# Detecting Fraud in Development Aid

By Jean Ensminger and Jetson Leder-Luis*

Corresponding Author:
Jetson Leder-Luis
Boston University and NBER
595 Commonwealth Avenue #518E
Boston, MA 02115
Telephone: 617-353-2035
jetson@bu.edu

Jean Ensminger
California Institute of Technology
HSS MC 228-77
Pasadena, CA 91125
Telephone: 626-590-8070
jensming@hss.caltech.edu

* Acknowledgements: We particularly thank Avinash Dixit, Esther Duflo, Ben Gillen, Jonas Heese, Karla Hoff, Jonathan Katz, Pierre Liang, Ben Olken, Antonio Rangel, Eddie Riedl, Ethan Rouen and Robert Sherman for advice and support at critical junctures in this project. This paper benefited considerably from the comments of two reviewers and the editor, who took the time to write meticulously constructive reviews that guided our final rewrite and greatly improved the paper. In addition, we have benefitted from the comments of colleagues and seminar participants at presentations of earlier drafts of this paper at the ASSA meetings, The World Bank, Massachusetts Institute of Technology, California Institute of Technology, New York University, Center for Global Development, Duke University, Oxford University, and the University of California (Irvine). We thank the California Institute of Technology for generous funding of this research.

# Detecting Fraud in Development Aid


## Abstract

In developing countries, traditional antifraud measures including auditing may face barriers due to institutional resistance and practical difficulties on the ground. This is especially true in development aid, where aid organizations face incentives to suppress information about misappropriated funds and may operate with limited transparency and accountability. We develop new statistical tests to uncover strategic data manipulation consistent with fraud. These tests detect falsified cost reports and facilitate monitoring in difficult-to-audit circumstances, relying only on mandated reporting of data. While the digits of naturally-occurring data follow the Benford's Law distribution, humanly-produced data instead reflect behavioral biases and incentives to misreport. Our new tests distinguish intentional manipulation from benign misreporting and improve the statistical power of digit analysis. We apply this method to a World Bank development project in Kenya. Our evidence is consistent with higher levels of fraud in harder to monitor sectors and in a Kenyan election year when graft also had political value. The results are validated by a forensic audit conducted by the World Bank. We produce simulations that demonstrate the superiority of our new tests to the standards in the field, and provide evidence of the broad generalizability of Benford's Law.



(*JEL* Codes: H83, M42, C49, D73, O22)

## 1. Introduction

Firms and governments around the world incur large financial losses due to fraud. Organizations rely on the financial reporting of their agents, who can exploit asymmetric information to divert financial resources. This asymmetry can arise between a firm's owners and their employees, between regulators and firms, or between the public and the bureaucrats who serve them. Abundant resources are devoted to closing these information gaps and improving the quality of reported data, including disclosure regulation, audits, monitoring, whistleblowing and, increasingly, tests of the data themselves to determine their quality. All reported data contain both information about the underlying true values, and also signals about the quality of reporting (Leuz & Wysocki, 2016). These signals provide an avenue through which fraud can be detected, which we explore.

Developing countries, and the aid organizations that serve them, face additional challenges when it comes to financial impropriety. The primary mechanisms for detecting and deterring corruption and fraud, such as auditing, adherence to accounting principles, and criminal and civil liability for corrupt individuals, require strong institutional support as well as accountability when rules or norms are violated. Aid organizations that serve these countries face these challenges on the ground but also have strong incentives not to report their own failures, for fear of losing the support of donors. These agency issues, combined with the weak institutional environments in developing countries, have made the application of traditional anti-fraud policy in the development aid space largely unsuccessful.

In this paper, we provide a partial solution to the challenges of monitoring development aid expenditures and many other forms of reported data from countries and organizations in weak institutional environments. Digit analysis analyzes the patterns of reported data to detect fraud,

relying on the fact that humanly-generated data are different from naturally-occurring data. Humans face incentives to manipulate the data, as well as behavioral biases when producing data, while naturally-occurring data follow Benford's Law. We build upon earlier digit analysis work (such as Nigrini and Mittermaier (1997) and Amiram *et al.* (2015)) to improve statistical power and present tests that better reveal suspected intent to defraud. In the developing world context, one advantage of this method is that it can detect aberrant patterns that arise when humans fabricate data, and it does not require the cooperation of potentially complicit suspects beyond mandated reporting.

We apply these statistical tests to data from a World Bank development project in Kenya. The data contain details about development aid expenditures and numbers of beneficiaries served. Qualitative information based on hundreds of interviews points to high levels of graft from this project. In response to an external complaint, the World Bank conducted a two-year forensic audit of the project (World Bank Integrity Vice Presidency, 2011). The qualitative findings were confirmed by the forensic audit. The audit revealed that the Bank's financial controls, monitoring, and existing audit mechanisms were not capturing the extreme level of suspected fraud that existed. The World Bank audit flagged 66% of the district transactions as suspicious (49% as suspected fraudulent and 17% as questionable). One outcome of our method, which is the number of statistical tests failed per geographic district, is statistically significantly correlated with the level of suspected fraudulent and questionable transactions from the forensic audit. The correspondence of our statistical tests with the forensic audit provides internal validity of the digit analysis method that, to the best of our knowledge, has not previously been reported in the literature.

Naturally-occurring data and humanly-produced data are different along several dimensions. Humans face behavioral limitations in producing numbers (Chapanis, 1995), have incentives to pad values, and respond to the economic and political environments in which they manipulate data. In contrast, naturally-occurring data follow Benford's Law, a logarithmic distribution which gives probabilities of digits in each digit place, where low digits (1, 2, etc.) are more likely to appear closer to the front of a number.

We advance the existing digit analysis and Benford's Law literatures in several ways. First, we expand the statistical power of Benford's Law goodness of fit testing by considering all digit places in one test, rather than just one or two digit places, as is the norm in previous literature. By improving statistical power, we allow for additional disaggregation and triangulation of data categories, which is crucial to pinpointing fraud. Second, the existing Benford's Law literature has focused on aberrant patterns, but has limited capacity to distinguish between strategic misreporting, which seeks to gain profit for the fraudster and subvert detection, and benign misreporting or error. This issue is driven by the fact that Benford's Law predicts digit distributions from the front of the number (e.g., first digit, second digit), irrespective of the number's value (i.e., one thousand versus one hundred thousand). Basic tests of conformance to Benford's law are not sensitive to the value of the digit being manipulated. Our test considers the value of the number and allows us to distinguish patterns consistent with profitable misreporting. We supplement our 2 new tests with 8 other tests, including tests from the existing literature, that capture economic and political incentives to steal, as well as the behavioral patterns that arise when humans fabricate data. We validate our statistical tests using extensive simulations to show that our tests can successfully detect misreporting in a way that is not

specific to this case study, and also that the patterns we uncover are not driven by benign factors such as underlying prices.

Our work reveals other important substantive findings that underscore the challenges of monitoring in development aid. First, we find significant inflation of expenditures during the 2007 Kenyan presidential election year. This is consistent with our qualitative data that World Bank funds were being syphoned into the Kenyan presidential election campaign of 2007, which is widely accepted to have been a stolen election (Gibson & Long, 2009). Moreover, our tests reveal higher levels of manipulation in harder-to-monitor types of spending, consistent with a rational crime approach (Becker, 1968) and previous empirical results (see, (Olken B. A., 2007)).

Our digit-based method for uncovering fraud is complementary to other popular forms of anti-fraud machine learning, which have focused on reported values such as the debt-to-equity ratio, or institutional details like the presence of a Big 4 auditor (Perols, 2011). Therefore, we anticipate that this method will have wide applicability for both measuring earnings fidelity as well as detecting fraud beyond just development aid, such as investment in developing markets or in U.S. public firms. Indeed, the SEC has recently warned that investors in emerging markets face risk due to limited and unreliable financial reporting (U.S. Securities and Exchange Commission, 2020). Our method can be used for ongoing monitoring to achieve early detection of irregularities, and to assist audits by guiding sample selection for deeper investigation.

Our work relates to a large body of literature that has addressed audit quality, the organizational economics of fraud, and the incentives of auditors. Auditing faces the challenge that it is costly, and also that auditors are often employed by the very people that they monitor, generating conflicting incentives to report suspected impropriety. Goldman and Barlev (1974) discuss threats to auditor independence and the conflicts of interest they face, particularly from

management that controls their employment and wants a favorable report. Their paper came early in a large literature on auditor independence, which has since paid much attention to the financial and public sectors, but less to international aid flows, where the problem is arguably more dire. Our paper also sheds some light on the problem of audit quality in weak institutional environments. Krishnan *et al*. (2006) show that misreporting among nonprofits is driven by managerial incentives and disciplined by the use of outside accountants. In a recent paper on auditor and client relationships, Cook *et al.* (2020) demonstrate that the reputations of auditors tend to match with the reputations of the clients they serve.

Lamoreaux *et al*. (2015) find that World Bank development aid loans are higher for countries with better accounting quality, but accounting issues are overlooked in areas of strategic importance for U.S. interests; our work provides a micro-foundation for those macro results. Andersen *et al.* (2022) provide evidence of offshoring of World Bank funds; we discuss this paper, and World Bank attempts to suppress it, in the next section. In other related work, Duflo *et al*. (2013) provide an example of auditor capture in the developing world and show that monitoring of monitors is an effective way to combat fraud. Our method addresses the limitations of auditing that these literatures identify.

Digit analysis and Benford's Law have generated a long literature of statistical methods. We discuss this literature fully and its relationship to our analysis in Section 3.3.

## 2. Auditing and Development Aid

From 2010 to 2020, aid to developing countries totaled $1.7 trillion (OECD, 2022). Developed nations around the world make sizeable investments in projects to promote growth and development in poor countries. They do this through bilateral aid, such as USAID, and

through multilateral aid such as the World Bank and the European Union. The U.S. alone spends about $45 billion per year on these endeavors. In short, aid is a major worldwide industry, but with vastly different oversight incentives and institutions than the for-profit sector.

Development aid faces a fundamental challenge in fighting graft: aid money flows to the poorest parts of the world, which have the weakest institutional environments, greatly increasing the risk of embezzlement. Figure 1 shows the correlation between net aid flows and the Worldwide Governance Indicator measure of the perception of corruption levels by country in 2019 (Kaufmann & Kraay, 2020) (The World Bank, 2019). This figure makes two points. The slope of the linear regression between log aid dollars and corruption control is -0.95, ($p = 0.000$, 95% confidence interval [-1.3, -0.6]), indicating a statistically significant correlation. Moreover, 92% of aid flows to countries with a below-mean corruption control measure, indicating the scale of the threat that aid dollars face.

[FIGURE 1 HERE]

Monitoring mechanisms such as auditing are used to control and detect fraud in a variety of contexts (Anderson, Francis, & Stokes, 1993), and empirical evidence has shown that increased auditing is effective at eliminating graft in development aid (Olken B. A., 2007). The World Bank has historically relied on internal investigations and monitoring tools such as whistleblower hotlines and internal audits as its primary anti-fraud mechanisms (Aguilar, Gill, & Pino, 2000). However, the usefulness of these tools relies on the ability and willingness of development aid staff or beneficiaries to make internal reports, conduct investigations, disseminate those findings,

and take corrective action. Management must also make sufficient funds and staffing available to ensure adequate monitoring.

From a practical standpoint, there are many reasons why audits in developing contexts are challenging. Development aid projects span a variety of sectors, and include infrastructure building, goods and equipment, services such as health care or child education, and trainings for beneficiaries to improve their human capital on areas such as agriculture. These projects, which generally reimburse costs, face serious monitoring challenges. Infrastructure projects, such as the construction of a school or a well, can face issues with low quality material or over-invoicing. Auditing the quality of materials is challenging and may necessitate a quantity surveyor (Olken B. A., 2007). This is particularly difficult when the projects occur in rural, dangerous, and hard to access parts of developing countries. Trainings and services produce even less physical evidence, as the good produced is intangible human capital, attested to by beneficiaries who may be difficult to find, and can face retaliation from the project for negative statements to outside monitors.

Development organizations face conflicts of interest. Development organizations often depend upon the field-supervision of outside experts who are typically chosen by the staff member overseeing the project. Their employment on future missions may depend upon reports favorable to the project. Routine financial management and auditing is usually handled internally by understaffed departments. The World Bank Integrity Vice Presidency (INT) is responsible for the Bank's fraud investigations, and similar responsibilities are held by the Office of the Inspector General for USAID (OIG-USAID). In Fiscal Year 2021, World Bank INT received 4,311 complaints, but opened only 347 investigations, and produced only 35 sanctions or settlements (World Bank Group Sanctions System , 2021). Similarly, in Fiscal Year 2021, the

OIG-USAID reported $4.9 billion in audited funds out of its $19.6 billion budget, with only 142 investigations closed (U.S. Agency for International Development Office of Inspector General, 2021) (U.S. Agency for International Development Office of Inspector General, 2021). This paper uses data from a rare forensic audit of the World Bank: according to the then head of anti-corruption investigations at the World Bank (Stefanovic, 2018), no other field-verified, transaction-level, forensic audit of this scope has taken place for any World Bank project before or since this one. This is the only such audit on the World Bank Internal Investigations website (World Bank Integrity Vice Presidency, 2011).

A primary factor in the low rates of auditing in developing contexts is the lack of incentives to monitor, and the direct incentives not to disclose negative findings. The World Bank and other development organizations rely on funding from developed nations; in the U.S., aid is appropriated by Congress. Congress therefore faces a classic principal-agent problem, and do so under information asymmetry, as they are unable to properly monitor the effectiveness of these aid organizations. Aid is in this way a credence good (Dulleck & Kerschbamer, 2006): the principal, developed countries, must rely on the agent, the development aid organization, both to *administer* the aid, and also to *monitor their own performance*. And yet, when development aid organizations uncover waste, fraud, or abuse, they stand to lose the support of donors, and therefore face strong incentives to hide the results of their findings, or not find fraud in the first place.

Recent literature addresses the issue of the subversion of aid at a macro level and shows the incentives of development aid organizations to suppress these facts. Andersen, Johannesen and Rijkers (2022) show that aid disbursements to countries correspond to increases in deposits in offshore financial havens known for secrecy, amounting to 5-7.5% of aid flows. However, the

author of that study, Bob Rijkers, is an employee of the World Bank, and his attempts to publish this piece were initially blocked by World Bank officials. World Bank employees and consultants are contractually bound to receive approval prior to publishing. In this controversial case, the Bank's Chief Economist resigning unexpectedly and shortly following this incident (Jones, 2020), (The Economist, 2020). This case underscores both the magnitude of fraud in development aid as well as the missing incentives for development aid organizations to effectively monitor themselves and disclose their negative findings.

Qualitative data also point to high levels of graft and low levels of anti-fraud measures in development aid projects. Appendix B presents data from interviews concerning the World Bank Arid Lands project in Kenya, which is the project examined in this paper. Similar issues have been addressed qualitatively by Jansen (2013), who discusses the lack of oversight and incentives not to disclose negative findings in a natural resource management program in Tanzania funded by the Norwegian government, for which Jansen was the program officer. Jansen identifies the lack of external monitoring, and attempts to suppress internal monitoring, that are consistent with our qualitative evidence.

This paper proposes a partial solution to these challenges of monitoring, auditing, and misaligned incentives: the use of digit analysis to monitor development aid expenditures. Digit analysis requires development aid organizations to release data that they already collect. This disclosure could be mandated by donor nations who fund development aid organizations. Digit analysis does not require the cooperation of potentially complicit subjects and can be used to detect signals of fraud and to guide deeper investigations. By mandating data transparency, rather than pushing aid organizations to audit, donors can more easily ensure compliance. Digit analysis can also be conducted by third parties, such as in-country beneficiaries, academics, anti-

corruption organizations, and donor governments, who do not face the same conflicts of interest as those within the organizations.

## 3. Research Method: Theory and Motivation

Given the challenges of monitoring development aid, digit analysis provides a method to detect fraud that requires only that data be made available. Here, we present 2 new statistical tests that provide a method for examining falsified data in development aid and beyond. We then utilize these 2 new tests in applications to the project dataset that take full advantage of their power. We complement these 4 tests with 6 additional tests (see Appendix A) that further display the incentives and behavioral limitations of those who fabricate data.

We motivate our statistical testing with a theoretical framework for the incentives of those who are tasked with producing expenditure reports. Those who report, typically bureaucrats, face a decision either to accurately report spending or to fabricate such data. The statistical properties of the observed data result from this decision, and this theoretical framework provides predictions of the differences between legitimate and fabricated data.

### *3.1 The Statistical Properties of Truthfully Reported Data*

Using a set of receipts dedicated to a single transaction, such as the construction of a classroom, an honest bureaucrat calculates the sum of all the construction related receipts and enters the total in the report. These data follow the digit patterns of natural data, as they accurately reflect the data without human interference.

Benford's Law describes the natural distribution of digits in financial data. Benford's Law is given mathematically by (Hill, 1995):

$$P(D_1 = d_1, ..., D_k = d_k) = \log_{10}\left(1 + \frac{1}{\sum_{i=1}^{k} d_i \times 10^{k-i}}\right)$$

We have, for example, the probability that the first 3 digits are "452":

$$P(D_1 = 4, D_2 = 5, D_3 = 2) = \log_{10}\left(1 + \frac{1}{452}\right)$$

In the first digit place, Benford's Law produces an expected frequency of 30.1 percent of digit 1 and 4.6 percent of digit 9. In later digit places, this curve flattens, and by the 4th digit place the distribution is nearly identical to the uniform distribution, with expected frequency 10.01 percent of digit 1 and 9.98 percent frequency of digit 9 (Hill, 1995) (Nigrini & Mittermaier, 1997). Table 1 shows the full digit-by-digit-place table of expected frequencies under Benford's Law. Datasets known to follow Benford's Law include financial data and population data, but also everything from scientific coefficients to baseball statistics (Amiram, Bozanic, & Rouen, 2015; Diekmann, 2007; Hill, 1995) (Nigrini & Mittermaier, 1997).

[Table 1 here]

The intuition behind Benford's Law is revealed if one imagines it as a piling-up effect: increasing a first digit from 1 to 2 requires a 100 percent increase, while increase from a first digit of 8 to 9 requires a 12 percent increase (Nigrini & Mittermaier, 1997). Furthermore, Benford's Law arises from data drawn as random samples from random distributions (Hill, 1995). Because numbers that have been repeatedly multiplied or divided will limit to the Benford distribution (Boyle, 1994), financial data can be expected to follow this natural phenomenon (Hill, 1995) (Nigrini & Mittermaier, 1997).

The nature of expenditure data, which are based upon sums of numerous receipts that in turn include sums and multiplication of price times quantity, provide a theoretical basis for why we can expect Benford's Law to be the appropriate null hypothesis distribution for development expenditures. Appendix C presents simulations showing that line-item totals, like the ones we analyze here, conform to Benford's Law. Moreover, across ecologically, economically, and

demographically similar regions such as those represented in our data, we should expect similar patterns of digits when reporting is conducted honestly, even if Benford's Law did not hold.

*3.2 The Statistical Patterns of Manipulated Data*

Bureaucrats have an incentive to falsify expenditure data and embezzle both for personal gain and to satisfy kickback demands from superiors. Embezzlers weigh the costs and benefits of such behavior, including the probability of getting caught and the size of the penalty, in line with a rational decision to commit crime (Becker, 1968). In addition to prosecution, the costs of getting caught may include payoffs to auditors or others who detect their fraud, or career consequences imposed by their bosses. There may also be career consequences for refusing to participate in fraud perpetrated by one's superiors; this is especially common in systemically corrupt countries. Cheating behavior may also be inhibited by personal or social values that provide disutility to dishonest behavior.

When a bureaucrat decides to fabricate data, we expect that they will manipulate the data to maximize payout and minimize the probability of detection. This can consist of a variety of behaviors. Bureaucrats falsifying reports are often subject to budget constraints for categories of expenditure but have flexibility over the value of each activity within that category; this was true in the World Bank project we analyze. Money can be skimmed either by adding line items that were never paid out (for example, ghost employees or trainings that never happened), or by padding the line items of genuine activities. Padding can take many forms, including over-invoicing arrangements with contractors, in which case the outside party was aware, or by inflating the final expense in the report, which puts a premium upon keeping the reporting secret so that the contractors, beneficiaries, and other potential whistleblowers never know the official

expenditure claimed for a project.[1] In line with a rational decision to commit fraud, we can expect that reporters increase data tampering in response to greater incentives to steal, and attempt to produce data that appear random to subvert detection. Furthermore, we expect that bureaucrats expend lower effort in subverting detection for data that are less likely to be monitored.

Bureaucrats who choose to produce false data face behavioral limitations on their ability to successfully do so. When experimental subjects are asked to produce random numbers, studies consistently show patterns of human digit preferences. In a study where students were asked to make up strings of 25 digits, their results followed neither the Benford distribution nor the uniform distribution (Boland & Hutchinson, 2000). The patterns produced by the subjects varied greatly, with individuals exhibiting different preferences for certain digits. Other experiments have shown similar results of individual digit preferences, confirming the inability of humans to produce random digits (Chapanis, 1995; Rath, 1966).

