# Open Data from Authoritarian Regimes: New Opportunities, New Challenges

*Ruth D. Carlitz and Rachael McLellan*

Data availability has long been a challenge for scholars of authoritarian politics. However, the promotion of open government data—through voluntary initiatives such as the Open Government Partnership and soft conditionalities tied to foreign aid—has motivated many of the world's more closed regimes to produce and publish fine-grained data on public goods provision, taxation, and more. While this has been a boon to scholars of autocracies, we argue that the politics of data production and dissemination in these countries create new challenges. Systematically missing or biased data may jeopardize research integrity and lead to false inferences. We provide evidence of such risks from Tanzania. The example also shows how data manipulation fits into the broader set of strategies that authoritarian leaders use to legitimate and prolong their rule. Comparing data released to the public on local tax revenues with verified internal figures, we find that the public data appear to significantly underestimate opposition performance. This can bias studies on local government capacity and risk parroting the party line in data form. We conclude by providing a framework that researchers can use to anticipate and detect manipulation in newly available data.

Until recently, the production and dissemination of fine-grained administrative data has been the exclusive provenance of developed democracies. Over the past decade, however, a growing number of poorer and less democratic states have begun releasing more data to the public. Indeed, when it comes to availability and periodicity of key socioeconomic indicators, the World Bank ranks the current performance of electoral and closed autocracies nearly on par with that of democracies.[1]

The availability of fine-grained data from previously closed contexts appears at first to be a boon to scholars of authoritarian politics. With access to new data, scholars are now working to address what had seemed previously intractable questions. As a result, a growing number of quantitative studies in comparative politics draw conclusions from non-democratic contexts.[2] Moreover, the availability of such data makes it possible to rectify previous biases in cross-country studies, which have been hampered by a lack of data from higher income authoritarian states.[3] However, as we detail in this piece, scholars should treat these newly available data with caution.

First, concerns about data quality are pervasive in the developing world (Jerven 2013; Devarajan 2013). Official

*A list of permanent links to Supplemental Materials provided by the authors precedes the References section.*

**Data replication sets are available in Harvard Dataverse at: https://doi.org/10.7910/DVN/AUTKAV*

*Ruth D. Carlitz is Assistant Professor of Political Science at Tulane University, where she focuses on international development and African politics (rcarlitz@tulane.edu). Her research looks at government responsiveness from the "top down" (how governments distribute public goods) and the "bottom up" (what citizens can do to promote transparency and accountability). Her work has appeared in* Research & Politics, World Development, The Journal of Comparative Policy Analysis, *and* Development Policy Review. *She has also worked on evaluations commissioned by NGOs and funding agencies such as the World Bank and USAID.*

*Rachael McLellan is a Postgraduate Research Associate at Princeton University where she completed her PhD in 2020 (rachaelm@princeton.edu). Her work focuses on regime and opposition party strategy, local politics and political behavior in electoral autocracies. She won the* American Political Science Association (Democracy & Autocracy Section) Fieldwork Prize *in 2019 for her dissertation work on local state capacity and local electoral control in Tanzania.*

*We thank Dan Berliner, Jim Vreeland, and Dan Tavana for their thoughtful comments as well as panel participants at the European Political Science Association Conference 2019.*

doi:10.1017/S1537592720001346

statistics are often inaccurate or incomplete due to a range of bureaucratic or administrative factors. For instance, Sandefur and Glassman (2015) provide compelling evidence that official statistics systematically exaggerate development progress across multiple African countries, reflecting two interlinked principal-agent problems: First, governments misreport to foreign donors, particularly in the context of results-based aid. Second, governments are themselves frequently misled by frontline service providers tasked with simultaneously providing public services and reporting truthful data on the same.

Scholars have also begun to highlight the influence of domestic politics on data manipulation. Autocrats are under pressure to release data but they have control over what data is released and its veracity, control which they can exploit for their own ends. For instance, population data is often politicized as it frequently determines the distribution of federal resources. This threatens the reliability of these data because governments have an incentive to manipulate it (Akinyoade, Appiah and Asa 2017; Elemo 2018). Macroeconomic indicators are also frequently politicized—particularly in nondemocratic contexts (Martinez 2019; Rawski 2001; Tsai 2008; Wallace 2016).

We build on this literature by highlighting an additional set of motivations that autocrats have to engage in targeted manipulation of subnational data in order to undermine the reputation of would-be challengers. Such data manipulation fits into the broader set of strategies that authoritarian leaders use to legitimate and prolong their rule. An empirical example from Tanzania, comparing data on local government tax takings released to the public with verified internal figures, provides evidence of this phenomenon and its associated risks. We conclude with a framework that researchers can use to anticipate and detect manipulation in their data.

## Why Do Authoritarian Regimes Release Data?

The increased availability of fine-grained data from authoritarian regimes reflects two parallel trends that reinforce each other. First, the Millennium Development Goals and their successors, the Sustainable Development Goals, have oriented the international development community towards clearly defined targets, motivating a raft of statistical exercises to improve the tracking of economic and social indicators (Kelley and Simmons 2019; Jerven and Johnston 2015; Sandefur and Glassman 2015). These include a push for national governments to disseminate information under open data protocols. At the same time, the promotion of open government data has pervaded the policy agendas of governments around the world (Davies and Bawa 2012). For example, the voluntary Open Government Partnership, launched in 2011, counts seventy-eight countries and twenty subnational governments among its members, who together have generated over 3,100 commitments to make their governments more open and accountable.[4] Star performers include not only the usual suspects in Western Europe and North America, but also lower income, less democratic countries such as Kenya, Honduras, and Moldova.

