—the official collects bribes from the sampled firms; the statistical office publishes Z_1^p and—if $P_1 \neq 0$ —the government computes $r(Z_1^p)$.

¹⁴Proportional bribe collection requires that the official has the power to wrest larger bribes from larger firms. This is plausible since the cost of hold-ups caused by the official arguably rises in the scale of production.

¹⁵The literature (e.g., Pitlik et al. 2010) suggests that information is important in fighting corruption (e.g., as it helps monitor public spending and identify misuses of funds). Here, it is still plausible to assume that $\hat{\beta}$ is unaffected by p : the statistical office collects information on private firms and not on the government sector.

In the second period, the government sets P_2 ; all firms learn (y_{i1}^*, ζ_{i1}) and—if $P_1 \neq 0$ —infer state S ; taking P_2 and the available information on S into account, all firms i choose x_{i2} . From this point onwards, the sequence of actions is identical to that in the first period.

4 Equilibrium economic policy

4.1 Input choice

Before going backwards through the sequence of policy choices, we consider the firms' investment decisions. In period $t \in \{1, 2\}$, firm i solves the maximization problem

$$\max_{\{x_{it}\}} \left\{ p \left[\pi(1 - \hat{\beta}) E_t^i \{y(A_t, x_{it}) + \zeta_{it}\} + (1 - \pi) E_t^i \{y(A_t, x_{it}) + \zeta_{it}\} \right] \right. \\ \left. + (1 - p) E_t^i \{y(A_t, x_{it}) + \zeta_{it}\} - x_{it} \right\}, \quad (10)$$

where $E_t^i\{\cdot\}$ refers to the expectation formed by the firm just before it chooses x_{it} . The objective function in problem (10) reflects that, with probability p , firm i is sampled by the statistical office, in which case there is a chance π that it will be approached by a bribe-collecting official who would take a share $\hat{\beta}$ of total revenue. Problem (10) can be simplified to

$$\max_{\{x_{it}\}} \left\{ (1 - p\beta) \left(E_t^i \{A_t\} x_{it}^\alpha + \chi \right) - x_{it} \right\}. \quad (11)$$

Since $0 < \alpha < 1$, the objective function in maximization problem (11) is a strictly concave function of $x_{it} \in [0, \infty)$. The function's maximizer is given by

$$x_{it}^* = \left[\alpha (1 - p\beta) E_t^i \{A_t\} \right]^{1/(1-\alpha)}. \quad (12)$$

Using production function (1), one can calculate firm i 's output of the homogeneous good as

$$y_{it}^* = A_t \left[\alpha (1 - p\beta) E_t^i \{A_t\} \right]^{\alpha/(1-\alpha)}. \quad (13)$$

4.2 Second period

The final decisions of interest to be taken in period 2 are those by the firms on second-period investment. Those decisions are mostly analyzed in Section 4.1. What remains to be done here is to determine $E_2^i \{A_2\}$. When deciding on x_{i2} , firm i has just learned about (y_{i1}^*, ζ_{i1}) . As a result, if $P_1 \neq 0$, the firm can infer $S \in \{-1, 1\}$ from (y_{i1}^*, ζ_{i1}) ; moreover, having observed P_2 , it can identify A_2 with certainty (Eq. 2). Otherwise, if $P_1 = 0$, firms still reckon that S has taken the value -1 with probability $1/2$ and the value 1 with the same probability; thus, $E_2^i \{A_2\} = 1$ irrespective of the choice of

P_2 . To summarize:

$$E_2^i \{A_2\} = \begin{cases} 1 & : P_1 = 0 \\ A_2 & : P_1 \neq 0 \end{cases}. \quad (14)$$

The first decision to be taken in period 2 is that by the government on second-period policy, $P_2 \in \{-1, 0, 1\}$. Objective function (9) implies that, at this point in time, the government wants to maximize $E_2 \{z_{i2}^*\}$, where the expectation is formed at the beginning of period 2. First assume that the status-quo policy has been implemented in period 1 ($P_1 = 0$). For this case, Eq. (14) implies $E_2^i \{A_2\} = 1$. We therefore obtain

$$E_2 \{z_{i2}^*(P_2)\} = E_2 \{A_2(P_2)\} [\alpha (1 - p\beta)]^{\alpha/(1-\alpha)} + \chi. \quad (15)$$

Again, since there is no information on the realization of S in this case, we have $E_2 \{A_2(P_2)\} = 1$ for all $P_2 \in \{-1, 0, 1\}$. So the government is indifferent between the three options. Without loss of generality, we henceforth assume that it decides to keep the status-quo policy in place:

$$P_2|_{P_1=0} = 0. \quad (16)$$

Now suppose that a reform policy has been implemented in period 1 ($P_1 \neq 0$). In this case, taking into account Eq. (13) and Eq. (14), we obtain

$$E_2 \{z_{i2}^*(P_2)\} = E_2 \left\{ [A_2(P_2)]^{1/(1-\alpha)} \right\} [\alpha (1 - p\beta)]^{\alpha/(1-\alpha)} + \chi. \quad (17)$$

The expectation in Eq. (17) is now based on $r(Z_1^P)$, the ex-post probability that the first-period reform alternative is the beneficial one. Therefore:

Proposition 1 Suppose $P_1 \neq 0$. Then, in order to maximize $E_2 \{z_{i2}^*(P_2)\}$, the government chooses P_2 according to

$$P_2|_{P_1 \neq 0} = \begin{cases} P_1 & : r(Z_1^n) \geq 1/2 \\ -P_1 & : r(Z_1^n) < 1/2 \end{cases}. \quad (18)$$