It is possible that specific digit preferences are culturally influenced, in which case it is instructive to have a culturally representative baseline for comparison. Evidence of specific digit preferences from Africa comes from an examination of African census data. A phenomenon known as age heaping occurs when people are approximating their age; demographic records

[1] There was a premium placed upon keeping reporting data private in this project, even from other high-level project officers working in the same district office. One of the authors spent 2 years negotiating with the World Bank for access to these reports and was granted access only after intervention from the U.S. representative on the Board of the Bank on the grounds that the original project document promised that these data would be made public (World Bank, 2003). Even so, only about 2/3 of the reports were ever released.

show a preference for certain ages. Many Africans of older generations do not know their exact age, and their responses to census takers represent their best approximation. This is an example of humanly-generated data that shows specific digit preferences. Among the African censuses, we see a strong preference for the digits 0 and 5, with secondary strong preferences for 2 and 8, and disuse of 1 and 9 (Nagi, Stockwell, & Snavley, 1973; UN Economic and Social Council Economic Comission for Africa, 1986). These same digit patterns occur in our data; both 0 and 5 are so heavily overrepresented that we omit them in most of our analyses and analyze only digits 1-4 and 6-9. Nevertheless, we can rule out the idea that the patterns present in our data are the result of *legitimate* digit preferences for underlying price. Appendix C presents a simulation where underlying prices are contaminated with digit preferences, and yet line-item totals, like the ones we analyze here, still conform to Benford's Law.

*3.3 Digit Analysis Literature*

Our method builds upon the widespread but previously underpowered use of digit analysis for the detection of anomalies and fraud, and we expect that our method can improve upon existing applications of digit analysis. Digit analysis has been used in accounting to measure financial statement errors (Amiram, Bozanic, & Rouen, 2015), as well as in forensic auditing, where it is used for targeting deeper investigation (Nigrini & Mittermaier, 1997; Durtschi, Hillison, & Pacini, 2004). However, these applications rely on one- or two-digit-place comparisons, often in the first, second, or last digit, which limits statistical power and can run into sample size concerns.

A set of literature has described more advanced statistical procedures to perform Benford's Law testing. Nigrini and Miller (2009) consider a second-order test of conformance to Benford's Law, which considers the difference between ranked values in a dataset, these differences

themselves being tested for Benford conformance. Da Silva and Carreira (2013) use Benford's law to find specific subsets of the data with the greatest nonconformance, to assist auditors with further investigation. Barabesi *et al*. (2018) apply digit analysis tests to detecting customs fraud using a sequential tree-structured testing procedure called serial gatekeeping, testing multiple high-level hypotheses and then lower-level single-digit hypotheses. Cerioli *et al*. (2019) apply a different method to international trade data, using corrected test statistics that account for false positives since values in international trade data may not be Benford conforming. In each of these papers, the authors tune tests for conformance to the Benford distribution to improve power or target the test or the sample based solely on the Benford's Law distribution. Our work complements these studies by also considering the political incentives to divert funds, the specific behavioral limitations of those fabricating data, and the financial incentives to pad values in valuable digit places. Our test of all digit places removes the need for sequential testing, incorporating all the statistical power of Benford's law into one easy-to-use test that itself can be disaggregated to find appropriate subsamples.

Digit analysis has also had widespread application to other areas where there is value in detecting data manipulation. Digit analysis has been used extensively in the detection of election fraud (Mebane, 2008; Beber & Scacco, 2012; Mack & Stoetzer, 2019). Other areas where digit analysis has been successfully used include in the detection of IMF data manipulation (Michalski & Stoltz, 2013), campaign finance fraud (Cho & Gaines, 2012), scientific data fabrication (Diekmann, 2007), and enumerator integrity during survey research (Bredl, Winker, & Kötschau, 2012; Judge & Schechter, 2009; Schräpler, 2011). The ever-increasing value of data leads to greater incentives to manipulate that data and has led researchers to use digit analysis in a variety of new settings.

Our analysis also contributes to the accounting and economics literature focused on monitoring, anomaly detection, and the measurement of data quality. Du *et al.* (2020) measure the fidelity of firms' reported earnings using a hidden Markov model and show that this can predict external indicators of bad accounting, specifically Security and Exchange Commission comment letters and earnings restatements. Perols *et al.* (2017) provide another method for fraud detection using data analytic methods. These authors use the reported values from accounting statements of known fraudulent firms to classify other firms as suspected fraudulent. Our method complements these existing studies; while these papers rely on the *values* of self-reported data, our method relies instead on the *patterns* of such data. As such, we expect that our measurement could be incorporated into broader models of earnings fidelity in future accounting studies. With its focus on pattern analysis, our work is similar to Purda and Skillicorn (2015), who analyze the text of annual and interim corporate reports and show that language patterns can be used for statistical detection of fraud.

## 4. Data

### *4.1 World Bank Expenditure and Participant Data*

We analyze data from the Kenyan Arid Lands Resource Management Project (World Bank, 2003). This World Bank project ran from 1993 to 2010, eventually serving 11 arid districts and 17 semi-arid districts that were added after 2003. Our digit analysis is confined to the 11 arid districts, as these districts were the most homogeneous across ecological, economic, and demographic measures. This community driven development project spent $224 million USD targeting the most impoverished people in the heavily drought-prone regions of Kenya. It funded small infrastructure (such as schools, dispensaries, and water systems), income-

generating activities (such as goat restocking), drought and natural resource initiatives, and training exercises for villagers.

The expenditure and participant data used in these analyses were extracted from quarterly electronic project reports produced by each of the 11 districts. These reports break out the expenditures and numbers of male and female participants associated with most activities undertaken by the project in a given district and year. Each line-item expenditure represents the total expenditures for that project, for example: a classroom, a goat restocking project, or a well rehabilitation.

These districts were all subject to the same project rules and the same level of monitoring. They also share many similar characteristics: their economies depend primarily upon livestock, they are among the poorest and most drought-prone in Kenya; they are remote from centers of power, sparsely supplied with infrastructure (roads, schools, health services, access to clean water, and electricity); and their populations are poorly educated. These similarities are important because they allow us to assume that there were no legitimate reasons to expect differences in digit patterns across districts. Additional details about these data, as well as qualitative details about the nature of corruption in the Arid Lands project, are presented in Appendix B.

*4.2 Forensic Audit Data*

In 2009, following an external complaint, the World Bank's Integrity Vice Presidency (INT) began a broad forensic audit of the Arid Lands project that lasted 2 years and culminated in a

public report (World Bank Integrity Vice Presidency, 2011).[2] Auditors sampled 2 years' worth of receipts for 7 districts, 5 of which were arid districts examined in this analysis. They examined 28,000 transactions. The auditors worked from actual project receipts and supporting documents, such as cashbooks, bank statements, and vehicle logs. They also travelled to the districts to conduct interviews with suppliers to verify the legitimacy of suspicious transactions. The outcome measure we use for this comparison to our own results is the percentage of suspected fraudulent and questionable transactions by district.

## 5. Digit Tests and Results

We provide a set of 10 non-overlapping tests that capture different ways in which data can be manipulated. Two of our tests are new, and two more build upon our new tests to show specific examples of data manipulation. The remaining 6 tests are variations on existing tests in the literature and are presented in Appendix A. We collect the findings of all 10 tests together and compare against those of the World Bank forensic audit. To account for multiple tests, we

[2] The World Bank referred the Arid Lands case to the Kenyan Anti-Corruption Commission after completing a joint review together with the Kenya National Audit Office, which confirmed the findings and resulted in the Kenyan government's agreement to repay the World Bank $3.8 million USD for disallowed charges (World Bank Integrity Vice Presidency and Internal Audit Department, Treasury, Government of Kenya, 2011). It is noteworthy that the Kenyan Anti-Corruption Commission did not follow up and no one from the senior management in headquarters was prosecuted or fired. Such impunity is common in systemically corrupt countries and speaks to the need for donors themselves to be more vigilant. The World Bank did refuse to renew the project in 2010, even though it already had a Board date set for a 5-year renewal.

use a Bonferroni correction: we divide our desired significance level (.05) by the number of tests (10) and set a significant level of $p = .005$, used throughout our analyses. The summary of our tests' statistical significance is presented below; full details of the *p*-value and sample size for each test are provided in Appendix A.

*5.1. Increasing Statistical Power: All Digit Places Beyond the First*

A simple, powerful test of data manipulation is conformance of the observed digits to Benford's Law. Such tests are frequently performed in a single digit place, using the first, second, or last digit place (Diekmann, 2007; Beber & Scacco, 2012). In this new test we examine multiple digit places simultaneously. Compared with single-digit-place tests, a simultaneous analysis of multiple digit places increases sample size for statistical testing and therefore vastly increases statistical power. The increase in sample size afforded by simultaneous-digit-place analysis is especially helpful when analysis can benefit from data disaggregation, resulting in low *n*. Furthermore, testing individual digit places results in multiple-hypothesis testing issues, which a two-way chi square test avoids. Additionally, we omit the first digit when conducting this analysis, because individuals tampering with data may not have complete control over the leading digit or may avoid changing it to subvert detection. This has the potential of a more powerful fraud detector because the noise of the first digit, which may have been left clean strategically, is eliminated. The first digit test alone is presented in Appendix A.3.

We use a two-way chi square test to compare the contingency table of all digit places beyond the first against the Benford distribution. We omit 0 and 5 from this analysis, which may be subject to rounding for legitimate reasons, and which we handle separately in a test for excess rounding (see Appendix A.1). For each digit place ($2^{nd}$ digit, $3^{rd}$ digit, etc.), the frequency of each digit (1, 2, 3, 4, 6, 7, 8, and 9) is compared with the expected frequencies given in Table 1.

Figures 2AB present the data of all digit places beyond the first for expenditure (Panel A) and participant data (Panel B). The data are projected onto one axis for visualization. Among the expenditure data for all districts in Figure 2, Panel A, we see a strong preference for digits 2 and 8, underreporting of 1 and 9, and overall non-conformance to the expected Benford distribution ($p = 3.9 \times 10^{-15}$). Strikingly, these same digit patterns appear in the participant data (Panel B), and the result for all district data combined is again highly significant ($p = 5.7 \times 10^{-51}$). This pattern is also consistent with the humanly generated African census pattern described earlier.

In 8 of our 11 districts, we reject the null hypothesis that all digit places conform to Benford's Law for both the expenditure data and the participant data at the $p < 0.005$ level.

[Figure 2 Here]

The lack of conformance to the expected distribution, consistency with known humanly-generated data from African census studies, and similar patterns across both expenditure and participant data, are strong indicators that these data have been tampered with. Importantly, this test of multiple digit places subsumes a test of last digits alone, which we present as a robustness check in Appendix A.5.

5.1.1 Simulations on the Power of All-Digit-Place Testing

Our new single test explores patterns in all digit places simultaneously, rather than multiple tests of different digit places, which greatly improves statistical power. An extensive literature on forensic auditing and the use of digit analysis have promoted the use of single-digit-place tests to find evidence of fraud, or to select samples of data for additional review or auditing. These tests focus on the statistics comparing a single digit, such as the first digit, second digit, or last digit, to Benford's Law (see, e.g. Nigrini and Mittermaier (1997) and (Beber & Scacco, 2012)). The Association of Certified Fraud Examiners 2018 report **Invalid source**

**specified.** provides a comprehensive overview of such analyses, encouraging auditors to compare single-digit tests to Benford's law, as well as a test of the last or last-two digit places against the uniform distribution. Here, we use a simulation to exhibit the relative power of this test as opposed to single-digit-place testing.

Our simulation proceeds as follows. We generate Benford-conforming data between 4 and 8 digits long, (i.e., between 1,000 and 99,999,999), with each of 6 simulated districts having 1,000 observations of data. We simulate 3 "bad" districts in the data, districts A, B, and C, which each have a preference for 2 digits chosen independently. For example, district A might prefer 3 and 7, while district B might prefer 2 and 5. For each bad district, each observation is originally generated as conformant to Benford's Law, but there is a 20% chance that they manipulate the data by replacing a digit in that observation with their preferred digit. There are also 3 "good" districts, D, E, and F, which produce Benford-conforming data with no digit preferences. An ideal test would be able to distinguish good districts from bad districts and successfully flag districts A, B, and C for further review, while not flagging districts D, E, and F.

To test the current standard in the literature, we first consider tests of first digits, second digits, third digits, and last digits in each of these districts. Given a desired significance threshold of $p < 0.05$, we must correct for multiple testing by dividing by the number of tests (4) and achieve a significance threshold of 0.0125.

Table 2, Panel A presents the results of these tests. The sample size for each test is 1,000. Panel A shows the issues with single-digit testing. In the first digit, no districts are statistically significant, and in the second digit, only district A stands out. No districts are statistically significant in the third digit. Pooling data from different districts similarly fails to detect aberrant patterns using these tests. In the last digit, District F is inappropriately flagged as suspicious.

Raising the statistical significance threshold back to 5%, that is, ignoring the Bonferroni correction for multiple tests, does not fix this issue; indeed, it would flag district E as suspicious in the last digit as well. District B fails no tests despite being (statistically) equally as manipulated as districts A and C.

[Table 2 Here]

Importantly, last digits here are tested against the uniform distribution, as is promoted by the literature (see Beber *et al* (2012)). District F, which has no manipulation, fails this test. The uniform distribution is generally appropriate for last digits, but last digits may have slight tendencies towards Benford's law when they are also part of short numbers. As seen in Table 1, in a 3-digit number, the last digits are third digits, which are not uniformly distributed. Here, the smallest number is 1,000, so the last digit place is the 4th digit place for these numbers. Therefore, the last-digit test here produces a false positive, and indeed is marginally significant ($p < 0.10$) for every district.

Table 2, Panel B presents an alternative testing regime, where we consider our new test of all digit places by district. This is a single test, and the appropriate statistical significance threshold is 5%. The sample sizes vary slightly because exact values are simulated, so some districts have more digits than others due to the random length of numbers. The three manipulated districts fail this test (A, B, and C), as they should, and none of the unmanipulated districts do (D, E, and F), as they should.

Appendix C.2. extends the results of this analysis to different sample sizes ($n$) and different rates of manipulation ($p$) using the same setup of 6 districts. These extended simulations show that the all-digit-places test outperforms many single-digit-place tests, with a

higher true-positive rate and a lower false-positive rate among a range of sample sizes and manipulation rates.

This simulation shows how the all-digit-places test substantially outperforms single-digit testing along many dimensions. Signals of fraud may be present in different digit places, but individual-digit-place tests fail to combine these signals in statistically powerful ways. When performing single-digit testing, each test must be compared to a significance threshold, but each test fails to incorporate corroborating information in different digit places. Our new multiple-digit-places test exactly solves this issue, improving the sample size and power of each test, and picking up digit preferences that are observable when a reporter exhibits them over different digit places.

*5.2. Strategic Intent: Padding Valuable Digit Places*

The first test demonstrates that the data do not conform to Benford's Law but does not demonstrate the directionality of how people are manipulating the digits. Evidence that data are being fabricated consistently in the direction of increasing payment to the embezzlers is important evidence of intent, which is a critical component to the distinction between fraud manipulation and accidental data error. While there may be a strong correlation between firms and individuals whose paperwork is sometimes incomplete or missing, and actual embezzlement, it is not necessarily the case that sloppy bookkeepers are misappropriating funds. This may be even more relevant in the developing world where staff are likely to be less well educated. For this reason, evidence that points to consistently profitable deviations from expected digit distributions, or evidence of strategic efforts to avoid detection, bring us a step closer to deducing intent to defraud.

As discussed in Section 3.2, bureaucrats falsifying data can be expected to inflate values in order to receive greater illicit reimbursement. We identify padding of expenditures by measuring overuse of high digits based on the monetary value of the digit place. We hypothesize that individuals fabricating data do so strategically, and therefore place additional high digits in the more valuable digit places.

Benford's Law governs the distribution of digits by the number of positions from the left (1st digit, 2nd digit). However, the value of a digit depends on the digit's position from the right (e.g., 1s, 10s, 100s place), and this value determines the incentive to manipulate a digit. Therefore, basic tests of conformance to Benford's law are not sensitive to the value of the digit being manipulated.

To overcome this limitation, we compute the expected mean under Benford's Law by digit place *from the right* (10s, 100s), using the length of the numbers in our dataset to match left-aligned digit places and right-aligned digit places. We compare the observed mean of our data to the expected mean under Benford's Law. This is a difference of means statistic, for which a positive value indicates a mean greater than the expected mean under Benford's Law. We then perform a Monte Carlo simulation of 100,000 Benford-distributed digits in each digit place, compare the difference-of-means statistic of the project data to the simulated data, and find the probability of observing our results under the Benford distribution. Appendix A.6 contains technical details of this process.

Figure 3 shows the padding tests among both World Bank and simulated data against the Benford expected distribution. The 0 line indicates the Benford mean; anything above the line represents an overuse of high digits, and anything below the line represents an underuse. The World Bank project data (Panel A) in the 10,000s place exceed 100 percent of the 100,000

simulated Benford-conforming datasets ($p = 1.0 \times 10^{-5}$). We also see a significantly high mean ($p = 2.3 \times 10^{-4}$) in the thousands place. At the district level there is statistically significant evidence of padding in the 10,000's place for 8 of 11 districts. Ten thousand Kenyan shillings was worth approximately $150 USD in 2007.

[Figure 3 here]

Perhaps the most interesting finding in Figure 3A, which points to intention to conceal, is the decline in the use of high digits as one goes from the 10,000s to the 1,000s, 100s, 10s, and 1s places. This is consistent with a strategy of padding extra high digits in the high value places and compensating by *underutilizing* high numbers in the low digit places. The human data generators may have been trying to avoid detection from an auditor or supervisor, who might otherwise have noticed the presence of too many high numbers in any given table in the report. In contrast, Figure 3B, which uses simulated data that conform to Benford's Law, show no such pattern, and the deviation from Benford's Law is randomly distributed around 0.

In sections 5.3 and 5.4 we provide examples of how our two new tests can be applied to reveal the effects of behavioral limitations (all digit places but the first) and political incentives (padding valuable digit places).

*5.3. Behavioral Limitations: Unpacking Rounded Numbers*

Project staff had an incentive to inflate the number of participants in training activities because they claimed food expenses for each participant at 100 Kenyan Shillings (about $1.50 USD) per person, per day. The authors of the annual district reports also had reason to expect that participant data would not be as carefully scrutinized as expenditure data. First, the impact of participants on expenditures was obscured because it was only one component of the full costs of a single training exercise. Second, training exercises in remote villages are very difficult to

verify because their final product is knowledge, which leaves limited physical evidence. With the threat of oversight reduced, we speculate that less effort was devoted to covering up data fabrication.

We further surmise that officers fabricating participant data may have begun with an embezzlement target in mind, undertaking low-effort fabrication and reporting a round total number of participants to meet that target. This total number of participants was then split into males and females, as was required for reporting. Therefore, we expect greater indicators of data fabrication when the total number of male and female participants sums to a round number.

To test this, we analyze the distribution of all but first digits of numbers of total participants (males and females) when their sum ends in a 0 versus a non-0 digit. We perform the multiple-digit-places-test on these two samples, as an application of our new method, using all digits beyond the first. Theoretically, the breakout of participant data by gender should show statistically identical digit distributions between these conditions. However, we see a much higher instance of 2s and 8s and low incidence of 1s and 9s when the gender specific data come from a pooled number that ends in 0 (Figure 4A, left). This pattern is consistent with humanly-generated data and not with naturally-occurring data. There is still evidence of human generation in the data when the gender total is not round, Figure 4A right ($p = 1.9 \times 10^{-6}$), but the statistical significance is even higher in the rounded data, Figure 4A left ($p = 2.6 \times 10^{-64}$ in the sample of all districts). For 8 out of 11 districts, we reject the null hypothesis that the total of male and female participant data are Benford conforming ($p < 0.005$).

The validity of this test hinges on the fact that, under Benford's Law, data from two Benford distributions where the sums happen to end in a round number still follow the Benford distribution. This is not a trivial idea; it is possible that, by conditioning on the *sum* of two

numbers drawn from Benford distributions, the digits of the data that produce that sum have some legitimate reason to come from a different distribution.

To validate this, we simulate independent Benford conforming "male" and "female" participant values between 2 and 4 digits, and sum them. We then condition on whether that sum is rounded or not. Panel B of Figure 4 shows the result of this simulation. We find no divergence from Benford's Law evident in simulated data; both the left and right panels (totals ending in 0 or not) show conformance to Benford's law. This is evidence that the patterns found in the World Bank Data (Panel A) are the result of human manipulation.

This test shows the power of the all-digit-places test for disaggregating data to pick up specific patterns consistent with the behavioral limitations of those producing data. Analyses of single digit places struggle with disaggregation due to low sample sizes. Analyzing multiple digit places solves this issue.

[Figure 4 here]

*5.4. Election Year Effects*

In forensic accounting, auditors may examine the time-dimensionality of irregular expenditures, and recent work has shown the value of such analyses in detecting corporate accounts misreporting (Cheng, Palmon, Yang, & Yin, 2022)(Fleming, Hermanson, Kranacher, & Riley Jr., 2016). Our next test makes use of our padding test to examine the timing of padded digits while simultaneously providing evidence consistent with intent to defraud.