That said, democracies tend to be more transparent than authoritarian regimes (Hollyer, Rosendorff, and Vreeland 2011). Autocrats have an incentive to resist transparency because it can be dangerous for their survival. Making information freely available allows citizens to not only update their beliefs about government performance, but also their beliefs about what other citizens believe (Hollyer, Rosendorff, and Vreeland 2015). Journalists, politicians, and civil society can leverage open data to criticize the regime and encourage mobilization against it, which can foment dissatisfaction with the regime (Reuter and Gandhi 2011).

However, autocrats do release data. An emerging literature suggests that authoritarian regimes may have incentives to allow or even promote certain forms of transparency to help them win elections (Maerz 2016). Little (2017) shows that the release of manipulated information on good performance makes it more likely that voters will coordinate to support the regime even if they do not sincerely believe the information. Transparency can also increase elite cohesion as a strategic response to greater threats from the masses (Hollyer, Rosendorff, and Vreeland 2018). Berliner (2014) argues that the passage of freedom of information laws allows incumbents (in democracies and autocracies) to ensure that they will not be shut out of access to government information and tools of monitoring if they lose power in the future. Subnational analysis from Mexico confirms that incumbents are particularly likely to pass such reforms when their grasp on power is less secure (Berliner and Erlich 2015).

The key challenge for scholars of authoritarian politics wishing to take advantage of data released under such initiatives is the heightened potential for manipulation. Whereas democratic leaders may also be tempted to manipulate official figures in their favor, countervailing institutions like a free press, independent judiciary, and impartial bureaucracy make this relatively difficult. It is arguably easier for autocrats to manipulate data before it is released, altering it in politically expedient ways. Furthermore, as we will show, incentives to manipulate data may vary across subnational units—creating particular threats to inference for analysis at this level.

## Evidence of Risks from Tanzania

Tanzania exemplifies the range of ways that autocrats can strategically censor and manipulate economic indicators for their own benefit. The country's ruling party, Chama Cha Mapinduzi (CCM), has been in power since independence in 1961 and there is little separation between the party and the state (Morse 2014, 2018).[5] The country has taken an increasingly authoritarian turn since the election of its current president John Magufuli in 2015 (McLellan 2018). Magufuli has pursued an unprecedented number of interventionist economic policies alongside an outright attack on political freedoms and opposition parties. To protect his economic record, Magufuli and his government have engaged in a number of tactics to obscure potentially damaging information, including blocking the publication of key domestic and international reports (Cotterill 2019; Collord 2019; African Arguments 2019).

However, as in many other low- and middle-income electoral autocracies, international organizations and donor governments exert significant leverage. As a result, the country has adopted a range of measures to increase data transparency in recent years. These include the publication of a raft of administrative data as part of the country's open government partnership with the World Bank (World Bank 2015).[6]

The release of such data can pose problems for the ruling party, particularly if it changes voters' beliefs about the regime and opposition parties' competency. Hegemonic parties like CCM maintain support by convincing voters they are the only party which can successfully rule (Magaloni 2006; Guriev and Treisman 2015). This image may be threatened by the release of data showing that opposition parties outperform them. In what follows, we provide evidence suggesting that the CCM regime manipulates public data to discredit opposition parties. Since 2005, opposition share of the vote has been growing. In 2015, opposition parties won 40% of the presidential vote and 45% of the legislative vote. Chadema, the main opposition party, relies on good performance in local government to win over voters. Hence, the Tanzanian regime is increasingly concerned with managing local as well as national competition to stay in power (McLellan 2020).

As evidence of how such dynamics affect the release of official statistics, we compare internal local tax collection data from Tanzania for the 2016/2017 fiscal year[7] with that released online through the Local Government Revenue Collection Dashboard, an open data portal established in 2017 and hosted on the website of the President's Office for Local Government and Regional Administration (TAMISEMI). These data are published by presidential appointees in TAMISEMI who are directly accountable to the president, making manipulation feasible.

Tanzanian local governments raise taxes to supplement central transfers and fund key services (roads, education, health, water) to foster development and garner political support. As in other decentralized countries, local tax revenue serves as an important indicator of local government performance, which can be important to win over voters (Lucardi 2016). The government began releasing data on local government tax takings as part of the aforementioned transparency initiatives, and this data has emerged as a yardstick by which Tanzanian politicians and the press appraise local government performance (Malanga 2019).

While opposition parties use tax takings to show they are delivering on their local mandate, the central government uses these figures to herald the success of key incumbent areas (Chidawali 2018). Newspapers regularly report on revenue collection, influencing how voters and elites view the relative credibility of the regime and opposition parties.[8] The political salience of these figures and the regime's influence over the bureaucrats who produce the data make it particularly vulnerable to manipulation.

### *Suspicious Digits*

As a first step for ascertaining evidence of manipulation, we leverage established methods for detecting fraud in official statistics. In general, these methods make certain assumptions about the data-generating process and then analyze whether official government statistics deviate from what would be expected in the absence of manipulation. These include various forms of digit analysis (Beber and Scacco 2012; Mebane 2010; Deckert, Myagkov, and Ordeshook 2011).

One common benchmark, Benford's Law, proposes that the leading digits in non-manipulated sets of numbers should conform to a distribution where the number 1 is the most likely to occur and the number 9 is the least likely to occur. We therefore run first digit analysis on both tax datasets, finding that the internal data is more consistent with theoretical expectations about digit distribution than the public data.[9] Such results suggest cause for concern, but they are far from conclusive.