Proof See Appendix A. □

4.3 First period

If a reform policy has been implemented, the final activity in period 1 is the computation of the ex-post probability $r(Z_1^P) \equiv \Pr[P_1 = S | Z_1^P]$. When doing this computation, the government considers the firms' investments earlier in the period and the resulting implications for the output of the homogeneous good. The government knows that the firms have solved the maximization problem stated in Section 4.1 and accordingly that the level of output is as specified in Eq. (13). Moreover, it follows that $E_1^i \{A_1\} = 1$ since state S takes each of its two possible values with probability $1/2$. So it is clear that

$$y_{i1}^* = A_1 [\alpha (1 - p\beta)]^{\alpha/(1-\alpha)}. \quad (19)$$

Given this, and considering Eq. (5) and Eq. (7), the government understands that Z_1^p follows a normal distribution with a mean that depends on S and P_1 :

$$Z_1^p \sim N \left\{ A_1 [\alpha (1 - p\beta)]^{\alpha/(1-\alpha)} + \chi, \theta/p \right\}, \quad (20)$$

where $A_1 = 1 + \sqrt{\gamma}$ if $P_1 = S$ and $A_1 = 1 - \sqrt{\gamma}$ if $P_1 = -S$. Since each of these two possibilities materializes with probability $1/2$, Bayes' rule implies

$$r(Z_1^p) \equiv \frac{(1/2) \cdot f(Z_1^p | P_1 = S)}{(1/2) \cdot f(Z_1^p | P_1 = S) + (1/2) \cdot f(Z_1^p | P_1 = -S)}, \quad (21)$$

where $f(Z_1^p | \cdot)$ denotes the corresponding normal density. Using functional forms, we obtain

$$r(Z_1^p) = \left\{ 1 + \exp \left[2 \frac{\sqrt{\gamma}}{\sigma/n} (\hat{Z}_1^p - \chi) (\hat{Z}_1^p - Z_1^p) \right] \right\}^{-1}, \quad (22)$$

where

$$\hat{Z}_1^p = [\alpha (1 - p\beta)]^{\alpha/(1-\alpha)} + \chi. \quad (23)$$

According to Eq. (22), $r(Z_1^p)$ is a strictly increasing function of Z_1^p , rising from 0 (if $Z_1^p \rightarrow -\infty$) to $1/2$ (if $Z_1^p = \hat{Z}_1^p$) to 1 (if $Z_1^p \rightarrow \infty$). In combination with Proposition 1, this implies that $P_2 = P_1$ if $Z_1^p \geq \hat{Z}_1^p$ and $P_2 = -P_1$ otherwise.

Moving backwards to the firms' actual investment decisions, it is sufficient to refer to the above discussion and repeat that x_{i1}^* and y_{i1}^* are given by Eqs. (12) and (13), respectively, with $E_1^i \{A_1\} = 1$ irrespective of the actual policy decision.

The first decision to be taken in period 1 is that by the government on first-period policy. To inform this decision, the government compares the value of its objective function, $V = E_1 \{z_{i1}^* + z_{i2}^*\}$, under the status quo to the value under any of the two reform alternatives. According to Eq. (16), $P_1 = 0$ implies $P_2 = 0$. So, if the government opts for the status quo in period 1 ($P_1 = 0$), we obtain $A_1 = A_2 = 1$. From this, it follows that the expected lifetime total revenue by the representative firm i is given by

$$E_1 \{z_{i1}^* + z_{i2}^* | P_1 = 0\} = 2 [\alpha (1 - p\beta)]^{\alpha/(1-\alpha)} + 2\chi. \quad (24)$$

Due to the symmetric setup, the government is indifferent between the two reform alternatives. Without loss of generality, we henceforth assume that $P_1 = 1$ if the government decides to abandon the status quo. In this case, P_2 is specified by Eq. (18) and the expected lifetime total revenue by the representative firm i is given by

$$E_1 \{z_{i1}^* + z_{i2}^* | P_1 = 1\} = \left[1 + E_1 \{[A_2(P_2)]^{1/(1-\alpha)}\} \right] [\alpha (1 - p\beta)]^{\alpha/(1-\alpha)} + 2\chi. \quad (25)$$

Moreover:

Lemma 1 Suppose the government opts for a reform policy in period 1 (e.g., $P_1 = 1$). Then,

$$E_1 \{[A_2(P_2)]^{1/(1-\alpha)}\} = \hat{A}^l + \Pr[P_2 = S] (\hat{A}^h - \hat{A}^l) > 1, \quad (26)$$

where $\Pr[P_2 = S]$ denotes the chance that in period 2 the beneficial reform policy is chosen and

$$\hat{A}^l \equiv (1 - \sqrt{\gamma})^{1/(1-\alpha)} \quad \text{and} \quad \hat{A}^h \equiv (1 + \sqrt{\gamma})^{1/(1-\alpha)}. \quad (27)$$

Proof See Appendix A. □

The results derived so far lead to the following conclusion:

Proposition 2 *In period 1, the government prefers reform (e.g., $P_1 = 1$) to the status quo. In period 2, the government's policy choice is described by Eq. (18).*

Proof The first statement of the proposition follows from Eqs. (24) and (25) and Lemma 1. The second statement follows from the first and Proposition 1. □

In period 1, there are two factors that make the government prefer a reform policy to the status quo. First, if a reform policy is implemented, the government gains information about “what works”; this, in turn, allows for a better informed policy decision in period 2. Second, y_{i2}^* is convex in A_2 (see Eqs. 13 and 14). For this reason, the government prefers taking a “symmetric risk” to obtaining the expected value with certainty.

5 Statistical capacity

5.1 Optimal statistical capacity

Besides economic policy, the government determines technical statistical capacity, p , with a view to maximizing V . The key magnitude in its decision problem is the probability with which the beneficial reform will be implemented eventually (Eqs. 25) and (26).

Proposition 3 *At the beginning of period 1, i.e., at the moment when the government decides to implement a reform policy (Proposition 2), the probability that the implemented second-period reform is in fact the beneficial one, $\Pr[P_2 = S]$, is given by*

$$I(C; \gamma, \theta) \equiv \Phi\left(\sqrt{\gamma/\theta} \cdot C(p; \alpha, \beta)\right) \quad (28)$$

where $\Phi(\cdot)$ denotes the distribution function of the standard normal distribution and

$$C(p; \alpha, \beta) \equiv \sqrt{p} [\alpha (1 - p\beta)]^{\alpha/(1-\alpha)} > 0. \quad (29)$$

Proof See Appendix A. □

In what follows, we will call I the “informativeness of policy experimentation”.¹⁶ As can be seen from Eq. (28), informativeness depends on five parameters, some of them unrelated to the statistical office: other things equal, if a reform is more significant (higher γ), or if the side businesses are less volatile (lower θ), informativeness is higher. On the other hand, informativeness is influenced by the statistical office. In Eq. (28), $C(p; \alpha, \beta)$ captures the entirety of channels by which the statistical office affects informativeness. For this reason, we will call $C(p; \alpha, \beta)$ a measure of “comprehensive statistical capacity”. It is immediately apparent that the effect of technical statistical capacity on comprehensive statistical capacity is ambiguous if the bureaucracy’s vulnerability to corruption is not equal to zero ($\beta > 0$). This reflects that firms—observing a positive relationship between p and expected bribe demands—reduce investment in response to a rise in p (Eq. 12). Note that α affects $C(p; \alpha, \beta)$ because this parameter governs the elasticity of investment with respect to expected bribe demands. $C(p; \alpha, \beta)$ has the following important properties:

Lemma 2 *Comprehensive statistical capacity $C(p; \alpha, \beta)$ is a function of p on $[0, 1]$ that has a unique maximizer, $p^* \in (0, 1]$. Moreover, $C(p; \alpha, \beta)$ is strictly concave on $[0, p^*)$.*

Proof See Appendix A. □

What level of technical statistical capacity maximizes comprehensive capacity? The answer depends on the bureaucracy’s vulnerability to corruption:

Proposition 4 *If the bureaucracy is sufficiently vulnerable to corruption, the level of technical statistical capacity that maximizes comprehensive statistical capacity $C(p; \alpha, \beta)$ —and hence informativeness $I(C; \gamma, \theta)$ —is strictly less than 1. In formal terms: if*

$$\beta > \frac{1 - \alpha}{1 + \alpha}, \quad (\text{R1})$$

we obtain

$$p^* = \frac{1 - \alpha}{1 + \alpha} \frac{1}{\beta} < 1, \quad (30)$$

where we use p^ to denote the maximizer of $C(p; \alpha, \beta)$.*

Proof See Appendix A. □

A rise in technical statistical capacity has two opposing effects on comprehensive statistical capacity and hence on the extent to which the estimate of average total firm revenue, Z_1^P , is informative for the policy decision in period 2. On the positive side, a rise in technical statistical capacity reduces the variance of the exogenous component of Z_1^P (Eq. 7); all else equal, this helps informativeness. On the negative

¹⁶The complementary probability to $\Pr[P_2 = S]$ is equal to the sum of the probabilities of a type-I error ($P_1 = S$ is rejected although true) and of a type-II error ($P_1 = S$ is not rejected although false).

Fig. 5 Informativeness and technical statistical capacity ($P_1 = S$). *Note:* Both subfigures show distributions of $Z_1^p - \hat{Z}_1^p$. The red distributions are associated with a higher level of technical statistical capacity, p_h , the blue distributions with a lower level, p_l . The government commits a type-I error, i.e., mistakenly switches to the alternative reform policy, if $Z_1^p - \hat{Z}_1^p < 0$. In Subfigure (a), moving from p_l to p_h reduces the probability of a type-I error; in Subfigure (b), the chance of a type-I error rises

side, for any individual firm, a rise in technical statistical capacity implies a higher chance of being subjected to bribe collection. As a result, firms respond by reducing investment (Eq. 12)—which dampens the impact of reforms on production; all else equal, this harms informativeness. The strength of the negative effect rises in the bureaucracy's vulnerability to corruption. If β is sufficiently large, the negative effect starts to dominate at a strictly interior level of technical statistical capacity.

Figure 5 illustrates the ambiguous role of technical statistical capacity. In both subfigures, the underlying assumption is that the implemented first-period reform is the beneficial one ($P_1 = S$); moreover, in both subfigures, the shaded areas correspond to the probabilities of committing a type-I error ($P_1 = S$ rejected although true) under low (blue) and high (red) technical statistical capacity, respectively. Subfigure (a) shows a case in which higher technical statistical capacity (red line) improves informativeness: an increase in p from p_l to p_h reduces the probability of a type-I error. In other words, the reduction in the variance of the exogenous component of Z_1^p outweighs the reduction in the beneficial effect of the reform (leftward shift of the entire distribution). By contrast, Subfigure (b) shows a situation in which an increase in technical statistical capacity leads to a higher probability of a type-I error: the reduction in the variance of the distribution is dominated by its shift to the left.

However, informativeness—and the associated probability of choosing the beneficial reform alternative in period 2—is not the only variable the government has to consider when determining the level of p that maximizes V , the expected lifetime total firm revenue. The negative static effect of technical statistical capacity on firm investment (and hence production) must be taken into account, too. We therefore obtain the following result:

Proposition 5 *Consider a benevolent government that aims at maximizing economic performance by experimenting with a set of policies and that can implement any level of technical statistical capacity at no cost. Yet assume that the government faces*

a bureaucracy that is sufficiently vulnerable to corruption such that condition (R1) holds.

When choosing the level of technical statistical capacity, this government neither opts for 1 (maximum level) nor for $p^* < 1$ (the level that maximizes comprehensive statistical capacity and hence the informativeness of policy experimentation). Instead, the government finds it optimal to choose a level p^{**} that is (substantially) lower than p^* :

$$p^{**} < p^* = \frac{1 - \alpha}{1 + \alpha} \frac{1}{\beta} < 1. \quad (31)$$

Proof See Appendix A. □

Figure 6 illustrates how $I(C; \gamma, \theta)$ and $V = E_1 \{z_{i1}^* + z_{i2}^*\}$ depend on technical statistical capacity, assuming that condition (R1) holds. Both curves are hump-shaped. Proposition 5 predicts that V peaks at a strictly lower level of technical statistical capacity than C does. As a result, the peak of I must lie to the right of the peak of V .

In Fig. 6, p^{**} takes a particularly low value. So technical statistical capacity is weak, as is the case in many developing countries. The figure thus conveys the message that real-world instances of low technical statistical capacity should be interpreted with care. They are compatible with completely different versions of reality. Clearly, “poor numbers” may be an expression of neglect and/or a lack of resources and expertise. However, keeping technical statistical capacity weak may also be the best response by a capable government that is mindful of facing a corrupt bureaucracy and problems with data confidentiality like those illustrated in Section 2.2. “Flying blind” may be a well-considered choice.

5.2 The effect of corruption

An increase in the bureaucracy’s vulnerability to corruption amplifies the negative effect of statistical capacity on firm investment. As a result, the level of technical statistical capacity that maximizes comprehensive statistical capacity and informativeness, p^* , is a decreasing function of β (Eq. 30). Simulations suggest that a similar

Fig. 6 Technical statistical capacity, informativeness, and economic performance. *Note:* For the simulation we have chosen the following parameter values: $\alpha = 0.7$, $\beta = 0.3$, $\gamma = 0.5$

result holds for the level of statistical capacity that maximizes the expected lifetime total firm revenue, p^{**} . Figure 7 illustrates that p^{**} shifts to the left as β increases.