Interview data frequently cited the connection between syphoned project funds and the controversial presidential political campaign of 2007. The association between corruption and political campaigns has also been noted in other studies (Claessens, Feijen, & Laeven, 2008). The next test partitions our data by project year to examine whether the evidence is consistent

with higher rates of embezzlement in the presidential election year 2007. We look for padding of high-digit numbers by project year by using our new padding test, with expenditure data disaggregated by year. We compare 2007 to the Benford-conforming baseline and repeat our Monte Carlo statistic by year.

As we see in Figure 5, in 2007 (the only presidential election year) there was a statistically significant overuse of high digits in valuable digit places ($p = 0.001$). This is consistent with a greater incentive to embezzle to support political campaigns during a highly controversial presidential election year that led to extreme violence (Gibson & Long, 2009).

[Figure 5 here]

This example further demonstrates the power of the padding test. It can be deployed to examine data fabrication in conjunction with time-based analysis. This test is sensitive to the patterns of humans fabricating numbers profitably, and is powerful enough to pick up signals even when disaggregating data.

*5.5 Other Tests*

Appendix A presents the results of 6 other tests that also exhibit the behavioral limitations and economic incentives expected from fabricated data. These tests are each motivated by existing digit analysis literature and include tests for first-digit conformance to Benford's Law, rounding of numbers, repeated data, increased rounding in lesser monitored expenditures, the underuse of "digit pairs" (e.g., 22 as a substring in the number 422,347), and last digits. For the study of rounding and repeats, we compare districts to each other, relying on the fact that patterns ought to be similar across districts, even when the appropriate baseline level of rounded or repeated data is unknown. These tests all corroborate that the World Bank data are highly

manipulated and allow us to graphically and statistically examine different signals of this behavior.

*5.6 Summary of Tests*

Table 3 compiles the results of all 10 tests for each district. To address type 1 error due to the number of tests we conduct, we perform a Bonferroni correction and divide our desired significance level (0.05) by the number of tests (10). This sets a significance level of 0.005. These 10 tests avoid almost all overlap and pinpoint different aspects of data manipulation. In the bottom row, we sum the number of failed tests by district, which averages 5.7 out of 10, and ranges from 3 to 8.

[Table 3 here]

## 6. Establishing Validity: Comparing Digit Analysis to The World Bank Forensic Audit and to Qualitative Data from the Field

The existence of both an independent forensic audit for this World Bank project and qualitative data from the field provides us with a unique opportunity to establish the internal validity of our new tests and to affirm the usefulness of digit analysis more broadly.

*6.1 The World Bank Forensic Audit*

The measure of failed digit tests presented in Table 3 is statistically significantly correlated with the results of the World Bank's forensic audit. Table 4 compares the results of our digit analyses by district to the results of the World Bank forensic audit (World Bank Integrity Vice Presidency, 2011). The World Bank audit found that 4 of the 5 districts for which we have both digit and audit results had 62-75 percent suspected fraudulent or questionable expenditures. In our digit analysis, we rejected the null hypothesis for those same 4 districts in 6 to 7 of our 10 digit tests. The remaining district, Tana River, had lower levels of suspected fraud in the audit

than the other districts (44 percent), and we rejected the null on 3 of our 10 tests. A Pearson's correlation test of the 5 districts for which we have both digit tests and the World Bank audit shows a correlation of 0.928, and a 95% confidence interval of [0.255, 0.995]. We reject the null hypothesis of no correlation at the 5% significance level, with $p = 0.0227$. The World Bank's forensic audit confirms the findings from our digit analysis tests.

[Table 4 here]

We also find significant digit violations in all of the unaudited districts we examine, which is consistent with the conclusions of the auditors that these problems were systemic throughout all sectors and all districts of the project. Of the remaining 6 districts that were not audited by the World Bank, we see that half (Mandera, Ijara, Baringo) have among the highest number of digit analysis violations (8, 7, and 6) in our sample. This underscores the potential gains of using digit analysis as a diagnostic for targeting costly auditing techniques to the areas of greatest suspicion.

## 6. Conclusion

Increased monitoring and oversight are important for development aid to reach its goal of helping the world's poor. Auditing development aid expenditures faces immense challenges, both in terms of the realities of auditing on the ground in challenging environments, as well as the missing incentives for development aid organizations to root out fraud or disclose negative findings.

In this paper, we present new methods specifically targeted to detect data tampering in development aid and other weak institutional contexts. These methods rely only on mandated reporting of data, something that most organizations already require. These methods require minimal cooperation from those who may be implicated in the fraud and who may have an

incentive to impede an audit. We demonstrate our methods on data from a World Bank project in Kenya. Our statistical tests rely on expenditure reports to find patterns consistent with profitable misreporting and attempts to evade detection. An independent forensic audit of the same project, as well as qualitative interviews and new simulations, correlate with our digit analysis results, lending internal validity to the method and the substantive findings. This approach can reduce the information gap that enables misreporting between a principal and an agent.

Our method involves new statistical tests that will be broadly applicable to the study of misreported data beyond the context of development aid or developing countries in general. Our new test of padding valuable digits relates aberrant digit patterns to the monetary value of the digit place and uncovers patterns consistent with profitable deviations as well as attempts to evade detection. This test has the potential to differentiate intent to defraud from benign error. It can also be applied to test for evidence of intent to defraud with the timing of expenditure reporting. Another of our new tests, employing Benford's Law to analyze multiple digit places simultaneously, provides a statistically powerful test applicable to even relatively small datasets. The ability to work on smaller sample sizes allows more multi-dimensional analyses, such as our comparisons across districts, years, and sectors.

We extensively validate our methods using simulated data. First, we show that financial data should follow Benford's Law, even when underlying prices might be contaminated with digit preferences. This rules out the hypothesis that the patterns in our dataset are benign reflections of underlying patterns of prices set by vendors. Then, we show that our tests are powerful as compared to the standard in previous literature. Our simulations also show that our new simultaneous test of multiple digits outperforms single-digit place tests, allowing the user to

disaggregate data and pinpoint fraud. Furthermore, our new test of padding is validated by data simulated from Benford's Law, which show no such pattern.

The exact battery of 10 tests that we use is not a turnkey system for digit analysis. Some characteristics of this dataset, such as the comparison of expenditure to beneficiary tests, are particular to these data. The exact set of tests that can be performed on other datasets depends on both the incentives for manipulation in that dataset, as well as the specifics of the attribute data that are available. Our tests serve as an example of the power one can achieve with these techniques and allow us to validate new and existing digit analysis tests with the overall results of the World Bank forensic audit.

Readers may be concerned that publication of these methods will provide potential fraudsters with the means to beat the monitors. They need not worry. Engineering a Benford-conforming dataset is a more challenging statistical exercise than is ensuring that digits are uniformly distributed. It would require centralization across an organization, and matching of all supporting documentation, such as coordination of date-stamped receipts, cashbooks, vehicle logs, cancelled checks, and bank statements. Furthermore, each individual instructed to fabricate data would face an incentive to self-deal, which would undercut efforts to produce aggregate results consistent with Benford's Law. Such coordination would also expose leadership to a high risk of detection. It is not possible both to pad values, consistent with theft, and to conform to Benford's Law.

The substantive findings of this project attest to the need for better measures and identification of fraud in the developing world and under weak institutional environments everywhere. The World Bank's forensic auditors determined that 66% of the district transactions they examined were suspected fraudulent or questionable. Similarly, the districts we examined

failed between 3 and 8 of the 10 digit tests that capture different dimensions of data manipulation. We demonstrate that more suspicious patterns emerge in a presidential election year, consistent with allegations that World Bank funds were illegally diverted to fund political campaigns. Our method could have been used to conduct real-time monitoring of this project to reduce potential fraud and to target the forensic audit of this project on the worst offending districts, 3 of which were missed in the World Bank audit sample.

Digit analysis is especially beneficial in any circumstance where traditional forms of monitoring are challenging or expensive, but it is applicable to any setting where individuals have incentives to fabricate data. It can be used in for-profit or other nonprofit settings, including as an additional layer of protection in traditional corporate accounting. We foresee the use of our method in a variety of new applications as well. Firms that invest in developing markets may choose to use these methods to conduct their own form of monitoring. This method can also be used to test the authenticity of data supplied by governments in compliance with international environmental and financial agreements, or to verify pollution and labor data supplied for treaty compliance. In the modern world, where big data proliferates, stronger tools to analyze these data for signs of strategic and profitable manipulation will find increasing applicability.

## Bibliography


Aguilar, M. A., Gill, J. B., & Pino, L. (2000). *Preventing Fraud and Corruption in World Bank Projects: A Guide for Staff.* Washington, DC: The World Bank.

Amiram, D., Bozanic, Z., & Rouen, E. (2015). Financial statement errors: Evidence from the distributional properties of financial statement numbers. *Review of Accounting Studies, 20*, 1540-1593. doi:10.2139/ssrn.2374093

Andersen, J. J., Johannesen, N., & Rijkers, B. (2022, February). Elite Capture of Foreign Aid: Evidence from Offshore Bank Accounts. *Journal of Political Economy,, 130*(2), 388-425.

Anderson, D., Francis, J., & Stokes, D. (1993). Auditing, directorships and the demand for monitoring. *Journal of Accounting and Public Policy, 12*(4), 353-375.

Barabesi, L., Cerasa, A., Cerioli, A., & Perrotta, D. (2018, April). Goodness-of-Fit Testing for the Newcomb-Benford Law With Application to the Detection of Customs Fraud . *Journal of Business and Economic Statistics, 36*(2), 346-358.

Beber, B., & Scacco, A. (2012). What the numbers say: A digit-based test for election fraud. *Political Analysis, 20*, 211-234. doi:10.1093/pan/mps003

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy, 76*, 169-217. doi:10.2307/1830482

Berkman, S. (2008). *The World Bank and the Gods of Lending* . Sterling, VA, USA: Kumarian Press.

Boland, P., & Hutchinson, K. (2000). Student selection of random digits. *The Statistician, 49*, 519-529.

Boyle, J. (1994). An application of Fourier series to the most significant digit problem. *The American Mathematical Monthly, 101*, 879-886.

Bredl, S., Winker, P., & Kötschau, K. (2012). A statistical approach to detect interviewer falsification of survey data. *Survey Methodology, 38*, 1-10.

Cerioli, A., Barabesi, L., Cerasa, A., Menegatti, M., & Perrotta, D. (2019, January 2). Newcomb–Benford law and the detection of frauds in international trade. *Proceedings of the National Academies of Science, 116*(1), 106-115.

Chapanis, A. (1995). Human production of "random" numbers. *Perceptual and Motor Skills, 81*, 1347-1363.

Chavis, L. (2010, November). Decentralizing development. Allocating public goods via competition.". *Journal of Development Economics 93 (2): 264-74., 93*(2), 264-274.

Cheng, X., Palmon, D., Yang, Y., & Yin, C. (2022, January). Strategic Earnings Announcement Timing and Fraud Detection. *Journal of Business Ethics*.

Cho, W. K., & Gaines, B. J. (2012). Breaking the (Benford) Law: Statistical Fraud Detection in Campaign Finance. *The American Statistician, 61*, 218-223.

Claessens, S., Feijen, E., & Laeven, L. (2008). Political connections and preferential access to finance: The role of campaign contributions. *Journal of Financial Economics, 88*(3), 554-580.

Cook, J., Kowaleski, Z., Minnis, M., Sutherland, A., & Zehms, K. (2020, August). Auditors are known by the companies they keep. *Journal of Accounting and Economics, 70*(1).

da Silva, C. G., & Carreira, P. M. (2013). Selecting Audit Samples Using Benford's Law . *AUDITING: A Journal of Practice & Theory*, 53–65.

Debowski, L. (2003). Benford's Law Number Generator. Polish Academy of Sciences, Institute of Computer Sciences.

Deckert, J., Myagkov, M., & Ordeshook, P. (2011). Benford's Law and the detection of election fraud. *Political Analysis, 19*, 245-268. doi:10.1093/pan/mpr014

DeFond, M., & Zhang, J. (2014, November-December). A review of archival auditing research. *Journal of Accounting and Economics*, 275-326.

Diekmann, A. (2007). Not the first digit! Using Benford's Law to detect fraudulent scientific data. *Journal of Applied Statistics, 34*, 321-329.

Du, K., Huddart, S., Xue, L., & Zhang, Y. (2020, April-May). Using a Hidden Markov Model to Measure Earnings Quality. *Journal of Accounting and Economics, 69*(2-3).

Duflo, E., Greenstone, M., Pande, R., & Ryan, N. (2013). Truth-telling by third-party auditors and the response of polluting firms: Experimental evidence from India. *The Quarterly Journal of Economics, 128*, 1499-1545. doi:10.1093/qje/qjt024

Dulleck, U., & Kerschbamer, R. (2006, March). On doctors, mechanics, and computer specialists: The economics of credence goods. *Journal of Economic Literature, 44*(1), 5–42.

Durtschi, C., Hillison, W., & Pacini, C. (2004). The effective use of Benford's Law to assist in detecting fraud in accounting data. *Journal of Forensic Accounting, V*, 17-34.
Ensminger, J. (2017). *Corruption in Community Driven Development: A Kenyan Case Study with Insights from Indonesia.* U-4 Anti-corruption Resource Centre. Bergen, Norway: Chr. Michelsen Institute.
Fang, H., & Gong, Q. (2017). Detecting Potential Overbilling in Medicare Reimbursement via Hours Worked. *American Economic Review , 107*(2), 562–591.
Fleming, A. S., Hermanson, D. R., Kranacher, M.-J., & Riley Jr., R. A. (2016). Financial Reporting Fraud: Public and Private Companies. *Journal of Forensic Accounting Research , 1*(1).
Fortune. (2020, March 25 2022). *Global 500*. Retrieved from Fortune.
Gibson, C. C., & Long, J. D. (2009, September). The presidential and parliamentary elections in Kenya, December 2007. *Electoral Studies, 28*(3), 497-502.
Goldman, A., & Barlev , B. (1974, Oct). The Auditor-Firm Conflict of Interests: Its Implications for Independence . *The Accounting Review , 49*(4).
Guggenheim, S. (2006). Crises and Contradictions: Understanding the Origins of a Community Development Project in Indonesia. In A. Bebbington, M. Woolcock, S. Guggenheim, & E. Olson, *The Search for Empowerment: Social Capital as Idea and Practice at the World Bank* (pp. 111-44). Bloomfield, CT: Kumarian Press.
Guggenheim, S., & Wong, S. (2005). Community-Driven Development: Decentralization's accountability challenge. In *East Asia decentralizes: Making local government work* (pp. 253-67). World Bank: Washington, DC.
Hill, T. P. (1995). A statistical derivation of the significant-digit law. *Statistical Science, 10*, 354-363. doi:10.1214/ss/1177009869
Jansen, E. G. (2013). Don't rock the boat: Norway's difficulties in dealing with corruption in development aid. In T. Soreide, & A. Williams, *Corruption, grabbing and development: Real world challenges.* Cheltenham, UK: Edward Elgar Publishing Limited.
Jones, C. (2020, February 18). The World Bank paper at the centre of a controversy. *Financial times*.
Judge, G., & Schechter, L. (2009). Detecting problems in survey data using Benford's Law. *Journal of Human Resources, 44*, 1-24.
Kaufmann, D., & Kraay, A. (2020). *Worldwide Governance Indicators.* World Bank.
Krishnan, R., Yetman, M., & Yetman, R. (2006, March). Expense Misreporting in Nonprofit Organizations. *The Accounting Review, 81*(2), 399-420.
Lamoreaux, P., Michas, P., & Schultz, W. (2015). Do Accounting and Audit Quality Affect World Bank Lending? . *The Accounting Review, 90*(2), 703–738.
Leuz, C., & Wysocki, P. (2016). The Economics of Disclosure and Financial Reporting Regulation: Evidence and Suggestions for Future Research . *Journal of Accounting Research, 54*(2), 525-622.
Mack, V., & Stoetzer, L. (2019, April). Election fraud, digit tests and how humans fabricate vote counts - An experimental approach. *Electoral Studies, 58*, 31-47.
Mansuri, G., & Rao, V. (2013). *Localizing Development: Does Participation Work?* Washington, DC: World Bank.
Mebane, W. (2008). Election Forensics: The Second-Digit Benford's Law Test and Recent American Presidential Elections. In R. Alvarez, T. Hall, & S. Hyde, *Election Fraud: Detecting and Deterring Electoral Manipulation.* Brookings Institution Press. Retrieved from http://www.jstor.org/stable/10.7864/j.ctt6wpf99
Mebane, W. (2011). Comment on "Benford's Law and the Detection of Election Fraud". *Political Analysis*(19), 269-272.
Michalski, T., & Stoltz, G. (2013). Do Countries Falsify Economic Data Strategically? Some Evidence That They Might. *The Review of Economics and Statistics, 95*, 591-616.
Nagi, M. H., Stockwell, E. G., & Snavley, L. M. (1973). Digit preference and avoidance in the age statistics of some recent African censuses: Some patterns and correlates. *International Statistical Review, 41*, 165-174. doi:10.2307/1402833
Nigrini, M. (2012). *Benford's Law.* Hoboken, New, Jersey: John Wiley & Sons, Inc.
Nigrini, M., & Mittermaier, L. (1997). The use of Benford's Law as an aid in analytic procedures. *Auditing: A Journal of Practice and Theory, 16*.
Nirgini, M., & Miller, S. (2009). Data Diagnostics Using Second Order Tests of Benford's Law. *Auditing: A Journal of Practice & Theory, 28*(2), 305-324.
OECD. (2022, March 25). *Query Wizard for International Development Statistics.* Retrieved from OECD.org: https://stats.oecd.org/qwids/
Olken, B. A. (2007). Monitoring corruption: Evidence from a field experiment in Indonesia. *Journal of Political Economy, 115*, 200-249. doi:10.1086/517935

Olken, B. A. (2009). Corruption Perceptions vs. Corruption Reality. *Journal of Public Economics, 93*(7-8), 950-64.
Perols, J. (2011, May). Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. *Auditing: A Journal of Practice & Theory, 30*(2), 19-50.
Perols, J., Bowen, R., Zimmermann, C., & Samba, B. (2017). Finding Needles in a Haystack: Using Data Analytics to Improve Fraud Prediction . *The Accounting Review, 92*(2), 221-245.
Purda, L., & Skillicorn, D. (2015, Fall). Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection. *Contemporary Accounting Research, 32*(3), 1193-1223.
Rath, G. J. (1966). Randomization by humans. *The American Journal of Psychology, 79*, 97-103.
Republic of Kenya. (2006). *Arid Lands Resource Management Project (Phase II) Tana River District Progress Report 2003-2006.*
Schräpler, J.-P. (2011). Benford's Law as an instrument for fraud detection in surveys using the data of the Socio-Economic Panel (SOEP). *Journal of Economics and Statistics, 231*, 685-718. doi:10.2139/ssrn.1562574
Stefanovic, M. (2018). Former Manager of World Bank's External Investigations Unit, Integrity Vice Presidency. (J. Ensminger, Interviewer)
The Economist. (2020, February 15). The World Bank loses another chief economist. *The Economist*.
The World Bank. (2019). *Net official development assistance and official aid received* . Retrieved from World Bank Data: https://data.worldbank.org/indicator/DT.ODA.ALLD.CD
Transparency International. (2009). *Corruption Perceptions Index*. Retrieved August 10, 2020, from https://www.transparency.org/en/cpi/2009#
U.S. Agency for International Development Office of Inspector General. (2021). *Semiannual Report to Congress April 1, 2021-September 30, 2021.*
U.S. Agency for International Development Office of Inspector General. (2021). *Semiannual Report to Congress October 1, 2020-March 31, 2021.*
U.S. Securities and Exchange Commission. (2020, April 21). *Emerging Market Investments Entail Significant Disclosure, Financial Reporting and Other Risks; Remedies are Limited.* Retrieved July 30, 2020, from SEC Public Statements: https://www.sec.gov/news/public-statement/emerging-market-investments-disclosure-reporting
UN Economic and Social Council Economic Comission for Africa. (1986). Adjustment of Errors in the Reported Age-Sex Data from African Censuses . *Joint Conference of African Planners, Statisticians and Demographers.* Addis Ababa, Ethiopia.
United Nations Inter-agency Task Force on Financing for Development. (2021). *Data update to the 2021 Financing for Sustainable Development Report, following the 13 April release of 2020 ODA data.* New York.
Wong, S. (2003). *Indonesia Kecamatan Development Program. Building a monitoring and evaluation system for a large-scale community-driven development program.* Washington, DC: World Bank.
Wong, S., & Guggenehim, S. (2018). *Community-Driven Development: Myths and Realities.* World Bank Group: Social, Urban, Rural and Resilience Global Practice.
Woodhouse, A. (2002). *Village corruption in Indonesia: Fighting corruption in the World Bank's Kecamatan Development Program.".* Washington, D.C.: World Bank.
Woodhouse, A. (2012). *Governance Review of PNPM Rural, Community level analysis.* Jakarta, Indonesia: World Bank Indonesia.
World Bank. (2003). Project Appraisal Document on a Proposed Credit in the Amount of Sdr 43.6 Million (US $60m Equivalent) to the Republic ofKenya for the Arid Lands Resource Management Project Phase Two. *C.D. Eastern and Southern African Rural Development Operations, Africa Region, 31.*
World Bank. (2007). Project Appraisal Document on a Proposed Credit in the Amount of Sdr 57.8 Million (US $86.0 Million Equivalent) to the Goverment of Kenya for a Western Kenya Community Driven Development and Flood Mitigation Project.
World Bank Group Sanctions System . (2021). *Annual Report Fiscal Year 2021.* World Bank Group.
World Bank Integrity Vice Presidency. (2011). *Forensic Audit Report: Arid Lands Resource Management Project -- Phase II -- Redacted Report.* World Bank.
World Bank Integrity Vice Presidency and Internal Audit Department, Treasury, Government of Kenya. (2011). *Redacted Joint Review to Quantify Ineligible Expenditures for the Seven Districts and Headquarters of the Arid Lands Resource Management Program Phase II (ALRMP II) for FY07 & FY08.* Washington, DC: World Bank.