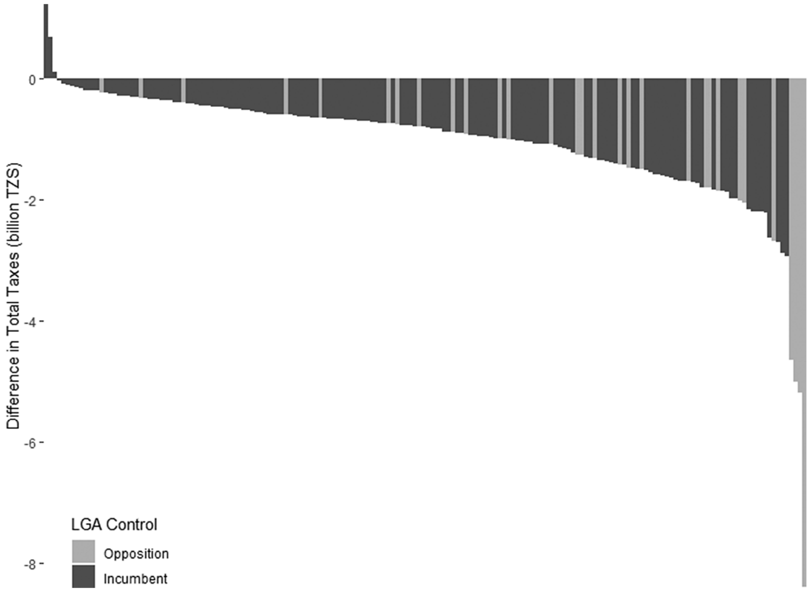
We next proceed to analyze the last digits of the tax data. Here, we would expect a uniform probability of the numbers 0–9 in sets of numbers that have not been manipulated. We analyze the last digits of the data and find that the distribution of last digits in the internal data conforms far better to the expected distribution of digits than the public data.[10] The results of this analysis are more compelling but still inconclusive.[11]

In sum, digit analysis of the public data suggests manipulation but is far from a smoking gun. None of the distributions discussed are statistically significantly different from the distribution that would be expected in the absence of manipulation. The results may be inconclusive because these methods analyze the entire distribution of the data. That means they are less sensitive to more targeted manipulation. As we will show, it is precisely such targeted manipulation that appears to be occurring in this case.

### *Evidence of Targeted Manipulation*

The fact that subnational performance is important to electoral competition in Tanzania may incentivize the regime to manipulate particular observations. Specifically,

**Figure 1**
**Difference between total revenue reported in internal and public data by LGA**



there are incentives to overstate the performance of local governments controlled by the ruling party and understate that of opposition-controlled localities. When we compare the internal and public figures, those released by the central government appear to underestimate local tax raising in the vast majority of Tanzania's local government authorities (LGAs). Figure 1 plots the differences between total revenue reported in internal and public data by LGA.[12] If these differences were fairly consistent, the most plausible explanation would be simple accounting problems.[13] However, there is substantial variation in the magnitude of the differences between the internal data and that released to the public. These differences range from an underestimate of 40 billion shillings ($16.5 million) in Kinondoni Municipal Council in Dar es Salaam to an overestimate of 1.25 billion shillings ($550,000) in Kibaha District Council. Per capita differences are similarly skewed. Moreover, these differences appear to vary systematically.

Further investigation suggests that the pattern that emerges can be most plausibly explained as a strategic response by the central government to the dynamics of political competition. This is in keeping with previous efforts to skew macroeconomic indicators in a way that is favorable to the regime. For instance, in 2017 a leading opposition MP was arrested and charged with sedition after his party published an analysis of official economic data from the Bank of Tanzania, suggesting that the official GDP growth statistics had been manipulated (Taylor 2017).

Opposition local governments' tax takings are underestimated to a far greater (and statistically significant) extent than those of incumbent local governments. In particular, the public data vastly underestimates the tax takings of prominent opposition strongholds like Moshi, Mbeya, Arusha, and central Dar es Salaam. Importantly, where the public data overestimates tax takings, it is in prominent regime strongholds like the administrative capital Dodoma.[14]

These differences in the data affect LGA rankings and thus the credit that opposition and ruling parties can claim for good performance. Table 1 shows the top five LGAs in total and per capita collection for the public and internal data. The top five LGAs according to the internal data are all opposition councils (shown in bold italics).[15] This is unsurprising in the Tanzanian context where there is evidence that opposition councils are strategically disfavored with central government transfers (Weinstein 2011), making them more reliant on their own revenue sources.

Opposition councils also dominate the top five in the public data but there are some notable shifts. First, Dodoma moves into the top five for total revenue collection from eleventh place in the internal data. This is noteworthy given that the transformation of Dodoma into a major hub is a government priority.[16] Second, Chadema councils fall out of the top five in per capita collection completely when the public data is used. In contrast, LGAs from other opposition parties remain among the top performers.

To provide further evidence, we regress CCM control on the difference between a local government's per capita local tax collection in the public data and the internal data.[18] We find that CCM control of a local government significantly reduces the difference between publicly released and internal figures.[19] The performance of opposition local governments is more likely to be underestimated. Our analysis suggests the effect is larger in opposition strongholds where the opposition share of councilors is higher, but this result is not robust to the more conservative log transformation used.

### Table 1
### Top performing local governments in 2016–2017

| Top 5 LGAs (total collection) | Top 5 LGAs (per capita collection) |
|---|---|
| Internal | Internal |
| 1. ***Kinondoni MC*** | 1. ***Kinondoni MC*** |
| 2. ***Arusha CC*** | 2. ***Mtwara MC*** |
| 3. ***Tanga CC*** | 3. ***Tanga CC*** |
| 4. ***Ubungo MC*** | 4. ***Arusha CC*** |
| 5. ***Mbeya CC*** | 5. ***Moshi MC*** |
| Public Data | Public Data |
| 1. ***Kinondoni MC*** | 1. ***Mtwara MC*** |
| 2. ***Arusha CC*** | 2. Kibaha DC (CCM) |
| 3. ***Ubungo MC*** | 3. Geita TC (CCM) |
| 4. Dodoma MC (CCM) | 4. ***Tandahimba DC*** |
| 5. ***Tanga CC*** | 5. Bagamoyo DC (CCM) |