Fig. 7 Effect of vulnerability to corruption on economic performance. *Note:* For the simulation we have chosen the following parameter values: $\alpha = 0.7$, $\gamma = 0.5$

Consistent with this, we find that a rise in β has negative consequences for the two outcomes of interest:

Proposition 6 *Suppose that the bureaucracy is sufficiently vulnerable to corruption so that condition (R1) holds. Then, (i) $C(p^{**}; \alpha, \beta)$ —and thus $I(C; \gamma, \theta)$ —are strictly decreasing functions of β ; (ii) $V = E_1 \{z_{i1}^* + z_{i2}^*\}$ is a strictly decreasing function of β .*

Proof See Appendix A. □

There is a large literature on the channels by which corruption affects economic performance. Proposition 6 puts the spotlight on a novel one. An increase in the bureaucracy's vulnerability to corruption reduces the informativeness of policy experiments (a process that is understood to involve the implementation, evaluation, and adjustment of reform policies). Two mechanisms are simultaneously at work. On the one hand, an increase in corruption dampens the economy's response to policy changes (Eq. 13). On the other hand, as can be inferred from Fig. 7, higher corruption induces the government to lower the level of technical statistical capacity—which makes navigating among the different reform alternatives even more challenging. This is also reflected in the economy's growth rate:

Proposition 7 *Suppose that the bureaucracy is sufficiently vulnerable to corruption so that condition (R1) holds. Then, the expected growth rate due to policy experimentation,*

$$E_1 \left\{ \frac{y_{i2}^* - y_{i1}^*}{y_{i1}^*} \right\} = \frac{1}{1 - \gamma} \left\{ \hat{A}^l + I(C; \gamma, \theta) \left(\hat{A}^h - \hat{A}^l \right) \right\} - 1, \quad (32)$$

is a strictly decreasing function of β .

Proof Follows immediately from Proposition 6. □

So the consequences of an increase in corruption are not limited to a mere level effect; higher corruption also flattens the path of production over time.

6 Informality and misreporting

In principle, sampled firms have an incentive to underreport total revenue since they must anticipate the possibility of proportional bribe demands. However, so far, we have maintained the assumption that the firms nonetheless report truthfully, e.g., because any misreporting may be detected and penalized by the statistical office. While this is a strong assumption, we can relax it without changing the main insights from our analysis. In this section, we sketch a modified version of the model in which firms can shift parts of their production into informal operations whose output is perfectly hidden—and thus not reported to the statistical office. The exposition here

concentrates on the impact of exogenous changes to technical statistical capacity on firm investment and the informativeness of policy experimentation.

Consider the following modifications. First, in addition to the standard technology represented by production function (1), firms have access to a second technology by means of which the homogenous good can be produced. The output of the second technology, in contrast to that of the standard technology, cannot be observed by any other actor in the economy. Therefore, firms will not report it. It is natural to regard the use of the second technology as an informal activity. Thus, if a firm relies on the second technology in addition to the standard one, it runs formal as well as informal operations.¹⁷ Second, in each period, firms must decide on the allocation of a fixed amount of a resource (e.g., managerial time) across their formal and informal operations. In this context, we continue to use the wording “investment”: the part of the resource that goes to the formal operations is called formal investment, the rest is called informal investment. Finally, the government no longer perfectly understands the firms’ decision problem. Yet it knows the standard technology and observes how much firms invest formally.¹⁸ As a result, following a reform in the first period, the government can still compute $r(Z_1^p)$, the ex-post probability that the implemented reform is the beneficial one.

For concreteness, suppose the second technology is represented by the production function

$$y_{it}^n = Bx_{it}^n, \quad (33)$$

where $B > 0$ captures productivity and x_{it}^n and y_{it}^n refer to, respectively, informal investment and the resulting informal output of the homogeneous good. Independence from policy is assumed for simplicity only. We use $\bar{x} > 0$ to denote the fixed amount of the resource to be allocated. Note that $x_{it}^n + x_{it}^f = \bar{x}$, where in this section x_{it}^f stands for formal investment (and y_{it}^f stands for the formal output that results via production function (1)). Parameter restriction

$$[\alpha(1 + \sqrt{\gamma})/B]^{1/(1-\alpha)} < \bar{x} \quad (34)$$

is sufficient to ensure that firms always run formal and informal operations in parallel. In what follows, we assume that restriction (34) holds. Moreover, without loss of generality, we impose $B = 1$. With these assumptions, the firms solve the following maximization problem:

$$\max_{\{x_{it}^f\}} \left\{ (1 - p\beta) \left(E_t^i \{A_t\} \left(x_{it}^f \right)^\alpha + \chi \right) + \left(\bar{x} - x_{it}^f \right) \right\}. \quad (11')$$

When we compare problem (11), the maximization problem in the baseline model, to problem (11'), it is immediately clear that the two objective functions are identical except for constant $\bar{x}$ in Eq. (11'). As a result, the two problems have identical

¹⁷Note that the exogenous contributions of the side businesses to total firm revenues remain formal. In practice, it is not uncommon for formal firms to have an informal “leg”. According to the ILO (2018), in many world regions, a significant share of informal workers are employed by formal firms.

¹⁸As in the baseline model, formal investment will be uniform across firms. We can assume that the statistical office collects data on formal investment in addition to the data on formal revenue.

solutions: $x_{it}^{f*} = x_{it}^*$, where x_{it}^* is given by Eq. (12). Restriction (34) guarantees that $x_{it}^{n*} = \bar{x} - x_{it}^{f*}$ is strictly positive.

How do the results regarding the impact of technical statistical capacity, p , differ between the baseline and the modified model? With regard to investment, there are similarities as well as differences. The similarity is that formal investment is still a monotonically decreasing function of p ; a higher chance of being subjected to proportional bribe collection continues to have a disincentivizing effect. The difference is that in the modified model total investment is unaffected by p ; a rise in p merely shifts investment into informal operations, away from formal ones. So the fall in output coming from formal operations is accompanied by a rise in informal output. In net terms, there is still a loss since the misallocation of investment across operations worsens.¹⁹ However, for identical parameters, the negative effect of a given increase in p on total output of the homogeneous good is smaller than in the baseline model.