## Tables

TABLE 1: EXPECTED DIGIT FREQUENCIES UNDER BENFORD'S LAW

| | | Digit Place | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| Digit | 0 | 0.0000 | 0.1197 | 0.1018 | 0.1002 | 0.10002 |
| | 1 | 0.3010 | 0.1139 | 0.1014 | 0.1001 | 0.10001 |
| | 2 | 0.1761 | 0.1088 | 0.1010 | 0.1001 | 0.10001 |
| | 3 | 0.1249 | 0.1043 | 0.1006 | 0.1001 | 0.10001 |
| | 4 | 0.0969 | 0.1003 | 0.1002 | 0.1000 | 0.10000 |
| | 5 | 0.0792 | 0.0967 | 0.0998 | 0.1000 | 0.10000 |
| | 6 | 0.0669 | 0.0934 | 0.0994 | 0.0999 | 0.09999 |
| | 7 | 0.0580 | 0.0904 | 0.0990 | 0.0999 | 0.09999 |
| | 8 | 0.0512 | 0.0876 | 0.0986 | 0.0999 | 0.09999 |
| | 9 | 0.0458 | 0.0850 | 0.0983 | 0.0998 | 0.09998 |

This table shows the expected frequency of digits in each digit place according to Benford's Law. (Nigrini & Mittermaier, 1997, p. 54)

TABLE 2: COMPARISON OF SINGLE DIGIT TESTS TO NEW ALL DIGITS TEST WITH SIMULATED DATA

| Panel A: Single Digit Tests | | | | | | | |
|---|---|---|---|---|---|---|---|
| | District A | District B | District C | District D | District E | District F | All Districts |
| First Digits | .0247 | .3849 | 0.2607 | 0.9681 | 0.5627 | 0.7321 | 0.5259 |
| Second Digits | **0.0009345** | 0.3319 | 0.3817 | 0.2 | 0.2157 | 0.1086 | 0.1378 |
| Third Digits | 0.02922 | 0.05461 | 0.4149 | 0.1289 | 0.06716 | 0.3711 | 0.1919 |
| Last Digits | **0.002284** | 0.08462 | **0.0002037** | 0.06975 | 0.0299 | **0.001027** | **1.778e-11** |
| *n* (per test) | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 6,000 |
| Panel B: All Digit Places | | | | | | | |
| | District A | District B | District C | District D | District E | District F | All Districts |
| All Digit Places | **6.885e-5** | **0.003257** | **.03098** | 0.1345 | 0.1993 | 0.411 | 0.2367 |
| *n* | 5,503 | 5,410 | 5,507 | 5,458 | 5,488 | 5,511 | 32,877 |

This table shows the result of simulated data, where single digits are tested separately (Panel A) and simultaneously (Panel B). Bolded values are statistically significant, corrected for multiple testing with a Bonferroni correction (0.05 divided by 4 tests in panel A for a significance level of 0.0125). Only districts A, B, C have manipulated data, but single-digit testing fails to detect this, while also inappropriately flagging District F in a last-digits test. Districts A, B, and C, which have manipulated data, are correctly identified by an all digit places test.

TABLE 3. SIGNIFICANCE OF DIGIT TESTS BY DISTRICT

| Fig | Digit Test | Mandera | Ijara | Wajir | Isiolo | Baringo | Garissa | Samburu | Marsabit | Moyale | Turkana | Tana | All Districts |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2A | All Digit Places Beyond the First: Expenditure | | | | | | | | | | | | |
| 2B | All Digit Places Beyond the First: Participant | | | | | | | | | | | | |
| 3 | Padding Valuable Digit Places | | | | | | | | | | | | |
| 4 | Unpacking Rounded Numbers: Participant | | | | | | | | | | | | |
| 5 | Election Year Effects: Expenditure | | | | | | | | | | | | |
| A1 | Rounding Digits: Expenditure | | | | | | | | | | | | NA |
| A2 | Repeating Numbers: Expenditure | | | | | | | | | | | | NA |
| A3 | Sector Effects: Expenditure | | | | | | | | | | | | |
| A4 | First Digit: Expenditure Data | | | | | | | | | | | | |
| A5 | Digit Pairs: Participant | | | | | | | | | | | | |
| | Number of Significant Tests p < 0 .005 (Out of 10) | 8 | 7 | 7 | 6 | 6 | 6 | 6 | 5 | 5 | 4 | 3 | |

p < 0.005 p >= 0.005

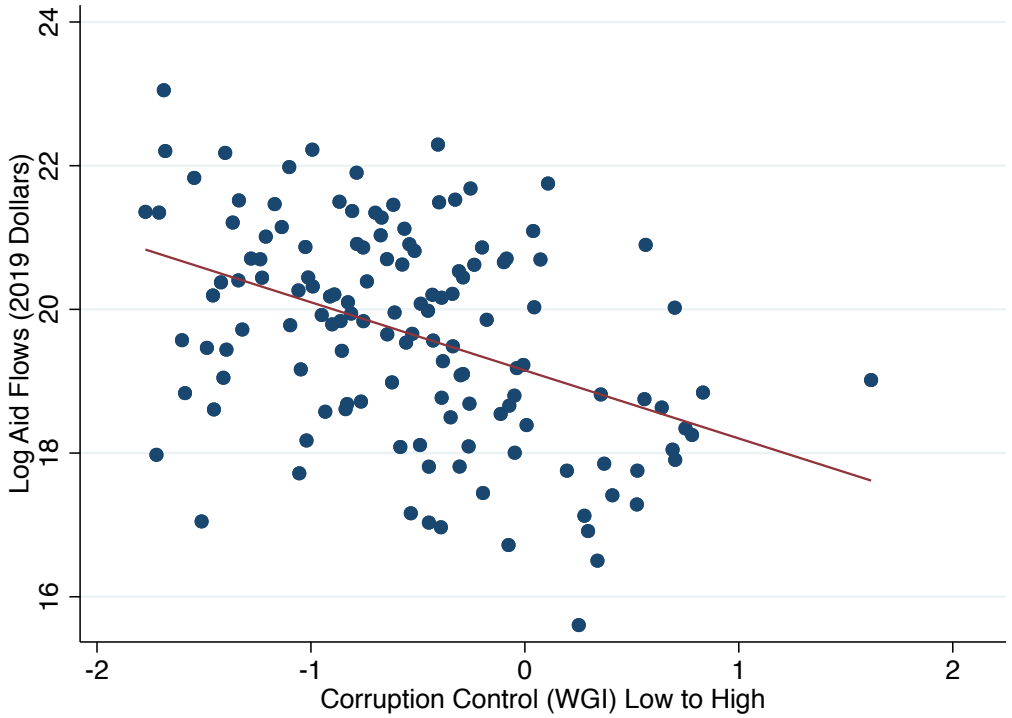
We run 10 digit tests on each of 11 districts. These tests are chosen to analyze different, non-overlapping aspects of the data. Given the large number of tests, a Bonferroni correction is used to establish 0.005 as the acceptable *p* – value for our tests. Failed tests at the 0.005 level are indicated in dark grey. Two tests, which compare rounding and repeats *across* districts, are not applicable for all districts combined. We tabulate the number of significant tests for each district in the bottom row. Exact *p*-values for each test are presented in Appendix Table A1.

TABLE 4. DIGIT TESTS BY DISTRICT COMPARED TO WORLD BANK INT FORENSIC AUDIT RESULTS

| | Digit Tests (Number Failed Out of 10) | INT Audit (Percent Suspected Fraudulent and Questionable Transactions) |
|---|---|---|
| Wajir | 7 | 75 |
| Isiolo | 6 | 74 |
| Samburu | 6 | 68 |
| Garissa | 6 | 62 |
| Tana | 3 | 44 |
| Mandera | 8 | Not Audited |
| Ijara | 7 | Not Audited |
| Baringo | 6 | Not Audited |
| Moyale | 5 | Not Audited |
| Marsabit | 5 | Not Audited |
| Turkana | 4 | Not Audited |

This table shows the number of digit analysis tests failed, out of 10, for each district in our data. Districts which fail greater levels of digit tests also have higher levels of suspected fraudulent and questionable transactions as measured by a forensic audit. A Pearson's correlation test of the 5 districts for which we have both digit tests, and the World Bank audit shows a correlation of 0.928, and a 95% confidence interval of [0.255,0.995]. We reject the null hypothesis of no correlation at the 5% significance level, with $p = 0.023$, $t$-statistic 4.33. The detailed results of each digit test are presented in Table 3 and Appendix A. The source for the INT forensic audit data is (World Bank Integrity Vice Presidency, 2011).

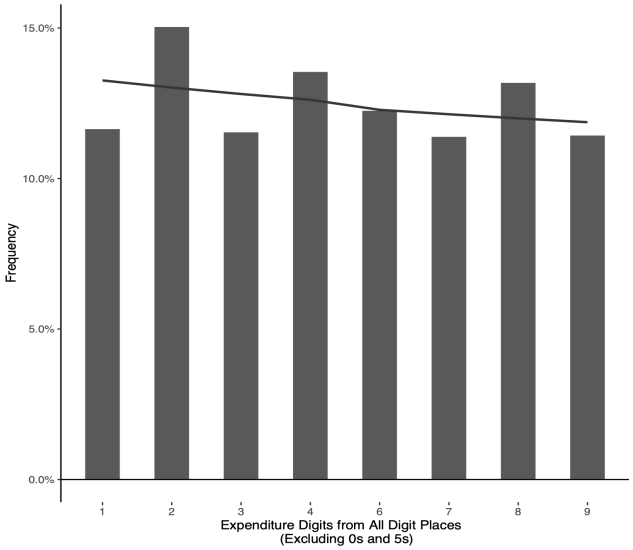
FIGURE 1: CORRUPTION CONTROL VS AID



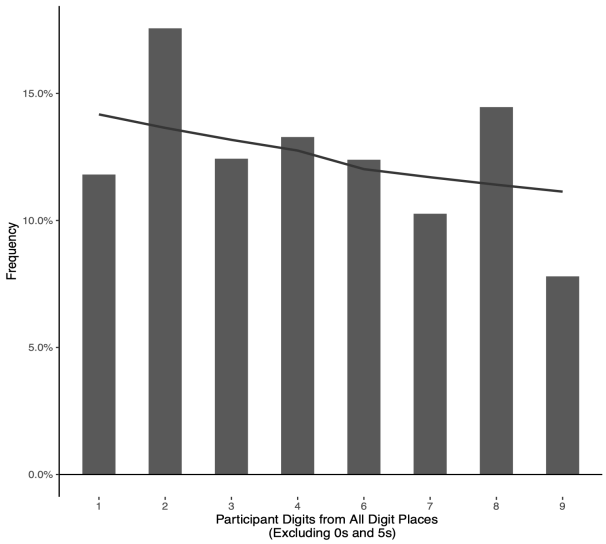

This figure plots the Worldwide Governance Indicator (WGI) control of corruption measure against log aid flows in 2019. WGI control of corruption measures "perceptions of the extent to which public power is exercised for private gain," standardized to mean 0 and standard deviation 1 (World Bank WGI, 2019); lower values correspond to lower controls and more corruption. Countries with worse corruption controls receive more aid. Of the $115 billion dollars of foreign aid to countries in these data, 92% of aid dollars flow to countries where corruption control is below the mean. The slope of the linear regression is -0.95, ($p = 0.000$, 95% confidence interval [-1.3, -0.6]). Log net aid flows are taken from the World Bank net official development assistance and official aid received and are measured in 2019 US dollars.

FIGURE 2: ALL DIGIT PLACES BEYOND THE FIRST AGAINST BENFORD'S LAW FOR EXPENDITURE AND PATICIPANT DATA

A: Expenditure Data



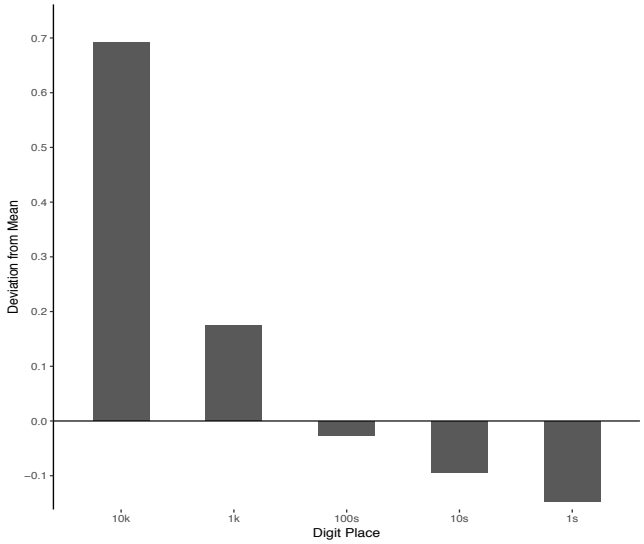

B: Participant Data



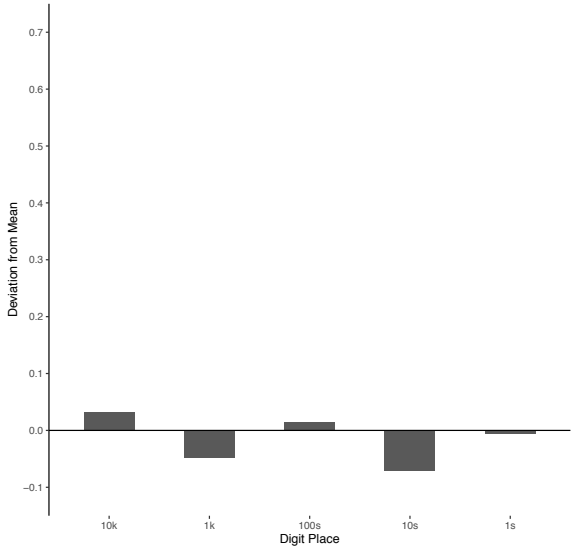

This figure presents all digits from beyond the first place, pooled, from expenditure data (Panel A) and participant data (Panel B) from all districts combined. The expected Benford's Law distribution is the solid line. Both tests are statistically significant, with $p = 3.9 \times 10^{-15}$; $n = 9371$ for the expenditure data (left) and $p = 5.7 \times 10^{-51}$; $n = 7385$ for the participant data (right). Notably, both datasets show preferences for even numbers, particularly 2 and 8. The digits 0 and 5 are omitted, due to heavy overuse that may be legitimate rounding. Tests of rounding are presented in Appendix A.

FIGURE 3: PADDING TEST OF MONETARY INCENTIVES WITH MONTE CARLO SIMULATION

Panel A: World Bank Data




Panel B: Simulated Data

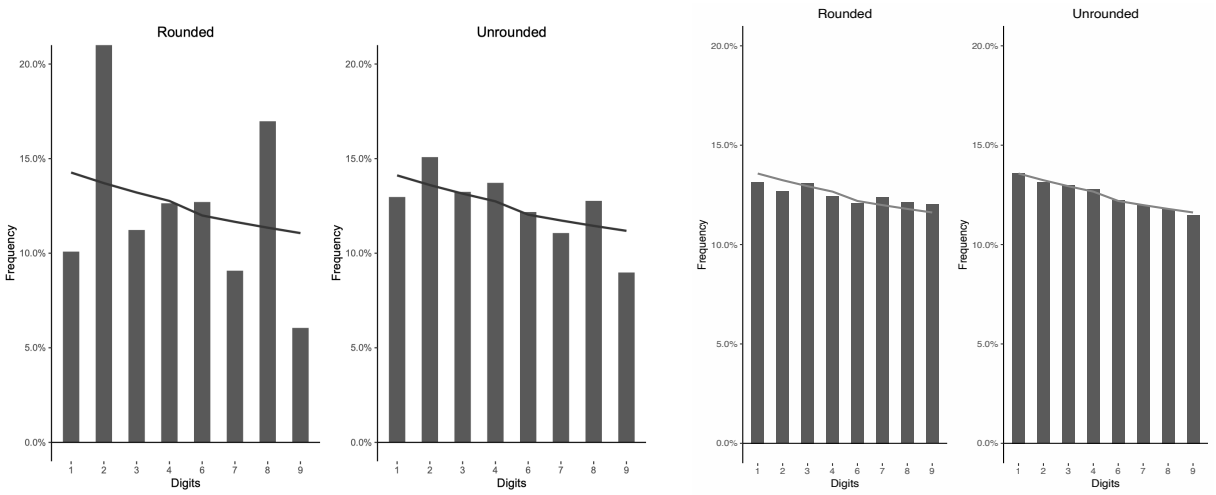



While traditional Benford's Law tests are performed from the left, i.e. first digit, second digit etc, here we test digits from the right, i.e. ones, tens, hundreds, which allows us to observe monetary incentives to pad more valuable digits. We compare the mean in each digit place from the right to the Benford expected mean in each sector. Zero reflects conformance to the Benford expected mean, and positive values indicate the mean is higher than Benford's Law predicts. The observed pattern in the World Bank Data (Panel A) is consistent with an intentional strategy of placing high digits in high digit value places and then underusing them in low digit value places to even out the digit distribution. We perform a Monte Carlo simulation of Benford-conforming datasets and compare our observed statistics to the simulated statistics to produce p-values. Compared to a sample of 100,000 simulations, using data from all sectors, we observe the following statistics for the World Bank Data: 10,000s place ($p = 1.0 \times 10^{-5}$), 1,000s ($p = 2.3 \times 10^{-4}$), 100s ($p = 0.33$), 10s ($p = 0.10$), 1s ($p = 0.061$). Panel B shows simulated Benford-conforming data with 10,000 observations. No such pattern emerges.

FIGURE 4: UNPACKING ROUNDED AND UNROUNDED DIGITS IN PARTICIPANT DATA

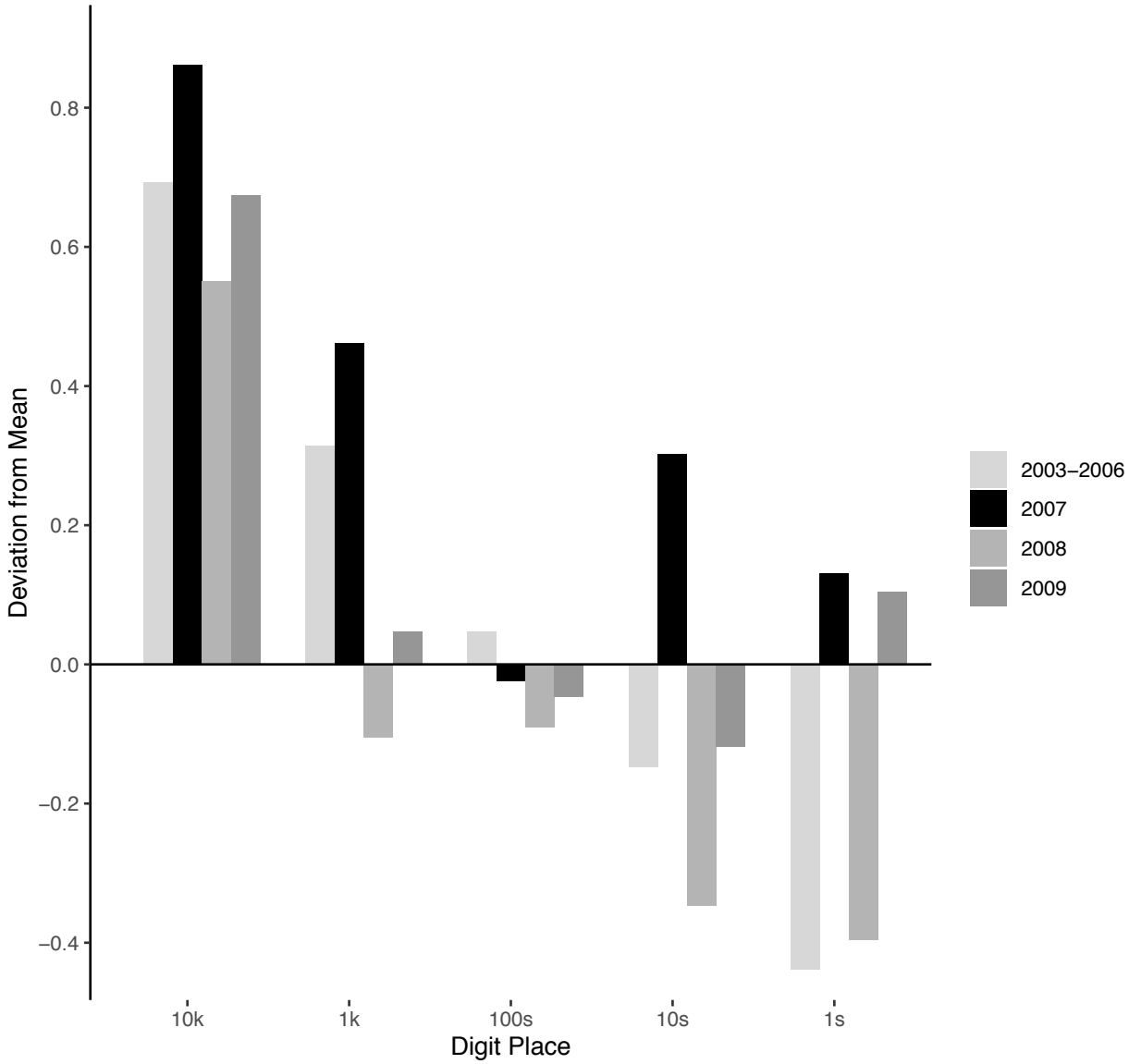
PANEL A: WORLD BANK DATA

PANEL B: SIMULATION




This presents a test of all digit places beyond the first digit among participant data (male and female pooled), when the total of male and female participants sums to a rounded number or an unrounded number. In the World Bank Data (Panel A), data that sum to a round number show higher preferences for even numbers, although both samples fail tests of conformance to Benford's Law: rounded data, $p = 2.6 \times 10^{-64}$; $n = 2975$, unrounded data, $p = 1.9 \times 10^{-6}$; $n = 4410$. We compare this to a simulation of $n = 50{,}000$ observations, where male and female numbers are generated independently in conformance with Benford's Law and then summed, and we analyze sums that happen to be rounded versus those that do not. The simulation is not statistically significantly different from Benford's Law, $p > 0.01$, and there are similar patterns between rounded and unrounded data.