Note: Opposition councils are shown with bold italics; CCM councils are labeled as such. These patterns suggest opposition performance has been strategically underestimated.[17]

For additional robustness, we look at whether opposition control predicts significant under-reporting of performance, which we define as those cases where the difference between public and internal is greater than a standard deviation of all the differences. [20] These are cases that we can more confidently say are intentional misreports. For total taxes, all significantly under-reporting LGAs are opposition LGAs. There are both opposition and CCM LGAs in the cases of significant under-reports of per capita tax revenues. Thus, we regress CCM control and share of opposition councilors on the likelihood of a significant underestimate. We find that CCM control significantly decreases the chances that an LGA significantly under-reports performance.[21] The chances of under-reporting increase with the share of opposition councilors. Taken together, these results suggest that the regime strategically underestimates opposition performance. The publicly released figures allow the ruling party to claim that they are performing well and undermine opposition parties' abilities to do the same.

### *Threats to Inference from Manipulated Data*

The manipulation our analysis suggests has significant implications for the inferences one can draw from the tax data as a whole. In the rest of this section, we present evidence of threats to inference for studies using local tax revenue as a dependent or independent variable. We compare the results of the models using the internal and the public data, demonstrating how partisan manipulation of data can bias a range of potential studies.

First, manipulation of the sort we detect may bias the results of any study that seeks to understand the effects of partisan control on local government performance. Table 2 plots the results of regression analysis investigating

### Table 2
### Effect of LGA partisanship on revenue raising

| | Dependent Variable: | | | |
|---|---|---|---|---|
| | Internal Data | | Public Data | |
| | Log (per capita tax revenue) | | | |
| | (1) | (2) | (3) | (4) |
| CCM majority | -0.281$^{**}$<br>(0.111) | — | -0.288<br>(0.242) | — |
| Share of opposition councilors | — | 0.580$^{***}$<br>(0.195) | — | 0.511<br>(0.419) |
| Observations | 179 | 173 | 179 | 173 |
| R$^2$ | 0.711 | 0.745 | 0.408 | 0.468 |
| Adjusted R$^2$ | 0.652 | 0.692 | 0.288 | 0.356 |
| Residual Std. Error | 0.412 | 0.392 | 0.897 | 0.842 |
| F Statistic | 12.118$^{***}$ | 13.864$^{***}$ | 3.395$^{***}$ | 4.163$^{***}$ |

$^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01
Note: Models include controls for population, administrative status, and region.

**Table 3**
**Effect of local tax capacity on service delivery**

| | Dependent Variable: | | | |
|---|---|---|---|---|
| | Test Scores | | Pass Rate | |
| | (1) | (2) | (3) | (4) |
| Per capita tax (internal) | 0.2701*<br>(0.1562) | — | 0.2679*<br>(0.1607) | — |
| Per capita tax (public) | — | 0.5076***<br>(0.1789) | — | 0.6066***<br>(0.1821) |
| Observations | 176 | 176 | 176 | 176 |
| $R^2$ | 0.6151 | 0.6280 | 0.6271 | 0.6472 |
| Adjusted $R^2$ | 0.5290 | 0.5448 | 0.5436 | 0.5682 |
| Residual Std. Error | 8.3393 | 8.1984 | 8.5779 | 8.3432 |
| F Statistic | 7.1416*** | 7.5439*** | 7.5137*** | 8.1975*** |

*p<0.1; **p<0.05; ***p<0.01
Note: Models include controls for population, administrative status, region, average level of education, and CCM control of the LGA. Per capita tax is in 1000s of TSH for easier interpretation. For every additional 1000 shillings collected, primary school test scores and pass rates in these tests increase by the coefficients shown.

the effect of LGA partisanship on local revenue collection using both the internally validated and publicly available data. The analysis based on internally validated figures shows that opposition LGAs collect significantly higher total and per capita taxes. Likewise, an increase in opposition share increases per capita and total tax collection.[22,23] However, the significance of these relationships disappears when the public data is used. A researcher using the internal data would conclude that opposition councils indeed collect more taxes while a researcher using the public data would not find any significant effect of opposition control.

Manipulation of this sort can also change the results of studies which ostensibly have nothing to do with party politics. Consider a researcher who is interested in how increased bureaucratic capacity affects tax collection. As part of their study, the researcher may test the effect of administrative status (rural, town, city, etc.) on tax collection.[24] Because a disproportionate number of opposition councils are urban, the results of this study may be biased because opposition performance and hence urban performance is underestimated. Indeed, the two datasets generate significantly different results.[25] Because the manipulation may not be consistent across units or over time, it is difficult to control for it directly or through fixed effects.

The manipulation we detect could also bias studies where local tax is used as the independent variable to proxy for state capacity. A scholar interested in how subnational variation in state capacity affects service delivery may look at the relationship between local tax takings and school performance. We show the results of this analysis in table 3. When we use the public data, we find a large and highly significant positive relationship between local tax capacity and two measures of school performance. When we use the internal data, the magnitude of the coefficients is halved and the relationship is no longer significant at the 0.05 level.