With regard to the informativeness of policy experimentation, there is no difference between the baseline and the modified model. The negative effect of p on formal investment continues to dampen the impact of reform policies. In fact, the level of p that maximizes informativeness, p^* , is unchanged (Eq. 30). Yet what level of p should be chosen in presence of informal operations? If the government were interested in formal-operations output only, Proposition 5 would continue to apply: the government would choose a p^{**} that is markedly smaller than p^* . But if the government factors in that losses in formal output are partly compensated by gains in informal output (which, however, it does not exactly observe), it may want to push p closer to p^* . By how much must depend, among other things, on the government's attitude towards informality. Regardless of this, we can summarize that the main insights from the baseline model are robust to the introduction of informality and misreporting.

7 Summary and conclusion

One approach to boost economic output is policy experimentation: tweak an existing economic policy, evaluate the consequences, and so discover “what works”. While economists have been interested in this approach for a long time, it has recently gained traction in the context of development policy. Clearly, how much policy makers can learn from a policy experiment—i.e., the degree of the experiment's informativeness—must depend on the accuracy with which the economy is measured. In developing countries, accuracy tends to be low. As a result, efforts to help developing countries improve their statistical capacities are of great importance. In this paper, we demonstrate that efforts with a sole focus on technical aspects—here understood as the scale of data gathering—need not be unambiguously helpful. The reason is a type of observer effect: raising technical statistical capacity to improve the measurement of the economy actually changes what is being measured. Why? If

¹⁹By reducing formal investment x_{it}^{f*} , a rise in p increases the wedge between $\alpha E_t^i \{A_t\} (x_{it}^{f*})^{\alpha-1}$ and B , the (expected) marginal products of formal and informal investments, respectively.

confidentiality breaches are to be expected and control of corruption is weak, more information gathering by the statistical office means that firms face a higher risk of being subjected to bribe collection. As a result, firms reduce their investments, thereby dampening the effect of policy changes. The relative strength of this harmful effect rises in the level of technical statistical capacity. At some point, it becomes the dominant force, implying that further improvements in technical statistical capacity make it harder—rather than easier—to discover “what works”.

Against this background, we argue that efforts with the aim of expanding data gathering should not be uniform but adapted to local circumstances. Broadly speaking, our analysis suggests that such efforts be aligned with the local institutional setting: in places where extensive bribe collection is the norm and data confidentiality in doubt, the extent of data gathering—and hence the precision with which the economy is measured—should be more limited. Viewed from a different angle, our analysis shows that efforts to address insufficient technical statistical capacity should come with additional arrangements. In particular, taking measures to reinforce data confidentiality would dampen the observer effect and so permit the analysis of a less distorted economy, including the economy’s response to policy changes. A list of such measures would certainly include the strengthening of cyber defenses. Yet, when dealing with an extractive bureaucracy in which a practice of expansive data sharing has taken root, institutional changes might be needed, too. They could range from fortifying the statistical system’s independence from the rest of the bureaucracy to more robust measures such as the outright outsourcing of statistical services to an independent international body. Finally, we believe that this paper may contribute to a better understanding of the intricate ways through which corruption retards economic growth. By stifling private economic activity, and by limiting the optimal degree of measurement precision, corruption makes it harder for policy makers to discover how to tailor economic policies to local circumstances.

Appendix A: Proofs

Propositions

Proof of Proposition 1 Obviously, the maximizer of $E_2\{[A_2(P_2)]^{1/(1-\alpha)}\}$ maximizes the expected second-period total revenue by the representative firm, $E_2\{z_{i2}^*(P_2)\}$. Together, Eq. (2) and the definition of $r(Z_1^n)$, stated in Eq. (8), imply

$$E_2\left\{[A_2(P_2)]^{\frac{1}{1-\alpha}}\right\} = \begin{cases} r(Z_1^n)(1 + \sqrt{\gamma})^{\frac{1}{1-\alpha}} + [1 - r(Z_1^n)](1 - \sqrt{\gamma})^{\frac{1}{1-\alpha}} & : P_2 = P_1 \\ 1 & : P_2 = 0 \\ r(Z_1^n)(1 - \sqrt{\gamma})^{\frac{1}{1-\alpha}} + [1 - r(Z_1^n)](1 + \sqrt{\gamma})^{\frac{1}{1-\alpha}} & : P_2 = -P_1 \end{cases} \quad (35)$$

Note that $E_2\{[A_2(P_1)]^{1/(1-\alpha)}\} \geq E_2\{[A_2(-P_1)]^{1/(1-\alpha)}\}$ if and only if $r(Z_1^n) \geq 1/2$. Moreover, because $[A_2]^{1/(1-\alpha)}$ is a strictly convex function of A_2 , we have $E_2\{[A_2(P_1)]^{1/(1-\alpha)}\} > 1$ if $r(Z_1^n) \geq 1/2$ and $E_2\{[A_2(-P_1)]^{1/(1-\alpha)}\} > 1$ if

$r(Z_1^n) \leq 1/2$. As a result, the government will set $P_2 = P_1$ if $r(Z_1^n) \geq 1/2$ and $P_2 = -P_1$ if $r(Z_1^n) < 1/2$. $\square$

Proof of Proposition 3 The probability that the beneficial reform policy is implemented in period 2 is composed of the probability that $r(Z_1^P) \geq 1/2$ if $P_1 = S$ and the probability that $r(Z_1^P) < 1/2$ if $P_1 = -S$. Equation (22) implies (i) that $r(Z_1^P) \geq 1/2$ is equivalent to $Z_1^P \geq \hat{Z}_1^P$; and (ii) that $r(Z_1^P) < 1/2$ is equivalent to $Z_1^P < \hat{Z}_1^P$. Thus,

$$\begin{aligned} \Pr[P_2 = S] &= (1/2) \Pr[Z_1^P \geq \hat{Z}_1^P | P_1 = S] + (1/2) \Pr[Z_1^P < \hat{Z}_1^P | P_1 = -S] \\ &= (1/2) \left\{ 1 - \Pr[Z_1^P < \hat{Z}_1^P | P_1 = S] + \Pr[Z_1^P < \hat{Z}_1^P | P_1 = -S] \right\}. \end{aligned} \quad (36)$$