FIGURE 5: ELECTION YEAR EFFECTS IN EXPENDITURE DATA




This figure performs the padding test by year. 2007 was a Presidential election year and has a statistically significant overuse of high digits in valuable digit places, even more than other years, (Ten thousands place, $p = 0.0001$; one thousands place, $p = 0.0001$).

## Appendix A: Additional Statistical Tests and Details

*A.1. Comparisons of District Patterns in Rounding and Repeating*

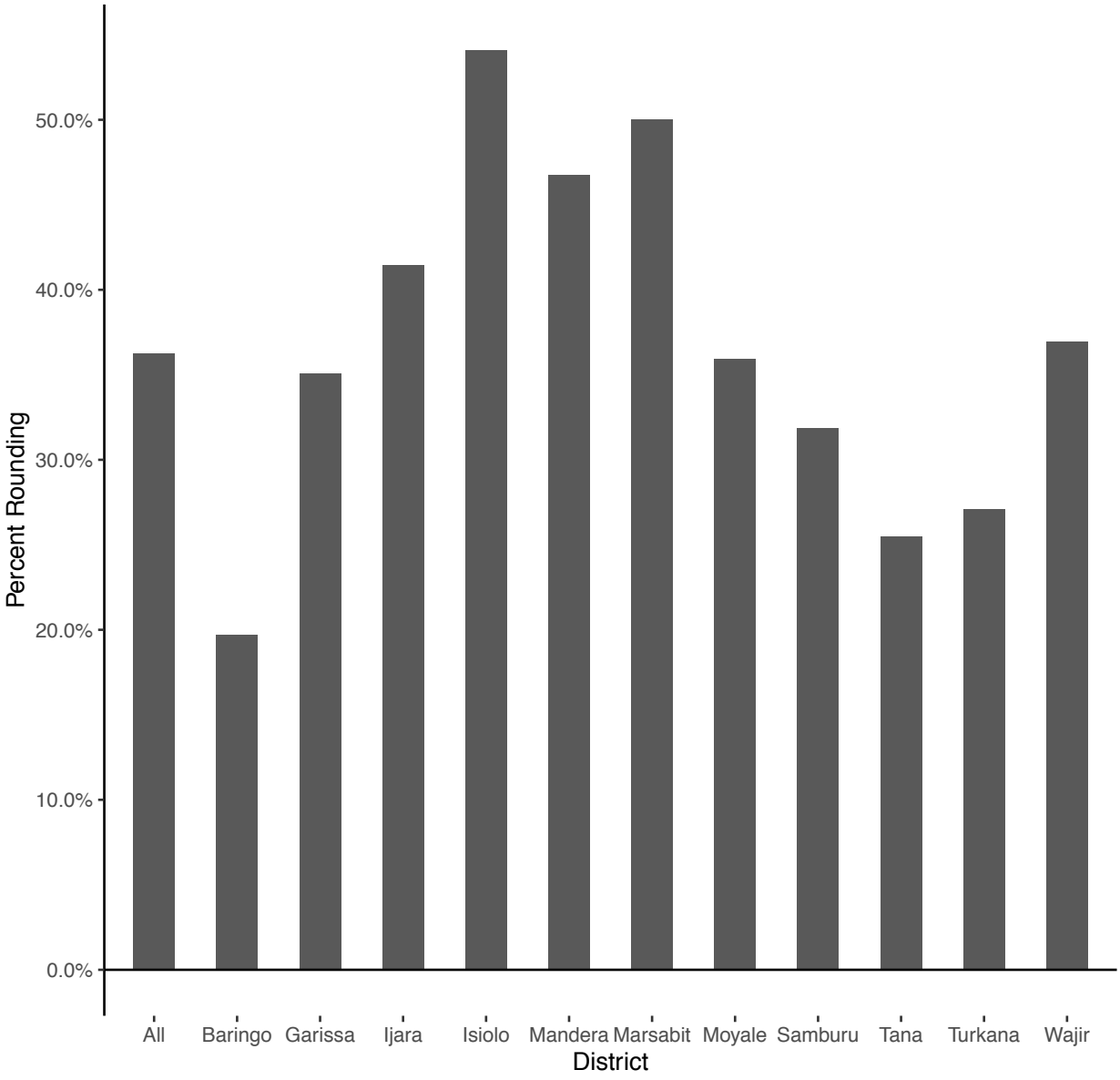
It is common for auditors to look for both high levels of rounded and repeated data, and these are often viewed as potential evidence of human tampering (Nigrini & Mittermaier, 1997). In the absence of theoretically acceptable levels of rounding and repeating, we compare districts to each other, as there is no known reason to expect differences among such ecologically, economically, and demographically similar districts.

The Kenyan shilling exchange rate was 66 Kenyan shillings to $1 USD in 2008. Its value was low enough that many receipt data would legitimately show high levels of 0s and 5s in the terminal digit place. However, one must bear in mind that these expenditure data represent sums of many receipts; it takes only one receipt ending in a non-0 or 5 to create a different terminal digit for the entire transaction, and it is these transaction totals that we are examining.

We count the number of rounded digits, tallying the number of trailing 0s (0, 00, 000, etc.), or digits in terminal strings of 5, 50, or 500, as a fraction of the number of digits in each line item. For example: the number 30,000 has 4 rounded digits out of 5 (80%); the number 12,350 has 2 rounded digits out of 5 (40%); and the number 11,371 has 0 rounded digits. Rather than indicating individual line items, counting rounded digits is a more sensitive indicator because it penalizes use of numbers such as 10,000 (4 rounded digits) more than the use of a number such as 10,600 (2 rounded digits).

Appendix Figure A1 shows the average percentage of rounded digits by district:

APPENDIX FIGURE A1: PERCENTAGE OF ROUNDED DIGITS IN EXPENDITURE DATA BY DISTRICT




This figure shows the percentage of digit places rounded in expenditure data by district. For each district, we compare the level of rounding to the level in all other districts and conduct a one-tailed t-test for excessive rounding. Ijara, Isiolo, Mandera, and Marsabit are statistically significant in their overuse of rounding as compared to other districts ($p < 0.005$).

While we don't know the empirically honest level of rounding that should occur in the dataset, there is good reason to expect that the same type of retailers, servicing the same type of contracts in economically, ecologically, and demographically similar districts, practiced the same rates of rounding. In the absence of an expected level of rounding, we compare districts to each other. For each district, we conduct a Welch's unequal variances t-test to compare the mean percentage rounding to all other districts. For example, the statistical test for Baringo compares the level of rounding in Baringo to the level of rounding in the 10 other districts combined. We conduct a one-tailed test to check for excessive rounding and define statistical significance at $p < 0.005$.

Exactly repeated numbers are also a red flag for auditors (Nigrini & Mittermaier, 1997). Our hypothesis is that embezzlers expended less effort in data fabrication when there was less reason to expect scrutiny. Repeated values are consistent with low-effort data fabrication. One such example is remote training exercises, which are particularly hard to verify.

A specific example from the Tana District Report of 2003-6 illustrates the problem of repeated data (Republic of Kenya, 2006). On page 49, we find 8 training exercises listed that took place in different villages for 3 weeks, each from March 5-27. The district had neither enough vehicles, nor enough training staff to run 8 simultaneous trainings. Among the 8 expenditures listed, we find the identical cost (245,392 Kenyan Shillings) listed for 3 different trainings, and another number (249,447) exactly repeated twice. Trainings are the summed costs of the per diems for 4-5 trainers and 1 driver (at different rates), the cost of fuel to the destination, stationary for the seminar, and 100 Kenyan Shillings per day, per trainee, for food costs. The number of trainees for each of these seminars is listed, and they range from 51 to 172.

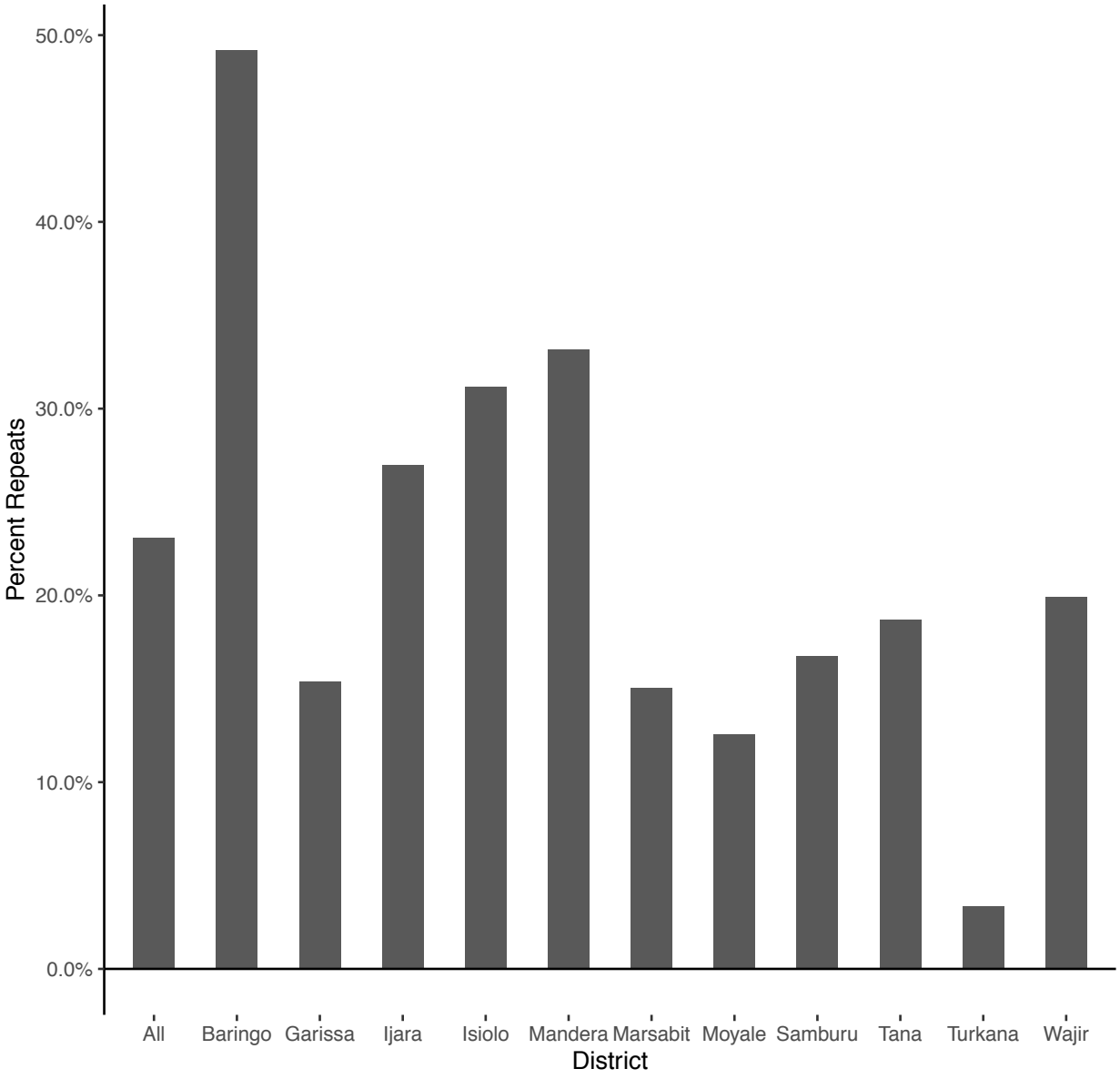
The expenses reported do not track the estimated food costs, as one would expect; indeed, the cost of food alone for 172 trainees should have exceeded all the amounts listed.

In our calculations, repeating numbers refer to the use of identical expenditure amounts for different activities. We define an exact repeat to be an expenditure matching year, district, sector, and expenditure value. There is no correction for rounding in the repeating data, as we wish to maintain the independence of our tests for rounding and repeating.

Figure A2 shows the results for the percentage of line items that repeat exactly. As with rounding, the empirically truthful level of repeating is unknown but there is no reason for patterns across districts to differ. We compare each district's average amount of rounding to all other districts, using a Welch's unequal variance t-test, and conduct a one-tailed test for excessive rounding as compared to all other districts. We see wide variation across districts: Baringo approaches 50 percent, while Turkana repeats about 5 percent. These two tests flag different districts in rounding and repeating behavior, indicating that they pick up different signals.

The difference in rates of repeating across districts may reflect differences in vulnerability to monitoring among different political territories. For example, Baringo is the home of former President Moi, who was still immensely powerful and may have provided political cover from project oversight. Systematic measurement of political "protection" is beyond the scope of this paper and is only one of several factors that may affect the sense of impunity under which district offices operated.

APPENDIX FIGURE A2: PERCENTAGE OF REPEATED ENTRIES IN EXPENDITURE DATA BY DISTRICT

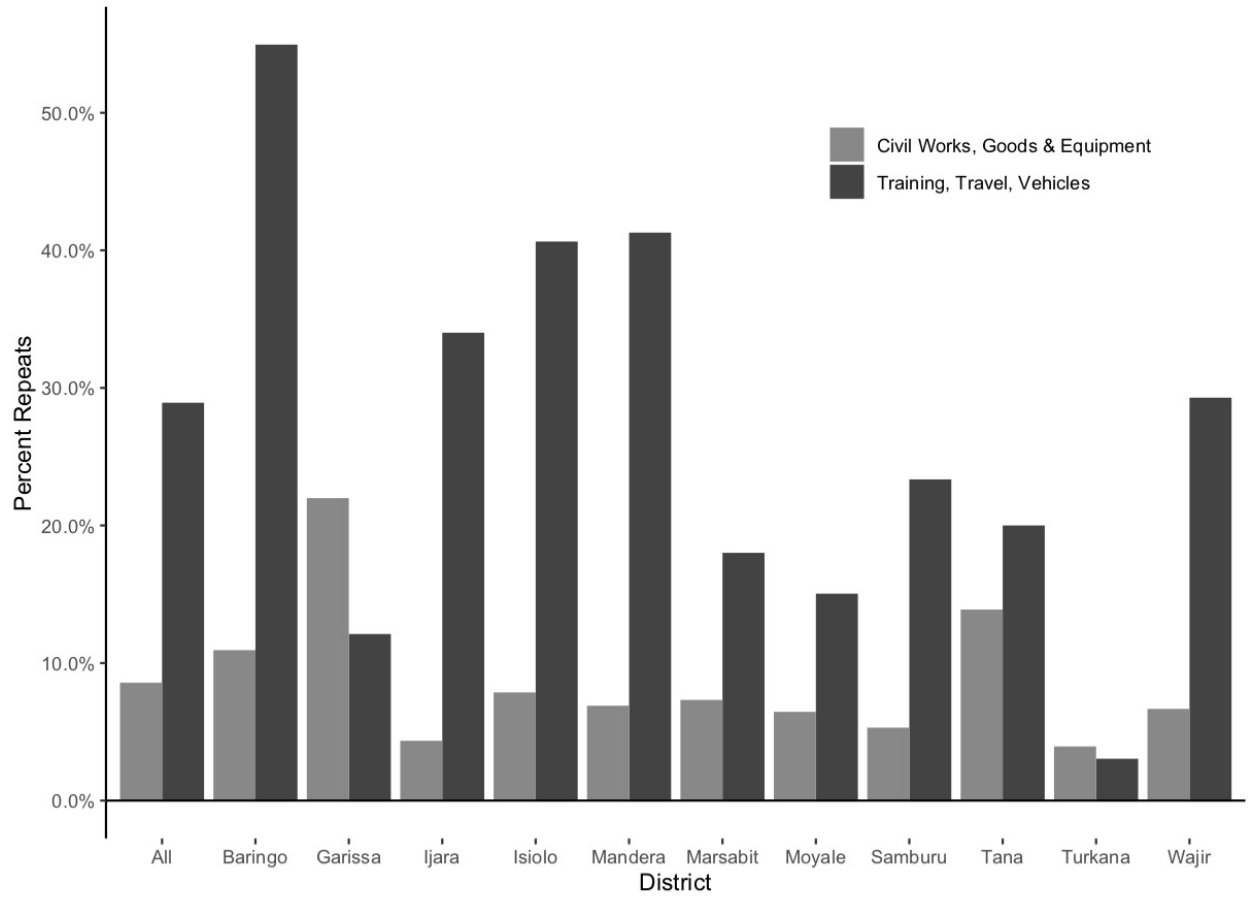



This figure shows the percent of exactly repeated expenditure entries by district for a given annual report. For each district, we compare the level of repeating to the level in all other districts and conduct a one-tailed t-test for excessive repeats. Baringo, Isiolo, and Mandera are statistically significant in their overuse of repeating compared to other districts ($p < 0.005$).

*A2. Sector Repeats*

Economic theory (Becker, 1968) and empirical work e.g. (Olken B. A., 2007) indicate that individuals are more likely to cheat when there is a lower risk of detection. The training and transport sectors of this project (travel, fuel, and vehicle maintenance) provided greater opportunities for individuals to pad expenditures relative to the civil works and goods and equipment sectors. While the latter left physical evidence of spending (such as a classroom), the former did not. For example, tracking down nomads who were reported as present for a training exercise in a remote village 2 years prior to an audit is all but impossible. Similarly, project fuel could have been diverted to private vehicles while leaving no trace. Therefore, we predict that individuals fabricating data for these sectors may have done so with less effort expended on deception. To detect this, we look for evidence of a greater incidence of repeated numbers among training, travel, and vehicle expenditures. We plot the percentage of repeated line items that match year, district, and amount, for each of the districts by sector. Appendix Figure A3 shows this result.

APPENDIX FIGURE A3: SECTOR EFFECTS IN EXPENDITURES




This figure plots the percentage of line-item expenditures repeated exactly, matching on district, year, and sector. We test whether harder-to-verify expenditures from training exercises, travel, and vehicles are more likely to be repeated than expenditures in civil works projects and purchases of goods and equipment. The districts of Baringo, Ijara, Isiolo, Mandera, Marsabit, Samburu, Wajir, and all districts combined show statistically significantly higher repeats ($p < 0.005$).

For each district, we conduct a Welch's unequal variance t-test of the number of repeats in the training and transport sector versus the civil works and goods and equipment sectors combined. Seven of 11 districts and the all-district test have statistically higher repeats in that sector. Turkana, Garissa, and Tana River Districts, where other sectors have higher percentages

of repeats, provide evidence that there is no structural reason for this phenomenon. While we don't know what the empirically honest level of repeating should be, there is no known legitimate reason for there to be more repeated line items in some districts than others. This test differs from the simple test of repeats because the sector test compares differences in repeating within a district, with the assumption that repeating should be constant across sectors. The simple repeat test compares the number of repeats across districts, on the assumption that repeating should be constant across regions that are economically, ecologically, and demographically similar.