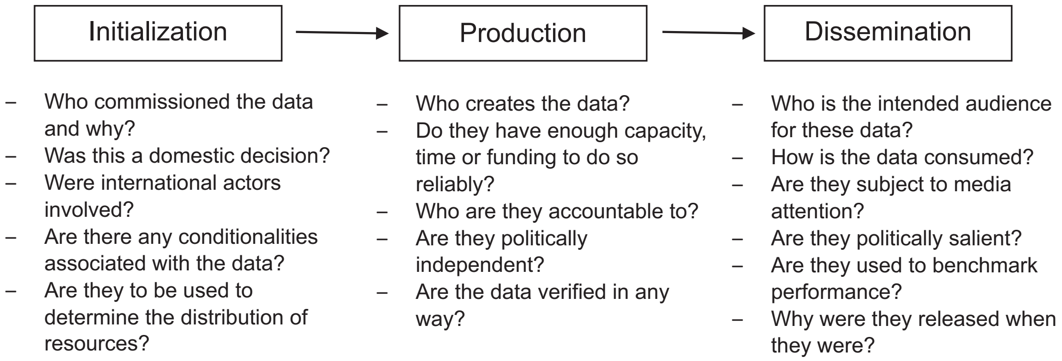
Across all three analyses, we find that strategically manipulated data biases the results. Perhaps the most pernicious are flawed inferences about the relationship between regime support and a variety of "good" outcomes for the regime. Not only does this risk parroting pro-government propaganda in data form, it can also lead to flawed inferences about why people support the regime and how the regime rewards supporters. Furthermore, public data may underestimate what the incumbent wants to hide, like opposition performance or discrimination in access to public resources. This makes it harder for anti-regime actors to claim credit or point to government abuses of state power. As a result, scholars may fail to appreciate the role of these actors in non-democratic systems.

We realize that most scholars do not have access to both public and internal figures to identify the kind of manipulation we find. In what follows, we therefore outline a framework for anticipating and identifying data manipulation in order to reduce the risk of flawed inferences such as those discussed here.

## Anticipating and Identifying Data Manipulation

This reflection is not intended to discourage scholars from taking advantage of newly available data from non-democratic regimes. However, we urge caution and reflection on the politics of data production and dissemination and suggest some additional steps to convince readers of the reliability of such data. Even researchers interested in questions that seem far removed from the landscape of political competition and control should take heed since that very landscape may determine the quality of available data.

**Figure 2**
**Political considerations at each stage of data production**



- Who commissioned the data and why?
- Was this a domestic decision?
- Were international actors involved?
- Are there any conditionalities associated with the data?
- Are they to be used to determine the distribution of resources?

- Who creates the data?
- Do they have enough capacity, time or funding to do so reliably?
- Who are they accountable to?
- Are they politically independent?
- Are the data verified in any way?

- Who is the intended audience for these data?
- How is the data consumed?
- Are they subject to media attention?
- Are they politically salient?
- Are they used to benchmark performance?
- Why were they released when they were?

### *Anticipating the Risks*

As a general rule, researchers should understand when data is at most risk of being manipulated and the types of flawed inference that falsified data may create. Wallace (2016) argues that data is most likely to be manipulated if it is politically sensitive and is released at politically sensitive times. We concur and argue that scholars should be mindful of the whole process involved in commissioning, producing, and releasing data to the public. Data may be more or less reliable depending on political incentives at each stage from initialization to dissemination. In figure 2, we propose a set of questions scholars should ask themselves when trying to understand the risks of manipulation.

First, researchers should consider who commissions the data and their motivation for doing so. This affects the audience(s) for whom the data is likely to be important, which in turn affects who stands to benefit or lose from the information that is released. For example, census data is collected to inform allocation of budgets and definition of electoral districts. Local performance data may be collected to benchmark central transfers. Elections in autocracies are partly held to gather information about the distribution of support that may then determine the distribution of resources. Data may also be collected to convey information to international actors. For instance, GDP figures matter for international reputation and investment. Data on primary school enrolment, HIV/AIDS, malnutrition, and other socioeconomic outcomes are relevant to tracking progress toward the Sustainable Development Goals, determining foreign aid envelopes, and informing external perceptions of domestic performance in low-income countries, which can in turn affect foreign direct investment. Data which are likely to influence the allocation of resources to a country or within it are particularly at risk of being manipulated.

Second, the production stage—who generates the data, their capacity, and their incentives—is likely to influence data quality and the ease with which it can be manipulated. Highly ambitious data collection projects that have little oversight and low budgets are more likely to be unreliable. Data produced by bureaucrats subject to political pressure are more likely to be manipulated as is the case in the example explored in the previous section. On the other hand, data that is endorsed by civil society organizations or international organizations with in-country presence are less likely to be manipulated because of additional oversight.

Third, dissemination – the audience for the data and its salience—is the last and perhaps most important stage of the process. Data that is primarily for internal use and released on open data platforms as a matter of course is significantly less likely to be manipulated than data that may be covered in the national press. Political salience is also an important consideration. Data may gain salience for a variety of reasons. Earlier, we showed how data can become salient when it measures subnational performance. In our example, the data is reported regularly in the press and is a politically salient benchmark for parties' performance. Note that performance may be measured using data that is idiosyncratic to a given country. In our Tanzanian example, local tax is an important metric of performance. In Rwanda, maternal mortality is regularly reported to measure local performance (Worley 2015). In Mexico, murder rates are fundamentally important for the credibility of state governments. Indeed, they are thought to be manipulated in some cases (Telesur 2017).

### *Know Your Data, Know Your Case*

As we have shown, data is not produced in a vacuum. Without a good grasp on the politics of data in their country of study, it is difficult for researchers to understand which data is most likely to be manipulated. Without case knowledge, it is also difficult for researchers to preempt the kinds of manipulation that incumbents might have incentives to make and, hence, what threats to inference they face when analyzing data. We therefore encourage researchers to gain familiarity with the context in which their data is produced—talking to and co-authoring with local scholars and case experts, as well as trying to validate subnational data with local bureaucrats, politicians, and members of civil society. Where possible, they should also make use of comparable questions in validated, independent data like the Barometer surveys or Demographic and Health Surveys. All these strategies allow scholars to generate a baseline against which they can appraise the public data they are using. These kinds of simple "sniff tests" protect researchers from publishing work based on manipulated data.