Taking into account Eq. (2), Eq. (5), and Eq. (19), Eq. (36) can be rewritten as

$$\begin{aligned} \Pr[P_2 = S] &= (1/2) \left\{ 1 - \Pr \left[(1 + \sqrt{\gamma}) [\alpha (1 - p\beta)]^{\alpha/(1-\alpha)} + \zeta_1^P < \hat{Z}_1^P \right] \right. \\ &\quad \left. + (1/2) \Pr \left[(1 - \sqrt{\gamma}) [\alpha (1 - p\beta)]^{\alpha/(1-\alpha)} + \zeta_1^P < \hat{Z}_1^P \right] \right\}, \end{aligned} \quad (37)$$

where $\hat{Z}_1^P$ is given by Eq. (23). Rearranging terms yields

$$\begin{aligned} \Pr[P_2 = S] &= (1/2) \left\{ 1 - \Pr \left[\sqrt{p/\theta} (\zeta_1^P - \chi) < -\sqrt{\gamma/\theta} \cdot C \right] \right\} \\ &\quad + (1/2) \Pr \left[\sqrt{p/\theta} (\zeta_1^P - \chi) < \sqrt{\gamma/\theta} \cdot C \right]. \end{aligned} \quad (38)$$

Since $\sqrt{p/\theta} (\zeta_1^P - \chi)$ follows a standard normal distribution, Eq. (38) is equivalent to

$$\Pr[P_2 = S] = (1/2) \left[1 - \Phi(-\sqrt{\gamma/\theta} \cdot C) + \Phi(\sqrt{\gamma/\theta} \cdot C) \right]. \quad (39)$$

Because the standard normal distribution is symmetric around zero, Eq. (39) simplifies to the expression given in the proposition. $\square$

Proof of Proposition 4 As $\Phi(\sqrt{\gamma/\theta} \cdot C)$ is a strictly increasing function of $C(p; \alpha, \beta)$, the maximizer of C , p^* , is also the maximizer of $I = \Pr[P_2 = S]$. From Eq. (51), it follows that $\partial C(1; \alpha, \beta)/\partial p < 0$ if condition (R1) is satisfied. Thus, according to Lemma 2, the unique maximizer p^* is strictly less than 1 and hence must be pinned down by the first-order condition $\partial C(p^*; \alpha, \beta)/\partial p = 0$. Equation (51) suggests that this condition is equivalent to

$$\frac{1}{2} \frac{(1 - p^*\beta)}{\sqrt{p^*}} - \frac{\alpha}{1 - \alpha} \frac{\beta}{N} \sqrt{p^*} = 0. \quad (40)$$

Rearranging terms yields the expression stated in the proposition. $\square$

Proof of Proposition 5 As the government opts for a reform policy in period 1 ($P_1 = 1$), $V(p; \cdot)$ is given by Eq. (25). Therefore,

$$\begin{aligned} \frac{\partial V}{\partial p} = & \phi \left(\sqrt{\gamma/\theta} \cdot C \right) \frac{\partial C}{\partial p} \sqrt{\frac{\gamma}{\theta}} \left(\hat{A}^h - \hat{A}^l \right) [\alpha (1 - p\beta)]^{\alpha/(1-\alpha)} \\ & - \alpha\beta \frac{\alpha}{1-\alpha} \frac{1 + E_1 \left\{ A_2^{1/(1-\alpha)} \right\}}{[\alpha (1 - p\beta)]^{(1-2\alpha)/(1-\alpha)}} \end{aligned} \quad (41)$$

or, equivalently, by

$$\begin{aligned} \frac{\partial V}{\partial p} = & [\alpha (1 - p\beta)]^{(2\alpha-1)/(1-\alpha)} \\ & \cdot \left\{ \phi \left(\sqrt{\gamma/\theta} \cdot C \right) \frac{\partial C}{\partial p} \sqrt{\frac{\gamma}{\theta}} \left(\hat{A}^h - \hat{A}^l \right) [\alpha (1 - p\beta)] - \alpha\beta \frac{\alpha}{1-\alpha} \left[1 + E_1 \left\{ A_2^{1/(1-\alpha)} \right\} \right] \right\}. \end{aligned} \quad (42)$$

First suppose that $\alpha \leq 1/2$ and consider Eq.(41). As p rises from 0 to the maximizer of $C(p; \alpha, \beta)$, p^* , the first line of Eq.(41) decreases monotonically from infinity to 0 (see the properties of $C(p; \alpha, \beta)$ stated in Lemma 2); for values of $p \in (p^*, 1]$, the first line is strictly negative. The second line of Eq. (41) increases monotonically (in absolute terms) as p rises from 0 to p^* (see Eq. (26) and Proposition 3), and remains strictly negative when p exceeds p^* (and rises towards 1). Together, the behavior of the first and the second line implies that $\partial V/\partial p$ —as p rises from 0 to p^* —falls monotonically from infinity to a value that is strictly less than 0; for values of $p \in (p^*, 1]$, $\partial V/\partial p$ remains strictly negative. As a result, on $[0, 1]$, there exists a unique $p^{**} < p^*$ such that $\partial V(p^{**}; \cdot)/\partial p = 0$. From this, it immediately follows that p^{**} is the unique maximizer of V . Now suppose $\alpha > 1/2$ and consider Eq. (42). A similar chain of arguments implies that, again, on $[0, 1]$, there exists a unique $p^{**} < p^*$ such that $\partial V(p^{**}; \cdot)/\partial p = 0$. As in the case $\alpha \leq 1/2$, p^{**} is the unique maximizer of V . $\square$

Proof of Proposition 6 To prove (i), first assume that $dp^{**}/d\beta > 0$. In this case, we consider $\partial V/\partial p|_{p=p^{**}} = 0$, the first-order condition that implicitly pins down p^{**} . Using Eqs. (41) and (51), and rearranging terms, we obtain

$$\frac{\hat{A}^h - \hat{A}^l}{2} \frac{\phi(\sqrt{\gamma/\theta} \cdot C) \sqrt{\gamma/\theta} \cdot C}{1 + E_1\{A_2^{1/(1-\alpha)}\}} \left(\frac{1}{p^{**}} - \frac{1+\alpha}{1-\alpha} \beta \right) = \frac{\alpha}{1-\alpha} \beta. \quad (43)$$

Expressions for C and $E_1\{A_2^{1/(1-\alpha)}\}$ are given by Eqs. 29 and 26, respectively. Taking these into account, we can transform Eq. (43) into

$$U(C) \frac{\hat{A}^h - \hat{A}^l}{2} \sqrt{\frac{\gamma}{\theta}} [\alpha(1 - p^{**}\beta)]^{\alpha/(1-\alpha)} \left(\frac{1}{\sqrt{p^{**}}} - \frac{1+\alpha}{1-\alpha} \beta \sqrt{p^{**}} \right) = \frac{\alpha}{1-\alpha} \beta, \quad (44)$$

where

$$U(C) = \frac{\phi(\sqrt{\gamma/\theta} \cdot C)}{1 + \hat{A}^l + \Phi(\sqrt{\gamma/\theta} \cdot C) (\hat{A}^h - \hat{A}^l)}. \quad (45)$$