*A.3. First Digits*

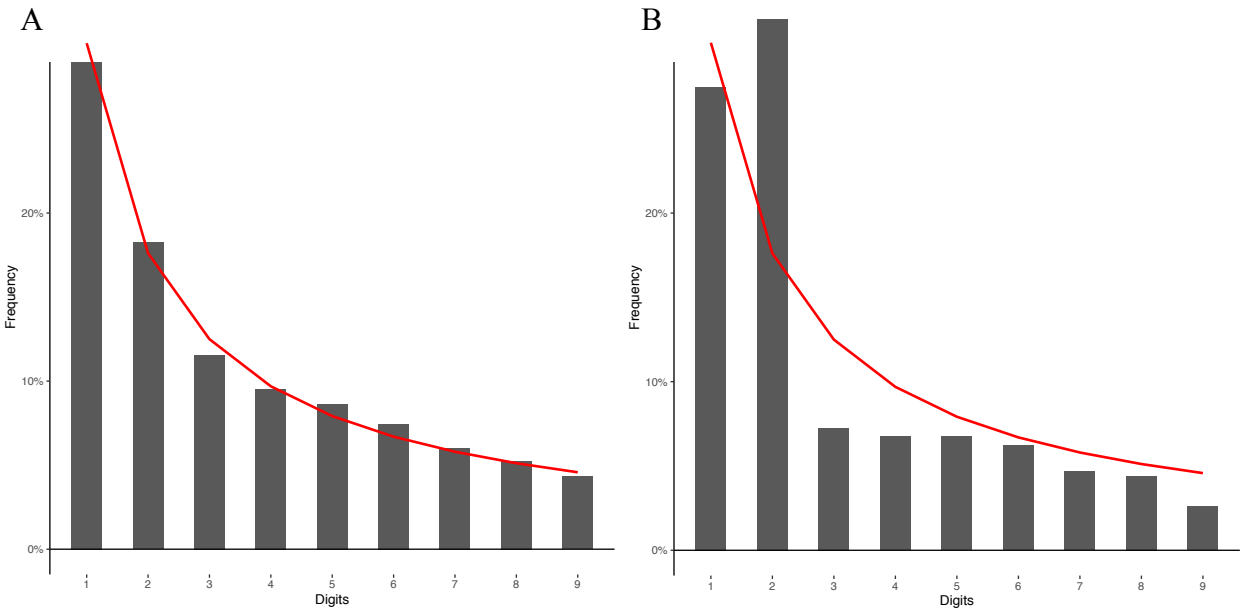
Next, we test conformance to the Benford distribution in the first digit place of the expenditure data, where we expect digits to follow (Hill, 1995):

$$\text{P(First Digit} = d) = \log_{10}\left(1+\frac{1}{d}\right)$$

Figure A4 plots this distribution as a solid line and shows the conformance of the first digits to Benford's Law. Data from the full sample of districts are not statistically significantly different from the expected distribution ($p = 0.089$) under a chi-square test. This supports the hypothesis that Benford's Law is the appropriate theoretical distribution for our dataset. Importantly, this does not necessarily mean that all the first-digit data are unmanipulated. First, people may resist tampering with the first digit to avoid detection. Second, pooled data may cancel out different individual signatures of manipulation and replicate Benford's Law (Diekmann, 2007). Appendix C reports simulations that exhibit this phenomenon; overall data conform to Benford's law, while data disaggregated by reporter may not. This becomes evident when we look at the data from individual districts where the reports were constructed. Panel B of Figure A4 shows the first digits from Ijara district, with $p = 2.3 \times 10^{-13}$. Ijara District uses the

digit 2 in the first digit place almost twice as often as predicted. Seven of our 11 districts are significantly different from Benford's Law at the $p < 0.005$ level.

APPENDIX FIGURE A4: FIRST-DIGIT EXPENDITURE DATA AGAINST BENFORD'S LAW



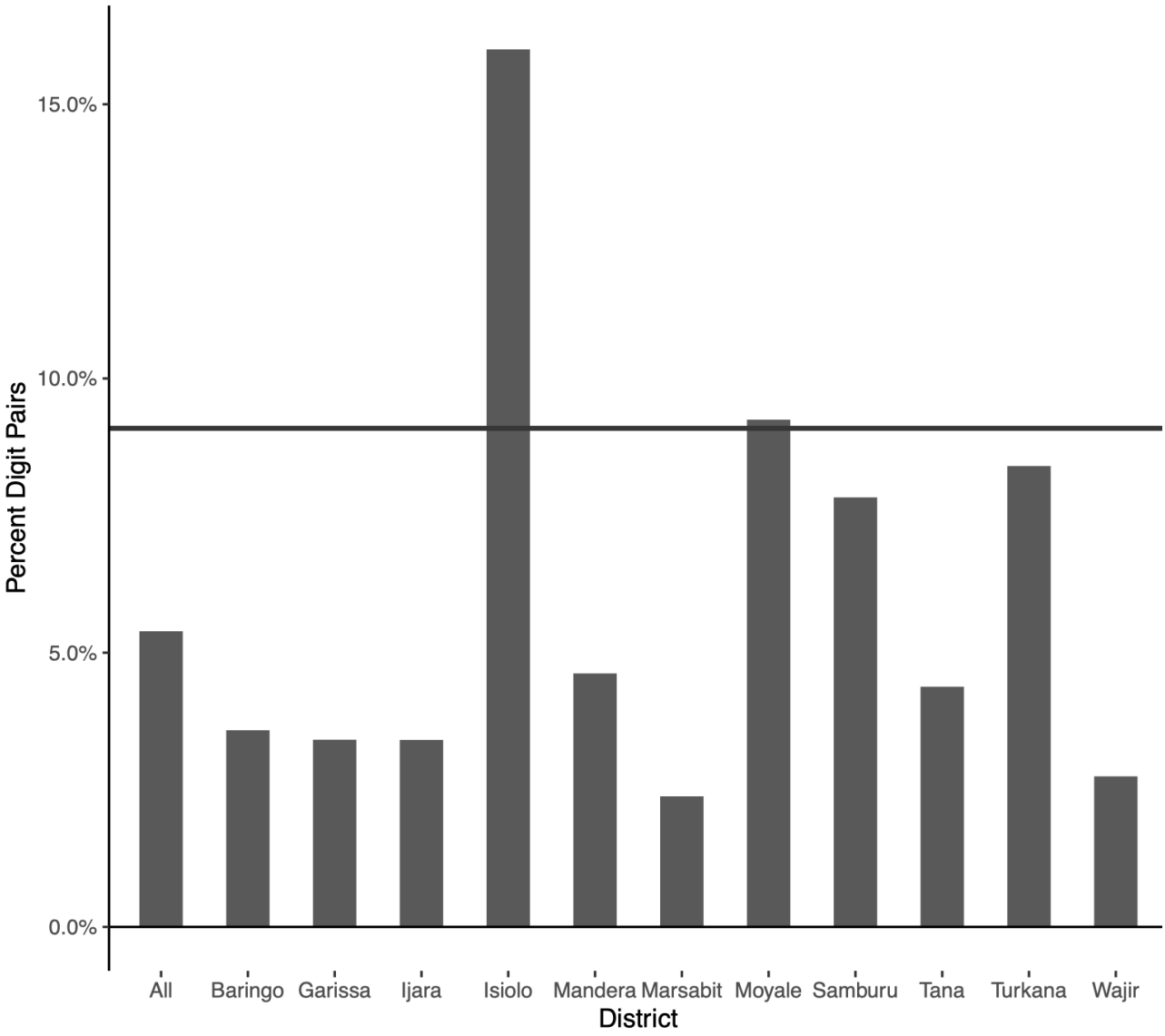

This figure presents the first-digit test as compared to Benford's Law for (A) All districts combined ($p = 0.089$; $n = 4339$) and (B) Ijara District only ($p = 2.3 \times 10^{-13}$; $n = 386$). The line represents expected distribution under Benford's Law. While the aggregate data conform to Benford's Law, individual districts can show preferences for digits. Appendix C presents simulations that support the phenomenon that digit preferences by different reporters can cancel out to Benford's Law.

*A.4 Digit Pairs*

Underuse of digit pairs, e.g., 11, 22…99, is a common feature of humanly produced data (Boland & Hutchinson, 2000; Chapanis, 1995). Other applications of digit analysis examine the last 2 digits (Nigrini, 2012), or explicitly test for digit pairs (Beber & Scacco, 2012).

Among the participant data, we expect a uniform distribution of terminal pairs, 9 of 99 pairs. We omit the pair 00 in case it is affected by rounding. We compare the observed number of digit pairs against the expected proportion using a binomial test, where the number of trials is the total combination of terminal digits observed. These data most typically record the number of women and men (listed separately) who showed up in response to an open invitation to appear for a training exercise in their village. To avoid use of first digits, we use participant data only if it has 3 or more digit places. This test is performed on the sum of male and female participants. A digit pair analysis of participant data is shown in Figure A5. Five of the 11 districts significantly underuse final-digit pairs in the participant data at $p < 0.005$ significance, as does the combined sample of all districts ($p = 2.5 \times 10^{-10}$).

APPENDIX FIGURE A5: DIGITS PAIRS IN THE LAST-TWO DIGITS FOR PARTICIPANT DATA BY DISTRICT




We test for underuse of digit pairs such as 11, 22, and 33 in substrings of numbers. Baringo, Garissa, Ijara, Marsabit, Wajir, and all districts underuse digit pairs under a binomial test with $p < 0.005$. The line represents the expected distribution of digit pairs under the uniform distribution, where 9 out of 99 substrings should be repeated values, (omitting the rounded substring 00 which is tested separately in Figure A1).

Due to the low value of the Kenyan shilling, rounding in the last digit places may be legitimate in expenditure data. Therefore, an equivalent analysis of expenditure data is not justified, as an underuse of digit pairs (e.g., 22) is confounded by a legitimate use of rounding (e.g., 20). For this reason, we confine our analysis to the beneficiary data, where there is no legitimate reason for rounding in the ones place, as participant data are reported as exact counts of people who show up.

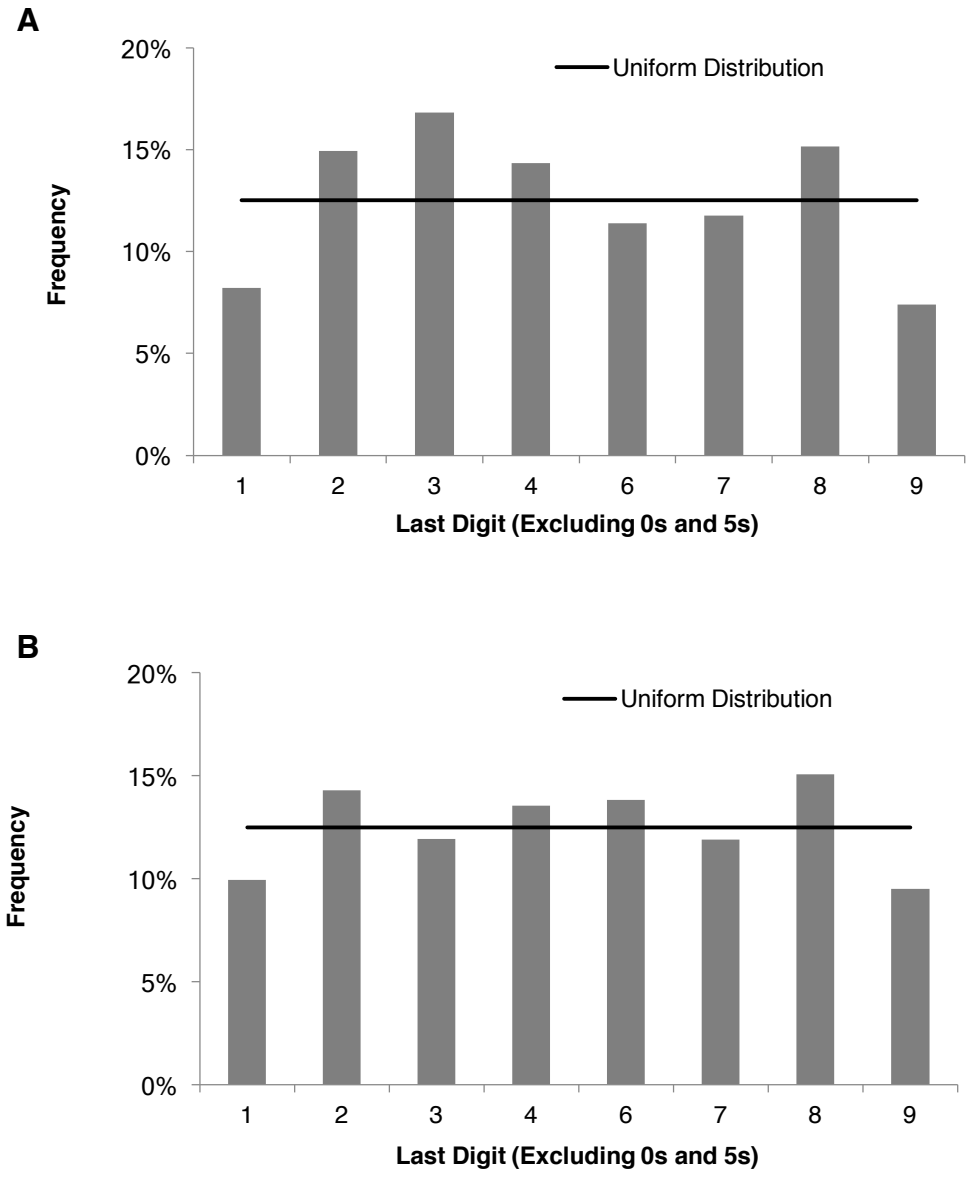
*A.5 Robustness Check (Excluded from Table 3 Summary of Tests by District): Last Digits*

Literatures on both forensic auditing and election fraud emphasize analysis of the terminal digits, which should be uniformly distributed if they represent the fourth digit place or beyond (Nigrini & Mittermaier, 1997; Beber & Scacco, 2012).

In the study of elections, the use of Benford's Law has been contested based on concerns over the distributions of data that produce voting counts (Mebane, 2011; Deckert, Myagkov, & Ordeshook, 2011; Beber & Scacco, 2012). However, these criticisms do not extend to our financial dataset or individual participant counts, both of which come from distributions that can be expected to conform to Benford's Law. Specific auditing guidelines over which types of data conform to Benford's Law includes these types of data (Durtschi, Hillison, & Pacini, 2004).

Results on the terminal digit are presented in Appendix Figure A6 and show exceptional statistical significance for both expenditure and participant data. We exclude this test from our aggregate analysis later because it is subsumed by the test of conformance to Benford's Law in our test of all digit places beyond the first. The simulation in Appendix D shows how tests of all digit places beyond the first outperform single-digit tests, including the last-digit test.

APPENDIX FIGURE A6: LAST-DIGIT EXPENDITURE AND PARTICIPANT DATA AGAINST THE UNIFORM DISTRIBUTION.




This figure presents a last-digit test as compared to the uniform distribution, which is standard in the digit analysis literature. Expenditure data (Panel A) are statistically significant, with $p = 1.5 \times 10^{-9}$; $n = 851$. Participant data in Panel B shows preferences for the same (even) digits and are also statistically significant ($p = 7.0 \times 10^{-26}$; $n = 5850$). 0s and 5s are excluded from both tests due to rounding, which is tested separately in Figure A1.

A.6 Additional Details for Padding in Valuable Digit Places Test

Figure 3 and Section 5.2 in the main text present the padding test consistent with financial incentives and its results on the World Bank Data. This statistical test checks for the presence of high digits in valuable digit places. Additional technical details about this process are presented here.

We compute the mean by digit place in the last 5 digits from the right among 5-, 6-, and 7-digit numbers. We eliminate 0s and 5s from this computation and re-weight Benford's Law as before. This gives us a mean for each of the 10,000s, 1,000s, 100s, 10s, and 1s digit places for those numbers that have all these digit places. In each digit place from the right (1s, 10s, etc.), we compute the Benford expected mean as follows: for 5-digit numbers, the Benford mean in the 10,000s place is the mean of the 1st digit; for 6-digit numbers, the Benford mean in the 10,000s place is the mean of the 2nd digit; etc. For each number length and digit place from the right, we can compute an expected mean under Benford's Law. We then combine our data from different string lengths, weighting the sample by how many numbers come from each length.

This process gives us a mean of the digit place from the right, as well as an expected mean of the digit place from the right under Benford's Law. The difference in these values is the difference in means statistic. Positive values indicate a weighted mean that exceeds the weighted Benford's Law mean, indicating padding with high-digit numbers. Negative values indicate a weighted mean that is below the weighted Benford's Law mean, indicating overuse of low digits.

To determine significance of each of our statistics, we perform a Monte Carlo simulation. We generate 100,000 observations of means drawn from the Benford distribution for the appropriate digit place. We remove 0s and 5s and compute the means by digit place from the right as well as the Benford expected mean, identically to the above. For each of the 100,000

observations, we produce a difference of means statistic. We then compare our observed difference of means statistic to these simulations. The *p*-values reported are the empirical cumulative distribution function (CDF) of our difference of means among the simulated statistics. That is, if our statistic exceeds 90% of the simulated values, its *p*-value is 0.10. For a simulation of K samples, there is a minimum *p*-value of 1/K.

We repeat this statistical process, with the data broken out by year, in Figure 5 in the main text.

Appendix A: Tables

## APPENDIX TABLE A1. SIGNIFICANCE OF DIGIT TESTS BY DISTRICT

| Fig | Digit Test | Mandera | Ijara | Wajir | Isiolo | Baringo | Garissa | Samburu | Marsabit | Moyale | Turkana | Tana | All Districts |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2A | All Digit Places Beyond the First: Expenditure | **3.6E-14** **846** | **2.6E-05** **769** | **1.9E-06** **1248** | 0.0082 437 | **7.3E-17** **1352** | **2.8E-08** **976** | 0.020 848 | **3.9E-04** **449** | **1.5E-14** **671** | 0.40 907 | **7.8E-04** **868** | **3.9E-15** **9371** |
| 2B | All Digit Places Beyond the First: Participant | **9.0E-18** **886** | **1.5E-10** **765** | **6.5E-15** **731** | **6.1E-11** **478** | **2.1E-04** **674** | **6.1E-18** **858** | **2.3E-05** **639** | 0.25 527 | 0.033 736 | **0.0037** **591** | 0.013 500 | **5.5E-51** **7385** |
| 3 | Padding Valuable Digit Places | **1.0E-05** | 0.0054 | **1.0E-05** | 0.131 | 0.024 | **0.0015** | **1.0E-05** | **1.0E-05** | **1.0E-05** | **1.0E-05** | **1.0E-05** | **1.0E-05** |
| 4 | Unpacking Rounded Numbers: Participant | **6.1E-21** **453** | **1.1E-10** **298** | **7.6E-11** **433** | **4.4E-13** **157** | 0.0085 248 | **5.9E-24** **459** | **3.9E-05** **179** | 0.014 222 | **0.0030** **179** | **3.1E-05** **205** | 0.057 142 | **2.5E-64** **2975** |
| 5 | Election Year Effects: Expenditure | 0.009 | 0.0098 | **0.0001** | 0.00605 | 0.0177 | **0.00155** | **0.0001** | 0.09075 | **0.0001** | **0.0001** | 0.01215 | **0.001** |
| A1 | Rounding Digits: Expenditure | **8.7E-32** | **1.8E-06** | 0.24 | **5.3E-33** | 1.0 | 0.86 | 1.0 | **1.9E-38** | 0.60 | 1.0 | 1.0 | N/A |
| A2 | Repeating Numbers: Expenditure | **2.6E-07** | 0.036 | 0.98 | **7.5E-04** | **4.0E-32** | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.98 | N/A |
| A3 | Sector Effects: Expenditure | **5.8E-21** **373** | **3.2E-16** **294** | **4.6E-14** **338** | **5.5E-13** **219** | **1.3E-16** **424** | 0.99 289 | **1.2E-07** **227** | **0.0035** **211** | 0.007 226 | 0.67 230 | 0.10 260 | **5.8E-69** **3091** |
| A4 | First-Digit: Expenditure Data | **1.4E-08** **489** | **2.3E-13** **386** | 0.37 578 | **5.5E-06** **308** | **1.4E-09** **488** | 0.029 430 | **5.7E-05** **359** | 0.011 293 | **1.9E-12** **319** | 0.071 357 | **0.0037** **332** | 0.089 4339 |
| A5 | Digit Pairs: Participant | 0.0070 238 | **0.0029** **176** | **4.9E-05** **255** | 1.0 125 | **5.9E-04** **251** | **1.2E-04** **293** | 0.35 166 | **0.0025** **126** | 0.59 173 | 0.48 119 | 0.030 137 | **2.4E-10** **2059** |
| | Number of Significant Tests $p < 0.005$ (Out of 10) | 8 | 7 | 7 | 6 | 6 | 6 | 6 | 5 | 5 | 4 | 3 | |

This table shows the *p*-value and sample size for each of 10 digit tests run on each of 11 districts, presenting additional details for Table 3 in the main text. The tests were chosen to analyze different, non-overlapping aspects of the data. Given the large number of tests, a Bonferroni correction was used to establish 0.005 as the acceptable *p* – value for our tests. Failed tests at the 0.005 level are indicated in bold. We tabulate the number of significant tests for each district in the bottom row.