Similarly, scholars should also get to know their data before diving into analysis. We recommend that researchers use their priors about the data generating process relevant to their area of interest in order to look for deviations from it that may suggest manipulation. Exploratory analysis—initial investigations to discover patterns, spot irregularities, and check assumptions with the help of summary statistics and graphical representations—is especially important when dealing with potentially risky data. If scholars have concerns about the quality of their data, they can use digit analysis and other fraud detection techniques. Other examples of such tests include the Whipple and Myers Indices, which provide summary measures of "heaping" on numbers ending in 0 or 5 (Borkotoky and Unisa 2014). These measures are most commonly applied to detect incorrect information on age in population data. Scholars have also developed a variety of regression-based approaches that involve the estimation of statistical models and examine irregularities, anomalies, or outliers (Alvarez and Katz 2008; Myagkov, Ordeshook, and Shakin 2009; Wand et al. 2001). We contend that these techniques, developed to detect falsification of population and election data, should be applied more broadly to administrative data released by authoritarian governments.

Ultimately, engaging with the "politics of data" can lend greater credence to the inferences drawn from newly available data. As in discussions of research design, where scholars make the case that their samples are representative or demonstrate balance across samples when leveraging natural experiments, scholars should also work to convince readers why and to what extent they are convinced of the quality of their data. Furthermore, they should acknowledge any possible bias and its suspected direction. By being upfront about such issues, scholars can demonstrate an awareness of risks and signal an effort to minimize them. This makes it easier for readers to transparently engage with their work.

Finally, we call on scholars and civil society to track data censorship and manipulation. Organizations like V-Dem and Freedom House should consider incorporating data censorship and manipulation into their existing work tracking other forms of suppression. Measures of data censorship are critical for scholars to understand the likelihood that the data they are using is compromised and therefore the level of caution they should exercise when using it. Furthermore, as data journalism becomes increasingly prominent, these kinds of measures are even more important for civil society and the press to be able to appraise what conclusions can be drawn from official statistics and what claims should be questioned.

## Supplementary Materials

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S1537592720001346.

## Notes

1 We compare average scores on the World Bank's Periodicity and Timeliness indicator for sixty-seven countries classified as electoral and closed autocracies by the *Regimes of the World* variable in the V-Dem Dataset v8 (Coppedge et al. 2018) with that of sixty-three countries classified as democracies. Autocracies score on average 79/100, whereas democracies score only slightly higher at 85/100 on average (World Bank 2012).

2 Since 2010, six top political science journals published 194 articles containing the words "subnational," "data," and "autocracy." This is nearly four times as many articles published containing similar keywords in the previous decade. These results are based on a Google Scholar search of *American Political Science Review, American Journal of Political Science, World Politics, Annual Review of Political Science, Comparative Political Studies*, and *British Journal of Political Science.*

3 See, e.g., Ross 2006, who shows how missing data from such countries led to flawed inferences regarding the effect of democracy on poverty reduction. More recently, Lall 2017 suggests that a "prodemocracy bias" in data availability has colored the debate over the existence of the political resource curse.

4 For more information, see https://www.opengovpartnership.org/about/.

5 Prior to 1977, the party was known as the Tanganyika African National Union (TANU) on the mainland.

6 This data has informed a number recent of working and published papers, including the authors' own work.

7 This data was collected directly from the President's Office for Regional Administration and Local Government by one of the authors. The reliability of this data was corroborated through interviews with bureaucrats and politicians in several regions.

8 Refer to online appendix figure A1.

9 Refer to online appendix figure A2.

10 Refer to online appendix figure A3.

11 This may be because of small sample sizes, which limit how confidently we can draw inferences throughout this paper. Here it is a particular problem because we are making claims about the extent to which a distribution of a small sample approximates an expected distribution. With a sample this small, it would be unlikely for even randomly generated numbers to conform exactly to the expected distributions; Lesperance et al. 2016. In the OLS analyses we present later, sample size is less of a problem as we are interested in comparing the results from samples with the same number of observations. It is possible that some results we show could be false positives or negatives because of statistical noise. However, we find consistent evidence of differences between the results of like models used to analyze each dataset across multiple measures and specifications.

12 We reproduce the same plot including the large outlier of Kinondoni in online appendix figure A4.

13 The open data included shillings to two decimal places. The smallest unit of Tanzanian shillings currency is 50 shillings. This suggests that the data entered had been converted back from U.S. dollars (or another currency). Variation between internal and public data is most likely because of simple currency fluctuations between the first conversion into US dollars for international consumption and back into Tanzanian shillings whenever the data was then posted online. This would account for the smooth variation between most of the differences. The relevant differences are those at the extremes of the distribution.

14 These differences could plausibly also be explained by systematic differences in the quality of bureaucrats sent to incumbent and opposition areas. If poor quality bureaucrats are sent to opposition areas, more human error could explain under-counting in opposition areas. However, interviews with local bureaucrats and politicians conducted by one of the authors between 2016 and 2018 suggest that the regime cycles high-quality bureaucrats to opposition areas to improve oversight.

15 We define opposition councils as those with a majority of opposition councilors.

16 Dodoma is the political capital of Tanzania but has long paled into irrelevance compared to the financial, cultural, and administrative capital of Dar es Salaam. President Magufuli has made it a priority to move ministries, embassies and other political offices to Dodoma; such moves are frequently heralded by pro-government newspapers; Lugongo 2019.

17 One might ask why the regime does not instead overestimate performance in the localities they control. We see this as a strategic calculation whereby overestimating revenues in incumbent-held areas could lead to greater demands for service provision and also raise questions about money that cannot be accounted for. This is in keeping with President Magufuli's anti-corruption zeal and public dislike for wasting money; BBC News 2019.

18 For this analysis and that which follows, our data is a cross-section with a relatively low sample size. Thus, we use simple ordinary least square models. When used as our dependent variable, we log transform local tax revenue because the distribution is leftward skewed.