Now consider the impact of a rise in β on Eq. (44), holding constant C . Obviously, the right-hand side of the equation increases; at the same time, since we assume $dp^{**}/d\beta > 0$, the left-hand side decreases. As a result, $U(C)$ must rise. Since $C > 0$ and Φ and ϕ denote, respectively, the distribution function and the probability density function of the standard normal distribution, it follows that C must fall to restore the equality of the two sides of Eq. (44). So, if $dp^{**}/d\beta > 0$, we must conclude that $C(p^{**}; \alpha, \beta)$ is a strictly decreasing function of β . Now suppose that $dp^{**}/d\beta \leq 0$ and consider

$$\frac{dC(p^{**}; \alpha, \beta)}{d\beta} = \frac{\partial C(p; \alpha, \beta)}{\partial p} \bigg|_{p=p^{**}} \frac{dp^{**}}{d\beta} + \frac{\partial C(p; \alpha, \beta)}{\partial \beta} \bigg|_{p=p^{**}}. \quad (46)$$

Observing Eq. (29), we can infer that $\partial C(p; \alpha, \beta)/\partial \beta|_{p=p^{**}} < 0$. Because $p^{**} < p^*$, and since $C(p; \alpha, \beta)$ is a strictly increasing function of p on $[0, p^*]$, we have $\partial C(p; \alpha, \beta)/\partial p|_{p=p^{**}} > 0$. Finally, because $dp^{**}/d\beta \leq 0$ by assumption, it follows that $dC(p^{**}; \alpha, \beta)/d\beta < 0$. So, once again, we must conclude that $C(p^{**}; \alpha, \beta)$ is a strictly decreasing function of β .

To prove (ii), note that Eqs. (25) and (26) and Proposition 3 imply that in equilibrium

$$V = \left[1 + \hat{A}^l + \Phi(\sqrt{\gamma/\theta} \cdot C(p^{**}; \alpha, \beta)) (\hat{A}^h - \hat{A}^l) \right] [\alpha(1 - p^{**}\beta)]^{\alpha/(1-\alpha)} + 2\chi. \quad (47)$$

Obviously, $\partial V/\partial p|_{p=p^{**}} = 0$. Therefore,

$$\begin{aligned} \frac{dV}{d\beta} \bigg|_{p=p^{**}} &= \phi(\sqrt{\gamma/\theta} \cdot C) \sqrt{\gamma/\theta} \frac{\partial C}{\partial \beta} (\hat{A}^h - \hat{A}^l) [\alpha(1 - p^{**}\beta)]^{\alpha/(1-\alpha)} \\ &\quad - \left[1 + \hat{A}^l + \Phi(\sqrt{\gamma/\theta} \cdot C) (\hat{A}^h - \hat{A}^l) \right] \frac{\alpha}{1-\alpha} [\alpha(1 - p^{**}\beta)]^{\alpha/(1-\alpha)-1} \alpha p^{**}. \end{aligned} \quad (48)$$

Considering the definition of C given in Eq. (29), it follows immediately that $\partial C/\partial \beta < 0$. Given this, we conclude that $dV/d\beta|_{p=p^{**}} < 0$. $\square$

Lemmas

Proof of Lemma 1 In period 2, the government will either keep the first-period reform policy in place or switch to the alternative reform policy (Eq. 18). Given the definition of A_2 in Eq. (2), and that of $\Pr[P_2 = S]$ in the lemma, the agents expect $A_2 = 1 + \sqrt{\gamma}$ with probability $\Pr[P_2 = S]$ and $A_2 = 1 - \sqrt{\gamma}$ with probability $1 - \Pr[P_2 = S]$. Thus,

$$E_1 \left\{ [A_2(P_2)]^{1/(1-\alpha)} \right\} = \Pr[P_2 = S](1 + \sqrt{\gamma})^{1/(1-\alpha)} + \{1 - \Pr[P_2 = S]\} (1 - \sqrt{\gamma})^{1/(1-\alpha)}, \quad (49)$$

an expression that can be rearranged to give the expression in the lemma.

To see that $E_1\{[A_2(P_2)]^{1/(1-\alpha)}\} > 1$, note (i) that $[A_2]^{1/(1-\alpha)}$ is a strictly convex function of A_2 , implying that $(1/2)(1 + \sqrt{\gamma})^{1/(1-\alpha)} + (1/2)(1 - \sqrt{\gamma})^{1/(1-\alpha)} > 1$; and (ii) that $\Pr[P_2 = S]$ is strictly greater than $1/2$ since the observation of Z_1^p permits learning. $\square$

Proof of Lemma 2 The partial derivative of C with respect to p is given by

$$\frac{\partial C}{\partial p} = \frac{1}{2} \frac{[\alpha(1 - p\beta)]^{\alpha/(1-\alpha)}}{\sqrt{p}} - \alpha\beta \frac{\alpha}{1 - \alpha} \frac{\sqrt{p}}{[\alpha(1 - p\beta)]^{(1-2\alpha)/(1-\alpha)}} \quad (50)$$

or, equivalently, by

$$\frac{\partial C}{\partial p} = \alpha^{\alpha/(1-\alpha)} (1 - p\beta)^{(2\alpha-1)/(1-\alpha)} \left[\frac{1}{2} \frac{1 - p\beta}{\sqrt{p}} - \frac{\alpha}{1 - \alpha} \beta \sqrt{p} \right]. \quad (51)$$

First suppose that $\alpha \leq 1/2$. Then, Eq. (50) implies that $\lim_{p \rightarrow 0} \partial C / \partial p = \infty$ and that $\partial C / \partial p$ is a strictly decreasing function of p on $[0, 1]$. We thus conclude that C is strictly concave on $[0, 1]$ and has a unique maximizer, p^* , that is strictly greater than 0. Now suppose $\alpha > 1/2$. Then, Eq. (51) implies that $\lim_{p \rightarrow 0} \partial C / \partial p = \infty$ and that—as p increases from 0— $\partial C / \partial p$ is a strictly decreasing function of p as long as $\partial C / \partial p \geq 0$. Equation (51) further suggests that $\partial C / \partial p$ can have at most one root on $[0, 1]$. Thus, again, C has a unique maximizer $p^* > 0$; moreover, it is strictly concave on $[0, p^*]$. $\square$

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11558-021-09421-5>.