## Appendix B: Qualitative Details from the Arid Lands Project

Qualitative research has the advantage that it can provide the substantive details necessary to understand how complex systems work. It provides the context to identify conflicting incentives and design flaws exacerbating the risk of fraud. While forensic audits and digit analysis help us identify specific instances and levels of likely fraud, they do not provide all the information required to design better monitoring systems to control future fraud. This section draws from thousands of pages of interviews with people familiar with this project's operation. To understand how the project operated on the ground, we draw upon the insider knowledge of project employees and beneficiaries, contractors, consultants, civil servants, World Bank employees, investigative reporters, politicians, and members of civil society (see Ensminger (2017) for more details).

The Arid Lands project functioned within the corrupt institutional environment of Kenya. In 2009, the Transparency International Corruption Perceptions Index ranked Kenya 146$^{th}$ out of 180 countries (Transparency International, 2009). Both then and now, Kenya qualifies as a systemically corrupt country: corruption and impunity are the norm, and the political system facilitates the theft of government resources, including those from projects such as this one.

Independent interviews and the INT forensic audit provide cross-corroborating details pointing to high-level government complicity in this theft.[1]

The specified flow of project funds was from the Kenyan Treasury to the project headquarters, then to districts and from there to the villages. According to diverse sources, the reality was that there were kickbacks flowing up and out at every level. Demands for kickbacks began with senior government officials external to the project. It is alleged that headquarters staff met some of those demands with funds embezzled from their headquarters budget. However, the project specified that the bulk of the funds had to be wired to the districts, which

[1] The relationship between the Arid Lands project and the Kenya Commercial Bank (KCB), a partially government-owned bank, serves to illustrate the kind of relationship that existed between the project and the Kenyan government. Most of the accounts of the project were held at KCB offices all over the country. The INT audit report details numerous ways in which the Kenya Commercial Bank appeared to be complicit with project staff in defrauding the program. The KCB also refused to turn over 49% of the cleared checks requested by the auditors, even though the terms of agreement between the Kenyan Government and the World Bank required cooperation with the World Bank's auditors. Among the checks that were turned over to the auditors there were many irregularities. The KCB cashed numerous checks made out to "Commissioner of VAT" that were never presented to VAT (World Bank Integrity Vice Presidency, 2011, pp. 6-7). In these cases, the original payee's name had been crossed out on the face of the check and another was substituted (World Bank Integrity Vice Presidency, 2011, p. 35). Several branches of the KCB provided altered bank statements to the auditors in which the words "cash withdrawal" were removed from the transaction description field.

posed a challenge for headquarters staff to get some of that money back. Interviewees report that this occurred in the form of monthly "envelopes" sent from districts back to headquarters as kickbacks. Some districts were able to avoid many such requests from headquarters because their districts were home to powerful national political actors who provided protection. But in many cases, this did not mean less embezzlement, just different recipients.

Even accounting for the corrupt environment in which this project operated, the fraud risk of this project was exacerbated by poor design. Two of these design flaws can be directly linked to resulting weakness in the monitoring systems: staff hiring and staff discretion in the choice of which villages received projects.

It is often said that the "tone at the top" matters. This is arguable even more the case when the surrounding institutional environment is systemically corrupt. Many of the senior staff in the Arid Lands project were seconded from their permanent ministry jobs, to which they expected to return, thus creating conflicts of interest and dual loyalty. This arrangement produced pressure to engage in fraud in collaboration with their home ministries. The project was effectively plugged directly into existing corruption networks that syphoned funds from the project upward to senior politicians and government civil servants. These features differentiate the project from a more successful World Bank community-driven development project in Indonesia (the KDP). Specifically, because the designers of the Indonesian project understood that they were operating in a similarly corrupt institutional environment, they went out of their way to create recruitment mechanisms independent of corruption centers in the government (Guggenheim, 2006).

Given the pressure on the top layer of Arid Lands management to kick funds upward, in addition to their desire for personal accumulation, it was important that they have obedient

subordinate staff beneath them, especially in the districts.[2] According to numerous sources, this was achieved by hiring staff who were underqualified for their jobs. Many did not have the minimum educational qualifications required for their jobs and were not subjected to competitive selection. Their high opportunity costs meant that they were more likely to comply with corrupt demands from headquarters.

[2] Interviewees consistently report that many Arid Lands staff at all levels were implicated in embezzlement, but they also note that not everyone participated or benefitted. Some staff and former staff were deeply troubled by what they knew was going on. Some chose to leave if they could find reasonable employment elsewhere. Others stayed but paid a steep price in terms of career advancement as a consequence of refusing to participate in the fraud. Many more were unwilling partners who were asked to do things like sign duplicate travel receipts even though they did not receive double reimbursement. The INT audit also found evidence of such double dipping. Receipts for the same activities were being submitted to both the project and the UN or other World Bank projects that were also collaborating with the project (World Bank Integrity Vice Presidency, 2011).

A second design flaw resulted from granting the district officers nearly complete discretion over the selection of villages receiving projects.[3] Project guidelines specified that selected villages would choose their own committees to manage the finances and monitor the project. In reality, the district officers were often approached by savvy villagers who agreed to collaborate with the officers in exchange for negotiated kickbacks from the village project (see Ensminger (2017) for details). As co-conspirators with the district offices, the village oversight committees aligned with the district staff against the interests of their own villagers. Many alternative designs would have improved upon this one. For example, the more successful Indonesian KDP project employed a competitive village selection model for projects (Chavis, 2010).

The design flaws in staffing and village selection contributed to many of the monitoring issues in the project. Village committees were tasked with monitoring their own projects, together with district project staff, but as we have noted, they were collaborating in the fraud. Villagers themselves faced information asymmetries and incentives that hindered

[3] In theory, there was a district steering committee (DSG) that also supervised project selection and project monitoring. Civil works projects were also supervised by government engineers who had to sign off on the plans and the work progress. According to sources, including members of the DSG and contractors from many different districts, both the DSG and the government offices that signed off on projects were compromised and ineffective (see Ensminger (2017) for more details).

whistleblowing.[4] First, it was not in the interest of either the district officers or the village committee to share the project specifications with the community. Without knowing what they were supposed to be receiving, it was impossible for villagers to know if funds were being misused. Second, the villagers were easily intimidated. The intended beneficiaries of micro projects were truly the world's poorest citizens living on less than $2 per day. They were grateful for any benefits from the project. It took years for individual villagers to begin to protest, but given the extent of complicity in the project, who were they going to complain to? Villagers who did complain were often bought off cheaply. If they persisted, the village was threatened that it would be cut-off from all future projects. This was the result of vesting monopoly discretion for the allocation of projects with the district offices; their leverage over villagers was all but absolute.

Given all the alleged embezzlement in this project, it is worth exploring how the World Bank's internal supervision processes failed to catch the ongoing fraud. Numerous Kenyan

[4] Senior staff in the Indonesian KDP project deployed several mechanisms to encourage both internal and external criticism of the project's performance, including on corruption. They designed an innovative mechanism for independent journalists to investigate and report on corruption in the project (Guggenheim, 2006) (Wong, 2003); academics were also invited to research and report on corruption (Olken B. A., 2007; Olken B. A., 2009). Finally, they commissioned several World Bank reports on corruption that were made public (Woodhouse, 2002) (Woodhouse, 2012). They used this research to inform their experimentation with different mechanisms of project design. This contrasted markedly from the secretive climate that interviewees described for the Arid Lands.

government and World Bank offices signed off on regular financial reviews. A task team leader (TTL) from the World Bank was assigned to overall supervision, and the TTL occasionally brought in missions of oversees experts. The Kenya National Audit Office conducted annual audits of all of the project's offices, and the TTL occasionally commissioned special audits from the Nairobi branch of international audit firms for subsets of districts.

One explanation for poor World Bank supervision is misaligned incentives. World Bank financial management staff, task team leaders, and outside missions are resource and time constrained. World Bank project managers themselves perceive that the Bank does not create the right incentives for them to engage in monitoring and evaluation (Berkman, 2008) (Mansuri & Rao, 2013, p. 302). To the extent that task team leaders are rewarded by the size of their project portfolios, finding evidence of large-scale fraud in one's own projects is not likely to be a career-advancing move. Conflicts of interest were also present in the task team leader's management of the outside experts brought in to provide periodic oversight. Many staff on this project commented that the same experts appeared time and again to oversee the project; they felt that fresh eyes that were less friendly with project management would have been more likely to see the problems. Outside experts may also face conflicts of interest, including real or perceived pressure to give positive evaluations to continue their relationship with the task team leader and to stay in good graces with the World Bank. These conflicts of interest are analogous to those between firms and outside auditors.

Both standard internal and external auditing of this project failed to catch most of the kinds of abuses flagged by the World Bank's forensic audit. Numerous interviewees described the friendly relations enjoyed between the project staff and the regular Kenyan auditors who visited headquarters and the districts annually. As described, there was more socializing than

examination of accounts, and the same auditors returned year after year. The project officers were less worried about professional or legal ramifications if the auditors found issues than they were that this would increase the leverage that auditors had over the office to extract a higher bribe to clean up the report. A particularly compelling report about the bribing of auditors came from a petrol station owner in a district capital: he explained that he always knew ahead of time when the project's auditors were about to arrive in town. He did business with the project and had large cash holdings. Just before the auditors arrived, the project staff would visit him to collect 200,000 Kenyan shillings (about $3000) to pay the auditors. These funds were repaid in over-invoiced petrol. The World Bank task team leader also ordered periodic audits of select districts from international firms in Nairobi. According to interviewees who were closely involved, those audits were just as unsatisfactory as the ones run by the Kenyan National Audit Office.

The qualitative investigation of this project points to many ways in which project design contributed to fraud risk and the reasons why standard World Bank supervision failed to catch it. What happened with the findings of the forensic audit speaks volumes about the enduring systemic nature of corruption in Kenyan institutions. Upon completion of their audit, the Integrity Vice-Presidency of the World Bank filed their report (World Bank Integrity Vice Presidency, 2011), conducted a joint exercise with the Kenya National Audit Office to validate their results, and also made that report public (World Bank Integrity Vice Presidency and Internal Audit Department, Treasury, Government of Kenya, 2011). In a highly unusual action, the Kenyan Government was required to repay $3.8 million USD of the inappropriately accounted funds. The World Bank also closed the project, which is highly unusual for a project that already had a board date set for its 5-year renewal. INT then submitted their supporting

audit evidence to the Kenyan Anti-Corruption Agency (KACC) for follow-up investigation. To the best of our knowledge, no further investigation was undertaken and no one from the project was indicted or prosecuted. Most senior staff are still in their posts and several of the most senior were promoted to higher level Presidential appointments upon the closing of the project.

## Appendix C: Simulations on Benford's Law

In this appendix, we show how Benford's Law is the appropriate null distribution for expenditure data even when human digit preferences may exist in underlying price data, and how our new statistical tests are more powerful tools for the detection of digit preferences than existing single-digit tests.

First, we consider whether Benford's Law is the appropriate distribution for financial data that arise when underlying *prices* are subject to digit preferences. That is, our goal is to exhibit that the evidence of misreporting in the World Bank transactions is not just a relic of manipulated underlying prices that could be a broader Kenyan phenomenon.

Janvresse and De La Rue (2004) show that Benford's law arises when data are drawn from uniform distributions whose maximum is a random number drawn from a log-uniform distribution. This "mixture" of uniform distributions of different magnitudes produces data conformant with Benford's law.

This first simulation proceeds as follows: we generate $n$ observations, where each observation is the sum of price times quantity among $k$ line items. The number of line items $k$ is different for each observation, and it is drawn from a uniform distribution between 1 and 100.

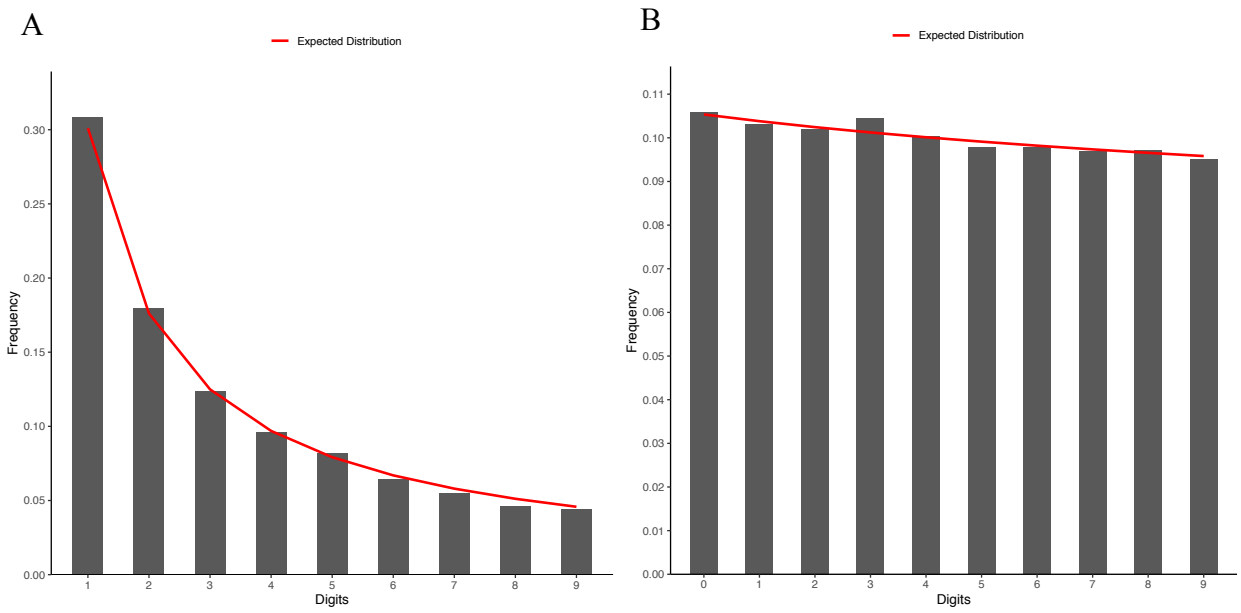
For each group of $k$ line items, we draw a maximum price from a log-uniform distribution between 1 and 10,000. Then, we draw $k$ prices from a uniform distribution with that maximum. For each price, we independently allow for contaminating digit preferences: with a 20% probability, each price has 1 digit replaced with either a 2 or an 8, reflecting a preference for even numbers. Finally, we draw a maximum price from a log-uniform distribution between 1 and 100, and draw $k$ quantities from a uniform distribution with that maximum. Quantities are

chosen without digit preference contamination. The observation is then the sum of the price times quantity values.

This setup reflects a realistic data-generating process where human preferences contaminate underlying price data, but both line items and quantities are from untampered uniform distributions. Prices from this dataset will not reflect a uniform or Benford distribution, because 20% of the data will have 1 digit replaced with either the digit 2 or the digit 8. However, in theory, the final observations should not reflect the underlying price preferences because they contain multiple line items that have been multiplied by quantities and summed.

This simulation confirms our theoretical predictions, and the reported data is still Benford-conforming. Appendix Figure C.1 shows the first-digit Benford's Law chi square test (Panel A), and the all-digits-beyond-the-first Benford's Law chi square test (Panel B). Both are Benford distributed, with $p = 0.2084$ and $p = 0.229$ respectively. Indeed, in this simulation, the most common digit in panel B is 3, not statistically significant, reflecting the fact that digit preferences for *2* or *8* in underlying price data, however legitimate, will not be reflected in the overall reported data. Rather, the high prevalence of 2 or 8 in the reported World Bank data are evidence of the manipulation of the reported data themselves.

Appendix Figure C1: Line Item Totals where Underlying Prices are Manipulated




This simulation demonstrates that even when we begin with price data that is randomly manipulated to contain extra 2s and 8s to reflect the digit preferences of vendors, once those receipts are multiplied by quantity and summed with others to arrive at a line-item entry for the project, they conform to Benford's Law. These figures test conformance to Benford's law in the first digit (Panel A), and all other digits (Panel B). Despite the preference for 2s and 8s in prices, overall data still conform to Benford's law, with $p > 0.2$ for each case. This shows that the manipulation visible in our project is not the result of unusual price preferences by vendors.

Second, we consider the importance of disaggregating data. We conduct a second simulation that asks whether manipulated data can be detected by a traditional Benford's law analysis that pools data from different reporters with different biases. Our simulation shows the value of disaggregating data and of pooling data beyond the first digit.

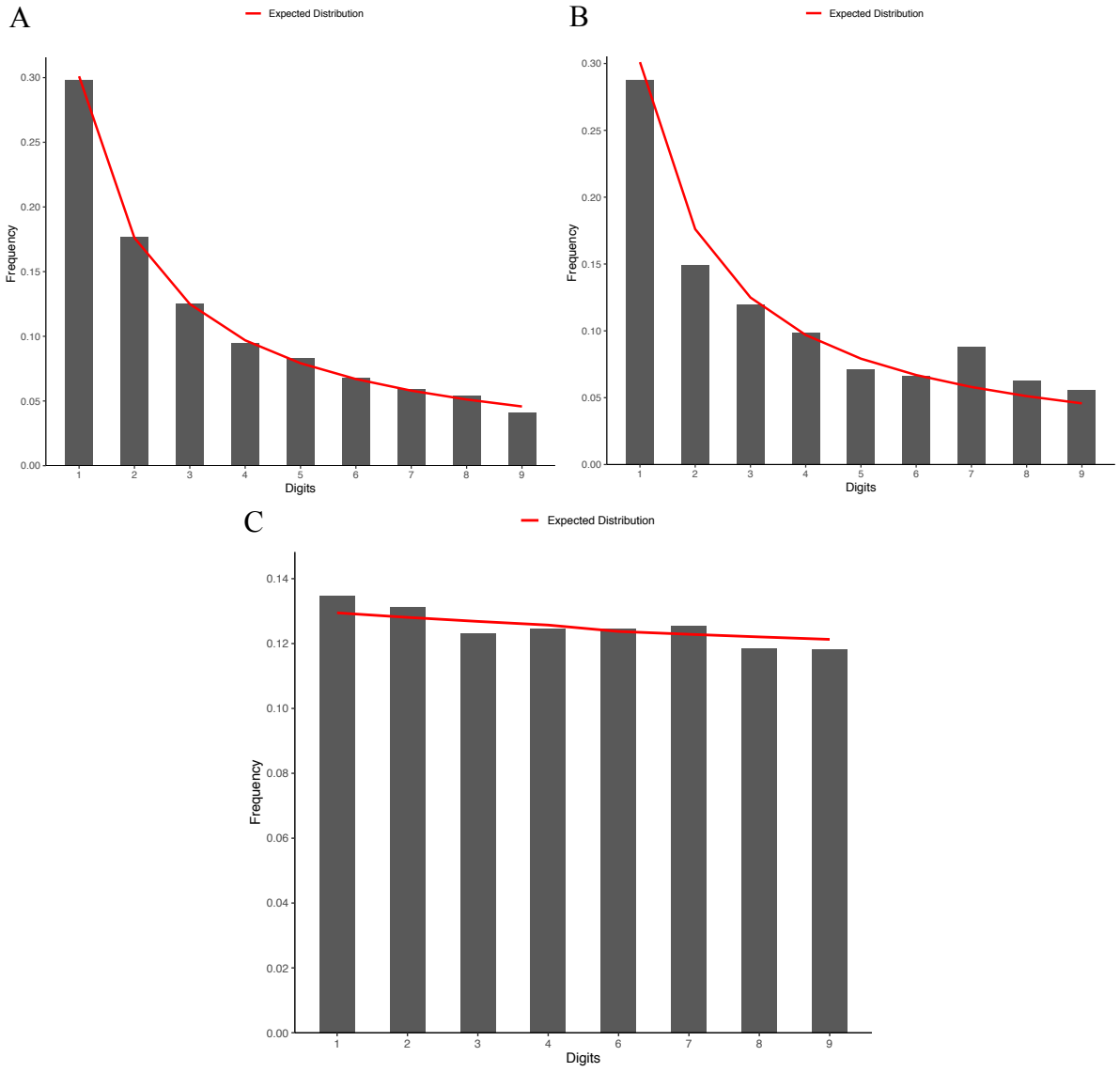
The second simulation proceeds as follows. We consider 10 "districts," representing distinct reporters, and each reports 1,000 observations, each corresponding to an item on a financial statement. We begin with Benford-conforming data for each district but allow each district's reporter to have 2 preferred digits, 0 through 9, independently chosen from each other.

With a 20% probability, each observation has 1 digit changed to a preferred digit from that district's reporter.

Appendix Figure C.2 shows the result of this second simulation. Panel A shows the first digits from all 10 simulated districts, which conform to Benford's Law in aggregate, $p = 0.2256$. Panel B shows the power of disaggregation, with the first digits of a single district, which are not Benford conforming, $p = 0.0008985$. Panel C presents the test of all-digits-beyond-the first for conformance to Benford's law from all districts with $p = 0.00487$.

This second simulation shows that, even when Benford-conforming data are contaminated by digit preferences, an overall test of first digits may fail to detect manipulation if different reporters exhibit different digit preferences that "wash out." Disaggregation, and the use of digits beyond the first place, can solve these issues and provide additional statistical power.