19 Refer to online appendix tables A1 and A2. We find similar results for the difference between total tax revenues in internal and public data.

20 No over-reports fit this criterion for either total local tax or per capita revenues.

21 Refer to online appendix table A3.

22 Results are robust to dropping Kinondoni.

23 We lose some observations because of missing data. However, the number of observations is the same across comparable models, so the loss of power does not pose a problem for inference.

24 The central government decides which LGAs are administratively upgraded based on population, population density, and their own discretion about "readiness." When LGAs are upgraded, they gain additional transfers and an increased workforce of bureaucrats, among other resources.

25 Refer to online appendix table A4.

## References

African Arguments. 2019. "Tanzania Search for Missing Millions Raises Questions over $1 Billion—African Arguments." Retrieved March 26, 2020 (https://africanarguments.org/2019/02/13/tanzania-search-missing-millions-reveals-missing-billion/)

Akinyoade, Akinyinka, Eugenia Appiah, and Sola Asa. 2017. "Census-Taking in Nigeria: The Good, the Technical, and the Politics of Numbers." *African Population Studies* 31(1): 3383–94. (https://aps.journals.ac.za/pub/article/view/997).

Alvarez, R. Michael, and Jonathan N. Katz. 2008. "The Case of the 2002 General Election." In *Election Fraud: Detecting and Deterring Electoral Manipulation*, ed. R. Michael Alvarez, Thad E. Hall, and Susan D. Hyde, 149–61. Washington, DC: Brookings Institution Press.

BBC News. 2019. "John Magufuli - Tanzania's 'Bulldozer' President in Profile." Retrieved March 26, 2020 (https://www.bbc.com/news/world-africa-34670983).

Beber, Bernd, and Alexandra Scacco. 2012. "What the Numbers Say: A Digit-Based Test for Election Fraud." *Political Analysis* 20(2): 211–34.

Berliner, Daniel. 2014. "The Political Origins of Transparency." *Journal of Politics* 76(2): 479–91.

Berliner, Daniel, and Aaron Erlich. 2015. "Competing for Transparency: Political Competition and Institutional Reform in Mexican States." *American Political Science Review* 109(1): 110–28.

Borkotoky, Kakoli, and Sayeed Unisa. 2014. "Indicators to Examine Quality of Large Scale Survey Data: An Example through District Level Household and Facility Survey." *PLoS One* 9(3): e90113.

Chidawali, Habel. 2018. "Revealed: What Gave Dodoma City a Boost in Revenue Earnings." *The Citizen*, August 2. Retrieved March 26, 2020 (https://www.thecitizen.co.tz/News/Land-sale–transparency-boosts-revenue-earnings-for-Dodoma-City/1840340-4694296-14j1m3ez/index.html).

Collord, Michaela. 2019. "Tanzania: The Politics of Being Auditor General." Retrieved March 26, 2020 (https://presidential-power.com/?p=9503).

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Agnes Cornell, Sirianne Dahlum, Haakon Gjerløw, Adam Glynn, Allen Hicken, Joshua Krusell, Anna Lührmann, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Juraj Medzihorsky, Moa Olin, Pamela Paxton, Daniel Pemstein, Josefine Pernes, Johannes von Römer, Brigitte Seim, Rachel Sigman, Jeffrey Staton, Natalia Stepanova, Aksel Sundström, Eitan Tzelgov, Yi ting Wang, Tore Wig, Steven Wilson, and Daniel Ziblatt. 2018. "V-Dem Dataset v8."

Cotterill, Joseph. 2019. "Tanzania President Blocks Critical IMF Report on Economy." *Financial Times*, April 18. Retrieved March 26, 2020 (https://www.ft.com/content/cb51db44-61f8-11e9-a27a-fdd51850994c)

Davies, Tim G., and Zainab Ashraf Bawa. 2012. "The Promises and Perils of Open Government Data (OGD)." *Journal of Community Informatics* 8(2): 1–8.

Deckert, Joseph, Mikhail Myagkov, and Peter C. Ordeshook. 2011. "Benford's Law and the Detection of Election Fraud." *Political Analysis* 19(3): 245–68.

Devarajan, Shantayanan. 2013. "Africa's Statistical Tragedy." *Review of Income and Wealth* 59: S-S15.

Elemo, Olufunmbi M. 2018. "Fiscal Federalism, Subnational Politics, and State Creation in Contemporary Nigeria." In *The Oxford Handbook of Nigerian Politics*, ed. Carl Levan and Patrick Ukata, 189–206. Oxford: Oxford University Press.

Guriev, Sergei, and Daniel Treisman. 2015. "How Modern Dictators Survive: An Informational Theory of the New Authoritarianism." Technical Report National Bureau of Economic Research.

Hollyer, James R., B. Peter Rosendorff, and James Raymond Vreeland. 2011. "Democracy and Transparency." *Journal of Politics* 73(4): 1191–205.

Hollyer, James R. 2015. "Transparency, Protest, and Autocratic Instability." *American Political Science Review* 109(4): 764–84.

Hollyer, James R. 2018. "Transparency, Protest and Democratic Stability." *British Journal of Political Science* 49(4): 1251–77.

Jerven, Morten. 2013. *Poor Numbers: How We Are Misled by African Development Statistics and What to Do About It*. Ithaca, NY: Cornell University Press.

Jerven, Morten, and Deborah Johnston. 2015. "Statistical Tragedy in Africa? Evaluating the Data Base for African Economic Development." *Journal of Development Studies* 51(2): 111–15.

Kelley, Judith G., and Beth A. Simmons. 2019. "Introduction: The Power of Global Performance Indicators." *International Organization* 73(3): 491–510.