Funding Open Access funding provided by Universität Luzern. No third-party funding.

Availability of Data and Material All observational data (Section 2.1) are in the public domain (see figure notes for data sources).

Code Availability Code (in Python) available from authors upon request.

Declarations

Conflict of Interests No conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Bai, J., Jayachandran, S., Malesky, E.J., Olken, B.A. (2019). Firm growth and corruption: Empirical evidence from vietnam. *The Economic Journal*, 129(618), 651–677.
- Banerjee, A., Hanna, R., Kyle, J., Olken, B.A., Sumarto, S. (2019). Private outsourcing and competition: Subsidized food distribution in indonesia. *Journal of Political Economy* 127(1).
- Barro, R.J. (1990). Government spending in a simple model of endogenous growth. *Journal of Political Economy*, 98(5), 103–125.
- Beekman, G., Bulte, E., Nillesen, E. (2014). Corruption, investments and contributions to public goods: Experimental evidence from rural Liberia. *Journal of Public Economics*, 115, 37–47.
- Berry, F., Iommi, M., Stanger, M., Venter, L. (2018). The status of GDP compilation practices in 189 economies and the relevance for policy analysis. *IMF Working Paper* 18(37).
- Binswanger, J., & Oechslin, M. (2015). Disagreement and learning about reforms. *Economic Journal*, 125(May), 853–886.
- Binswanger, J., & Oechslin, M. (2020). Better Statistics, Better Economic Policies? *European Economic Review*, 130, 103588.
- Brookings Institution (2018). Global cybercrimes and weak cybersecurity threaten businesses in Africa. <https://www.brookings.edu/blog/africa-in-focus/2018/05/30/global-cybercrimes-and-weak-cybersecurity-threaten-businesses-in-africa/>. Accessed January 2020.
- Colonnelli, E., & Prem, M. (2020). Corruption and Firms. SSRN Working Paper.
- della Porta, D., & Vannucci, A. (2012). The hidden order of corruption: An institutional approach. Farnham, UK: Ashgate.
- Devarajan, S. (2013). Africa's statistical tragedy. *Review of Income and Wealth*, 59(Special Issue), 9–15.
- Estrin, S., Korosteleva, J., Mickiewicz, T. (2013). Which institutions encourage entrepreneurial growth aspirations? *Journal of Business Venturing*, 28(4), 564–580.
- Fisman, R., & Svensson, J. (2007). Are corruption and taxation really harmful to growth? Firm level evidence. *Journal of Development Economics*, 83(1), 63–75.
- Freund, C., Hallward-Driemeier, M., Rijkers, B. (2016). Deals and delays: Firm-level evidence on corruption and policy implementation times. *World Bank Economic Review*, 30(2), 354–382.
- GADM database of Global Administrative Areas (2012). Version 1. <http://www.gadm.org/>. Accessed May 2012.
- Hainmueller, J., Mummolo, J., Xu, Y. (2019). How much should we trust estimates from multiplicative interaction models? simple tools to improve empirical practice. *Political Analysis*, 27(2), 163–192.
- Henderson, J.V., Storeygard, A., Weil, D.N. (2012). Measuring economic growth from outer space. *American Economic Review*, 102(2), 994–1028.
- Hodler, R., & Raschky, P.A. (2014). Economic shocks and civil conflict at the regional level. *Economics Letters*, 124(3), 530–533.
- ILO. (2018). *Women and men in the informal economy: A statistical picture*. Geneva: International Labor Organization.
- Immordino, G., & Pagano, M. (2010). Legal standards, enforcement, and corruption. *Journal of the European Economic Association*, 8(5), 1104–1132.
- Jerven, M. (2013). *Poor numbers*. Ithaca: Cornell University Press.
- Kerner, A., Jerven, B., Beatty, A. (2017). Does it pay to be poor? Testing for systematically underreported GNI estimates. *Review of International Organizations*, 12(1), 1–38.
- Kiregyera, B. (2015). *The emerging data revolution in africa*. Stellenbosch: SUN MeDIA.

- Malesky, E.J., & Samphantharak, K. (2008). Predictable corruption and firm investment: Evidence from a natural experiment and survey of Cambodian entrepreneurs. *Quarterly Journal of Political Science*, 3(3), 227–267.
- Manski, C.F. (2015). Communicating uncertainty in official economic statistics: An appraisal fifty years after morgenstern. *Journal of Economic Literature*, 53(3), 631–653.
- Muralidharan, K., & Niehaus, P. (2017). Experimentation at scale. *Journal of Economic Perspectives*, 31(4), 103–124.
- National Oceanic and Atmospheric Administration (NOAA) (2014). Version 4 DMSP-OLS Nighttime Lights Time Series. Boulder, CO: National Geophysical Data Center. <https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html>. Accessed January 2014.
- Olken, B.A., & Pande, R. (2012). Corruption in developing countries. *Annual Review of Economics*, 4(1), 479–509.
- Paunov, C. (2016). Corruption's asymmetric impacts on firm innovation. *Journal of Development Economics*, 118, 216–231.
- Pitlik, H., Frank, B., Firschow, M. (2010). The demand for transparency: An empirical note. *Review of International Organizations*, 5(2), 177–195.
- Privacy International (2019a). State of Privacy India. <https://privacyinternational.org/state-privacy/1002/state-privacy-india>. Accessed January 2020.
- Privacy International (2019b). State of Privacy Kenya. <https://privacyinternational.org/state-privacy/1005/state-privacy-kenya>. Accessed January 2020.
- Rodrik, D. (2010). Diagnostics before prescription. *Journal of Economic Perspectives*, 24(3), 33–44.
- Sandefur, J., & Glassman, A. (2015). The Political Economy of Bad data: Evidence from African Survey & Administrative Statistics. *The Journal of Development Studies*, 51(2), 116–132.
- Serianu (2017). Africa Cyber Security Report 2017. <https://digital4africa.com/wp-content/uploads/2018/04/Africa-Cyber-Security-Report-2017.pdf> Accessed December 2020.
- Svensson, J. (2003). Who must pay bribes and how much? evidence from a Cross-Section of firms. *The Quarterly Journal of Economics*, 118(1), 207–230.
- World Bank (2015). World Development report: Spotlight When corruption is the norm. *Part, 1*, 60–61.
- Zakharov, N. (2019). Does corruption hinder investment? Evidence from Russian regions. *European Journal of Political Economy*, 56, 39–61.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.