Appendix Figure C2: The Power of Disaggregation and Multiple Digit Places




Simulations show the effect of pooling data from 10 reporters with different digit preferences. Panel A shows the total first digit from all reporters, which is not statistically significant, $p > 0.2$. This shows that, even when data are manipulated, the effects can wash out when data are pooled and only the first digit place is considered. Panel B shows that disaggregation is powerful for detection by showing the data of one of the 10 reporters, which fails to conform to Benford's law, $p < 0.005$. This highlights the statistical power of disaggregation. Panel C shows that jointly considering digits beyond the first place is powerful, even when data are not disaggregated the manipulation is statistically significant, $p < 0.005$.

C.2 The Power of All Digit Places

Section 5.1.1 of the main text shows the power of our all-digit-places test when compared to single-digit-places tests. Here, we extend that result to show that our all-digit-place test outperforms single-digit-place tests when varying sample size $n$ or the rate at which data are manipulated, $p$.

We generate Benford-conforming data between 4- and 8-digits long, (i.e., between 1,000 and 99,999,999), with each of 6 simulated districts having $n$ observations of data. We simulate 3 "bad" districts in the data, districts A, B, and C, which each prefer 2 digits chosen independently. For each bad district, each observation is originally generated as conformant to Benford's Law, but there is a $p$ chance that they manipulate the data by replacing a digit in that observation with their preferred digit. There are also 3 "good" districts, D, E, and F, which produce Benford-conforming data with no digit preferences. An ideal test would be able to distinguish good districts from bad districts and successfully flag districts A, B, and C for further review, while not flagging districts D, E, and F. The hyper-parameters for the results presented in Table 2 of the main text are $n = 1{,}000$, reflecting 1,000 observations per district, and $p = 0.2$, a 20% rate of manipulating data among the districts that fabricate.

We consider a battery of the same 4 tests as presented in Section 5.1.1: first digits, second digits, third digits, and last digits. We vary $p$ within 0.1, 0.2, 0.3, 0.4 and 0.5, and we consider sample sizes 100, 500, 1000, 5000 and 10000. Because we are conducting 4 tests x 5 sample sizes x 5 probabilities of manipulation = 100 tests per district, we divide the desired significance level (0.05) by 100 (0.0005) to accomplish a Bonferroni correction for multiple testing.

Appendix Table C1 presents the results of these tests. We count the number of tests failed among the falsifying districts (A, B, and C) out of 12 as the true positive rate, and the

number of tests failed among the clean districts (D, E, and F) out of 12 as the false positive rate. As the sample size increases, and as the probability of manipulation increases, these tests perform better but not perfectly; indeed among 1,000 data points per district with 20% manipulation probability, single-digit tests fail only 75% of the time. Appendix Figure 3 plots the true positive rate and the false positive rate against the expected number of manipulated data points, $n \times p$. There are false positives, largely driven by last-digit testing, which as discussed in the main paper, can suffer from issues due to the last digit having some Benford rather than uniform characteristics.

In contrast, Appendix Table C2 presents the results of the same variation in $n$ and $p$ but using the new all-digit-places test. We conduct 5 sample sizes x 5 probabilities of manipulation = 25 tests per district, and so we divide the desired significance level (0.05) by 25 (0.002) to accomplish a Bonferroni correction for multiple testing. We find a very high rate of true positives, and a very low rate of false positives. Above an $n \times p$ of about 250 expected manipulated observations, the test successfully catches the manipulating districts; only very rarely are non-manipulating districts flagged (2 total times in 75 district tests). Appendix Figure C4 plots these results, showing the excellent performance of this powerful test.

The results of these simulations confirm that the all-digit-places test substantially outperforms single-digit-place testing. At all levels of manipulation and sample sizes, all-digit-places testing is higher powered, having a better true positive rate and a lower false positive rate.

Appendix Table C1: Results of Many Single-Digit Tests

| *n* | p | *n* times *p* | A Failed | B Failed | C Failed | D Failed | E Failed | F Failed | True Positive Rate | False Positive Rate |
|---|---|---|---|---|---|---|---|---|---|---|
| 100 | 0.1 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 100 | 0.2 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 100 | 0.3 | 30 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 100 | 0.4 | 40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 100 | 0.5 | 50 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 500 | 0.1 | 50 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 500 | 0.2 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 500 | 0.3 | 150 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0.08 |
| 500 | 0.4 | 200 | 1 | 1 | 1 | 0 | 0 | 0 | 0.25 | 0 |
| 500 | 0.5 | 250 | 1 | 1 | 1 | 0 | 0 | 0 | 0.25 | 0 |
| 1000 | 0.1 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1000 | 0.2 | 200 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1000 | 0.3 | 300 | 2 | 0 | 3 | 0 | 0 | 0 | 0.42 | 0 |
| 1000 | 0.4 | 400 | 2 | 3 | 2 | 0 | 0 | 0 | 0.58 | 0 |
| 1000 | 0.5 | 500 | 4 | 3 | 4 | 0 | 0 | 1 | 0.92 | 0.08 |
| 5000 | 0.1 | 500 | 2 | 1 | 1 | 1 | 1 | 1 | 0.33 | 0.25 |
| 5000 | 0.2 | 1000 | 3 | 2 | 4 | 1 | 1 | 1 | 0.75 | 0.25 |
| 5000 | 0.3 | 1500 | 4 | 4 | 4 | 1 | 1 | 1 | 1 | 0.25 |
| 5000 | 0.4 | 2000 | 4 | 4 | 4 | 1 | 1 | 1 | 1 | 0.25 |
| 5000 | 0.5 | 2500 | 4 | 4 | 4 | 1 | 1 | 1 | 1 | 0.25 |
| 10000 | 0.1 | 1000 | 2 | 2 | 3 | 1 | 1 | 1 | 0.58 | 0.25 |
| 10000 | 0.2 | 2000 | 4 | 4 | 3 | 1 | 1 | 1 | 0.92 | 0.25 |
| 10000 | 0.3 | 3000 | 4 | 4 | 4 | 1 | 1 | 1 | 1 | 0.25 |
| 10000 | 0.4 | 4000 | 4 | 4 | 4 | 1 | 1 | 1 | 1 | 0.25 |
| 10000 | 0.5 | 5000 | 4 | 4 | 4 | 1 | 1 | 1 | 1 | 0.25 |

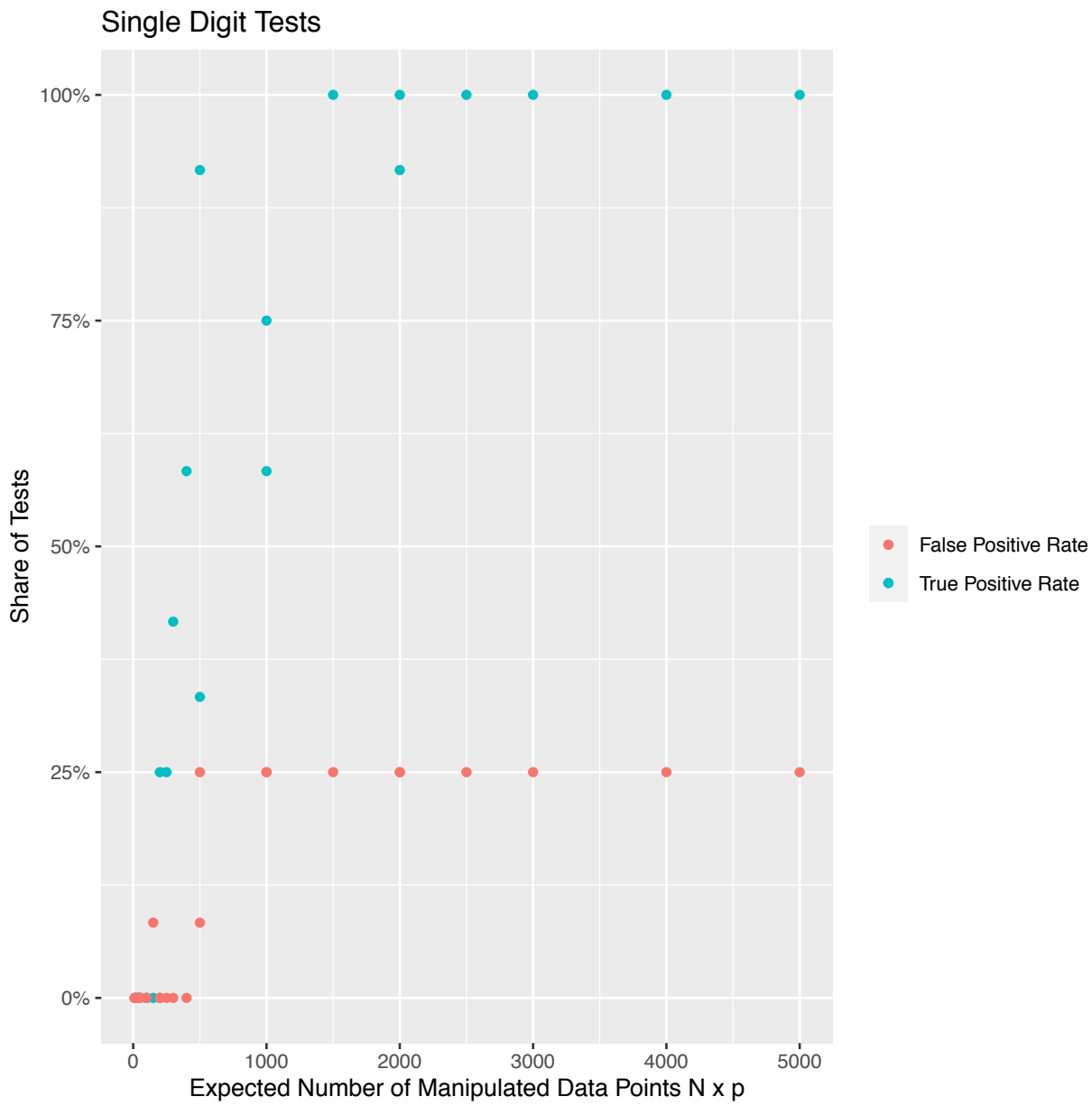
This table presents the results of single-digit-place testing among simulated data, using 4 tests: first digit, second digit, third digit, and last digit. There are 6 districts. Districts A, B, and C manipulate data, and districts D, E, and F do not. Each district produces data with sample size $n$, and districts A, B, and C manipulate data with probability of manipulation $p$. We count the number of tests failed per district as $n$ and $p$ are varied. The true positive rate is the number of tests failed among districts A, B, and C out of 12; the false-positive rate is the number of tests failed among districts D, E, and F out of 12. A Bonferroni correction is used to determine test failure, dividing the desired significance rate of 5% by the number of tests, which is 100 per district.

Appendix Table C2: Results of Many All-Digit-Place Tests

| *n* | *p* | *n* times *p* | A Failed | B Failed | C Failed | D Failed | E Failed | F Failed | True Positive Rate | False Positive Rate |
|---|---|---|---|---|---|---|---|---|---|---|
| 100 | 0.1 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 100 | 0.2 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 100 | 0.3 | 30 | 1 | 0 | 0 | 0 | 0 | 0 | 0.33 | 0 |
| 100 | 0.4 | 40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 100 | 0.5 | 50 | 0 | 1 | 0 | 0 | 0 | 0 | 0.33 | 0 |
| 500 | 0.1 | 50 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 500 | 0.2 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 500 | 0.3 | 150 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 500 | 0.4 | 200 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0.33 |
| 500 | 0.5 | 250 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 1000 | 0.1 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1000 | 0.2 | 200 | 0 | 0 | 1 | 0 | 0 | 0 | 0.33 | 0 |
| 1000 | 0.3 | 300 | 1 | 0 | 1 | 0 | 0 | 0 | 0.67 | 0 |
| 1000 | 0.4 | 400 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 1000 | 0.5 | 500 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 5000 | 0.1 | 500 | 1 | 0 | 1 | 0 | 0 | 0 | 0.67 | 0 |
| 5000 | 0.2 | 1000 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 5000 | 0.3 | 1500 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 5000 | 0.4 | 2000 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 5000 | 0.5 | 2500 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 10000 | 0.1 | 1000 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 10000 | 0.2 | 2000 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0.33 |
| 10000 | 0.3 | 3000 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 10000 | 0.4 | 4000 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 10000 | 0.5 | 5000 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |

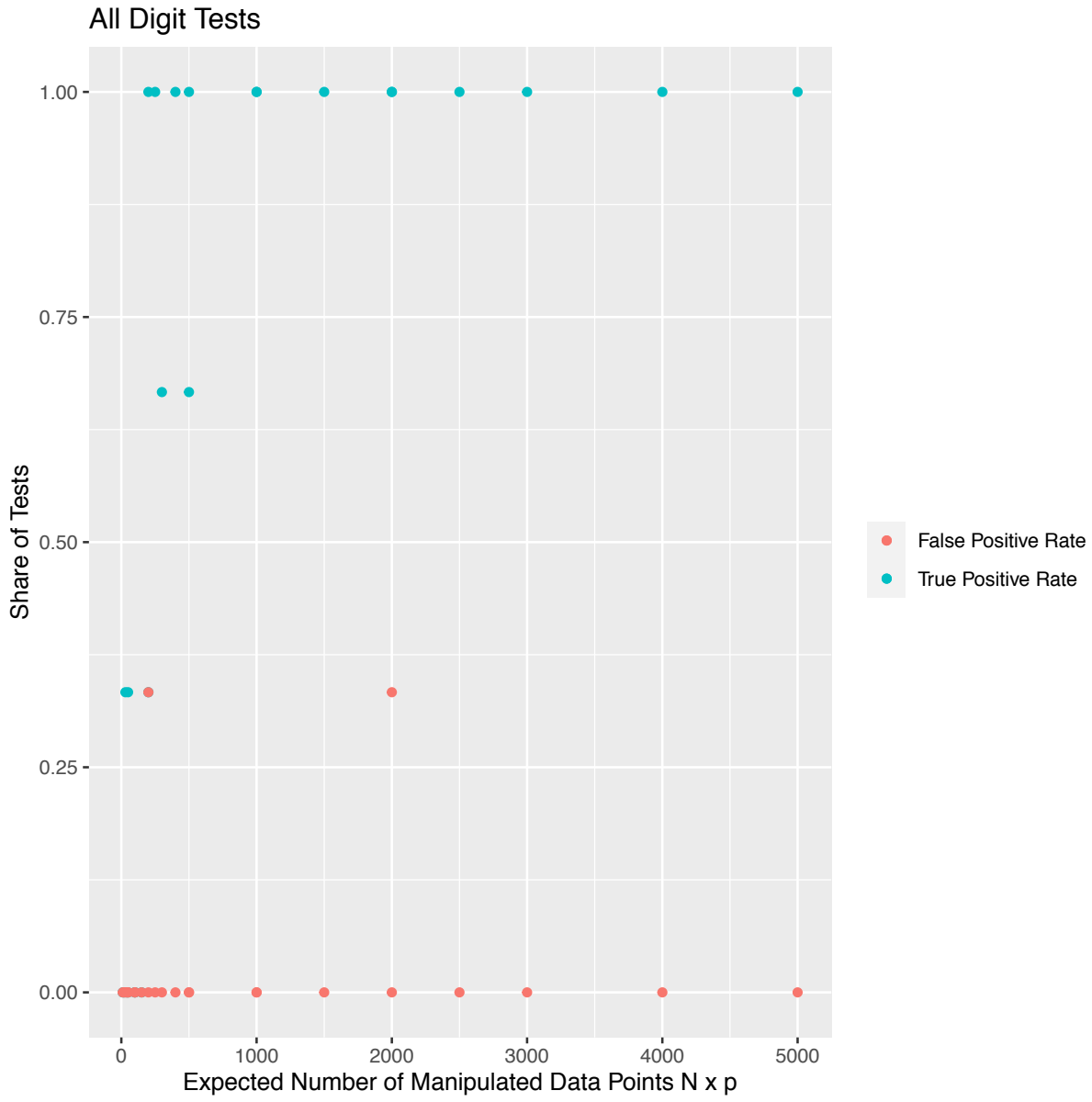
This table presents the results of single-digit-place testing among simulated data. There are 6 districts. Districts A, B, and C manipulate data, and districts D, E, and F do not. Each district produces data with sample size $n$, and districts A, B, and C manipulate data with probability of manipulation $p$. We count which districts either fail the all-digit-places test (1) or do not (0) as $n$ and $p$ are varied. The true-positive rate is the number of tests failed among districts A, B, and C out of 3; the false-positive rate is the number of tests failed among districts D, E, and F out of 3. A Bonferroni correction is used to determine test failure, dividing the desired significance rate of 5% by the number of tests, which is 25 tests per district.

Appendix Figure C3: Variation in True-Positive Rate and False-Positive Rate Among Many Single-Digit Tests




This figure plots the true-positive and false-positive rate of many single-digit tests against the expected number of manipulated data points. Single-digit place testing converges slowly to a perfect true-positive rate.

Appendix Figure C4: Variation in True-Positive Rate and False-Positive Rate Among Many All-Digit Tests




This figure plots the true-positive and false-positive rate of all-digit tests against the expected number of manipulated data points. All-digit-place testing outperforms single-digit-place testing and has a low number of false positives.

# Appendix Bibliography


Beber, B., & Scacco, A. (2012). What the numbers say: A digit-based test for election fraud. *Political Analysis, 20*, 211-234. doi:10.1093/pan/mps003

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy, 76*, 169-217. doi:10.2307/1830482

Berkman, S. (2008). *The World Bank and the Gods of Lending* . Sterling, VA, USA: Kumarian Press.

Boland, P., & Hutchinson, K. (2000). Student selection of random digits. *The Statistician, 49*, 519-529.

Chapanis, A. (1995). Human production of "random" numbers. *Perceptual and Motor Skills, 81*, 1347-1363.

Chavis, L. (2010, November). Decentralizing development. Allocating public goods via competition.". *Journal of Development Economics 93 (2): 264-74., 93*(2), 264-274.

Deckert, J., Myagkov, M., & Ordeshook, P. (2011). Benford's Law and the detection of election fraud. *Political Analysis, 19*, 245-268. doi:10.1093/pan/mpr014

Diekmann, A. (2007). Not the first digit! Using Benford's Law to detect fraudulent scientific data. *Journal of Applied Statistics, 34*, 321-329.

Durtschi, C., Hillison, W., & Pacini, C. (2004). The effective use of Benford's Law to assist in detecting fraud in accounting data. *Journal of Forensic Accounting, V*, 17-34.

Ensminger, J. (2017). *Corruption in Community Driven Development: A Kenyan Case Study with Insights from Indonesia.* U-4 Anti-corruption Resource Centre. Bergen, Norway: Chr. Michelsen Institute.

Guggenheim, S. (2006). Crises and Contradictions: Understanding the Origins of a Community Development Project in Indonesia. In A. Bebbington, M. Woolcock, S. Guggenheim, & E. Olson, *The Search for Empowerment: Social Capital as Idea and Practice at the World Bank* (pp. 111-44). Bloomfield, CT: Kumarian Press.

Hill, T. P. (1995). A statistical derivation of the significant-digit law. *Statistical Science, 10*, 354-363. doi:10.1214/ss/1177009869

Janvresse, E., & De La Rue, T. (2004). From Uniform Distributions to Benford's Law. *Journal of Applied Probability, 41*, 1203-1210.

Mansuri, G., & Rao, V. (2013). *Localizing Development: Does Participation Work?* Washington, DC: World Bank.

Mebane, W. (2011). Comment on "Benford's Law and the Detection of Election Fraud". *Political Analysis*(19), 269-272.

Nigrini, M. (2012). *Benford's Law*. Hoboken, New, Jersey: John Wiley & Sons, Inc.

Nigrini, M., & Mittermaier, L. (1997). The use of Benford's Law as an aid in analytic procedures. *Auditing: A Journal of Practice and Theory, 16*.

Olken, B. A. (2007). Monitoring corruption: Evidence from a field experiment in Indonesia. *Journal of Political Economy, 115*, 200-249. doi:10.1086/517935

Olken, B. A. (2009). Corruption Perceptions vs. Corruption Reality. *Journal of Public Economics, 93*(7-8), 950-64.

Republic of Kenya. (2006). *Arid Lands Resource Management Project (Phase II) Tana River District Progress Report 2003-2006.*

Transparency International. (2009). *Corruption Perceptions Index*. Retrieved August 10, 2020, from https://www.transparency.org/en/cpi/2009#

Wong, S. (2003). *Indonesia Kecamatan Development Program. Building a monitoring and evaluation system for a large-scale community-driven development program.* Washington, DC: World Bank.

Woodhouse, A. (2002). *Village corruption in Indonesia: Fighting corruption in the World Bank's Kecamatan Development Program.".* Washington, D.C.: World Bank.

Woodhouse, A. (2012). *Governance Review of PNPM Rural, Community level analysis.* Jakarta, Indonesia: World Bank Indonesia.

World Bank Integrity Vice Presidency. (2011). *Forensic Audit Report: Arid Lands Resource Management Project -- Phase II -- Redacted Report.* World Bank.

World Bank Integrity Vice Presidency and Internal Audit Department, Treasury, Government of Kenya. (2011). *Redacted Joint Review to Quantify Ineligible Expenditures for the Seven Districts and Headquarters of the Arid Lands Resource Management Program Phase II (ALRMP II) for FY07 & FY08.* Washington, DC: World Bank.