Lall, Ranjit. 2017. "The Missing Dimension of the Political Resource Curse Debate." *Comparative Political Studies* 50(10): 1291–324.

Lesperance, M., W.J. Reed, M.A. Stephens, C. Tsao, and B. Wilton. 2016. "Assessing Conformance with Benford's Law: Goodness-of-Fit Tests and Simultaneous Confidence Intervals." *PloS one* 11(3): e0151235.

Little, Andrew T. 2017. "Propaganda and Credulity." *Games and Economic Behavior* 102: 224–32.

Lucardi, Adrian. 2016. "Building Support from Below? Subnational Elections, Diffusion Effects, and the Growth of the Opposition in Mexico, 1984–2000." *Comparative Political Studies* 49(14): 1855–95.

Lugongo, Bernard. 2019. "Tanzania: Government Move to Dodoma Now At 86 Per Cent." *Tanzania Daily News (Dar es Salaam),* February 6. Retrieved March 26, 2020 (https://allafrica.com/stories/201902060401.html)

Maerz, Seraphine F. 2016. "The Electronic Face of Authoritarianism: E-Government as a Tool for Gaining Legitimacy in Competitive and Non-Competitive Regimes." *Government Information Quarterly* 33(4): 727–35.

Magaloni, Beatriz. 2006. *Voting for Autocracy: Hegemonic Party Survival and Its Demise in Mexico*. Cambridge: Cambridge University Press.

Malanga, Alex. 2019. "LGA Have Collected Only 55 Per Cent of Targeted Revenue." *The Citizen*, April 11. Retrieved March 26, 2020 (https://www.thecitizen.co.tz/News/LGA-have-collected-only-55-per-cent-of-targeted-revenue/1840340-5066774-eaftcr/index.html)

Martinez, Luis R. 2019. "How Much Should We Trust the Dictator's GDP Growth Estimates?" SSRN Scholarly Paper ID 3093296 Social Science Research Network, Rochester, NY: Retrieved March 26, 2020 (https://papers.ssrn.com/abstract=3093296)

McLellan, Rachael S. 2018. "Why Is Once-Peaceful Tanzania Detaining Journalists, Arresting School- Girls and Killing Opposition Leaders?" *Washington Post,* November 30. Retrieved March 26, 2020 (https://www.washingtonpost.com/news/monkey-cage/wp/2018/11/30/why-is-once-peaceful-tanzania-detaining-journalists-arresting-schoolgirls-and-killing-opposition-leaders/)

McLellan, Rachael S. 2020. "*The Politics of Local Control in Electoral Autocracies.*" PhD Dissertation, Politics Department, Princeton University, Princeton, NJ.

Mebane, Walter R. 2010. "Fraud in the 2009 Presidential Election in Iran?" *Chance* 23(1): 6–15.

Morse, Yonatan L. 2014. "Party Matters: The Institutional Origins of Competitive Hegemony in Tanzania." *Democratization* 21(4): 655–77.

Morse, Yonatan L. 2018. *How Autocrats Compete: Parties, Patrons, and Unfair Elections in Africa*. New York: Cambridge University Press.

Myagkov, Mikhail, Peter Ordeshook, and Dimitri Shakin. 2009. *The Forensics of Electoral Fraud*. New York: Cambridge University Press.

Rawski, Thomas G. 2001. "What Is Happening to China's GDP Statistics?" *China Economic Review* 12(4): 347–54.

Reuter, Ora John, and Jennifer Gandhi. 2011. "Economic Performance and Elite Defection from Hegemonic Parties." *British Journal of Political Science* 41(1): 83–110.

Ross, Michael. 2006. "Is Democracy Good for the Poor?" *American Journal of Political Science* 50(4): 860–74.

Sandefur, Justin, and Amanda Glassman. 2015. "The Political Economy of Bad Data: Evidence from African Survey and Administrative Statistics." *Journal of Development Studies* 51(2): 116–32.

Taylor, Ben. 2017. "Has Zitto Revealed Fake News on Tz Economic Growth?" Retrieved March 26, 2020 (https://mtega.com/2017/11/has-zitto-revealed-fake-news-on-tz-economic-growth)

Telesur. 2017. "Are Mexican States Lying about Crime and Homicide Rates?" Retrieved March 26, 2020 (https://www.telesurenglish.net/news/Are-Mexican-States-Lying-About-Crime-and-Homicide-Rates–20170419-0020.html)

Tsai, Lily L. 2008. "Understanding the Falsification of Village Income Statistics." *China Quarterly* 196: 805–26.

Wallace, Jeremy L. 2016. "Juking the Stats? Authoritarian Information Problems in China." *British Journal of Political Science* 46(1): 11–29.

Wand, Jonathan N., Kenneth W. Shotts, Jasjeet S. Sekhon, Walter R. Mebane, Michael C. Herron, and Henry E. Brady. 2001. "The Butterfly Did It: The Aberrant Vote for Buchanan in Palm Beach County, Florida." *American Political Science Review* 95(4): 793–810.

Weinstein, Laura. 2011. "The Politics of Government Expenditures in Tanzania, 1999–2007." *African Studies Review* 54(1): 33–57.

World Bank. 2015. "Tanzania Conference Places Open Data at Center of Development Agenda." Retrieved March 26, 2020 (https://www.worldbank.org/en/news/press-release/2015/09/04/tanzania-conference-places-open-data-at-center-of-development-agenda).

World Bank. 2012. "Note on the Statistical Capacity Indicator." (http://siteresources.worldbank.org/INTWBDEBTSTA/Resources/Note_on_Statistical_Capacity_Indicator_BBSC_Nov2012.pdf).

Worley, Heidi. 2015. "Rwanda's Success in Improving Maternal Health Population Reference Bureau." Retrieved March 26, 2020 (https://www.prb.org/rwanda-maternal-health).