Measuring Strategic Data Manipulation: Evidence from a World Bank Project

By Jean Ensminger and Jetson Leder-Luis

Corresponding Author:
Jean Ensminger
California Institute of Technology
HSS MC 228-77
Pasadena, CA 91125
Fax: 626-405-9841
Telephone: 626-395-4541
jensming@hss.caltech.edu

Jetson Leder-Luis
MIT Department of Economics
The Morris and Sophie Chang Building, E52-300
50 Memorial Drive
Cambridge, MA 02142
jetson@mit.edu

Measuring Strategic Data Manipulation: Evidence from a World Bank Project

By Jean Ensminger and Jetson Leder-Luis*

Abstract

We develop new statistical tests to uncover strategic data manipulation, and we apply these methods to a World Bank project in Kenya. Data produced by humans follow different patterns than naturally occurring data, which motivates an analysis of digit distributions. These new tests unmask profitable data fabrication and suggest efforts to subvert detection. We find evidence consistent with higher levels of fraud in poorly monitored sectors and in a Kenyan election year when graft also had political value. These methods are validated with results from a forensic audit of the same project, which found extensive levels of suspected fraudulent transactions. (*JEL* Codes: D73, O22, H83, C49)*

* Ensminger: California Institute of Technology, HSS MC 228-77, Pasadena, CA 91125, U.S.A. (email: jensming@hss.caltech.edu); Leder-Luis: Massachusetts Institute of Technology, Economics, The Morris and Sophie Change building, E52-300, 50 Memorial Drive, Cambridge, MA 02142, U.S.A. (email: jetson@mit.edu). We particularly thank Avinash Dixit, Esther Duflo, Ben Gillen, Jonathan Katz, Ben Olken, Antonio Rangel, and Robert Sherman for advice and support at critical junctures in this project. 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.

I. Introduction

Corruption is difficult to measure because those who commit illicit acts have an incentive to conceal their behavior. In this paper, we develop new digit analysis methods that exploit irregular digit patterns produced during strategic human data fabrication. These tests derive from two principles: that fraudulent data manipulation responds to economic and political incentives, and that humans are poor generators of random numbers. We apply our new tests to data from a Kenyan World Bank development project. This analysis is specifically tailored to detect patterns consistent with strategic cheating, and these methods also outperform the statistical power of existing digit analysis methods.

Our digit results point to serious levels of data fabrication. The scale of the problem is noteworthy, as is the finding that it intensified during the national election cycle when politics added a further incentive for graft. We are able to statistically validate our method by comparing variation across districts from our digit analysis (showing a failure rate of 30% to 70%, with 60% overall for all districts) to findings for the same districts from the forensic audit (showing 44% to 75% suspected fraudulent or questionable transactions, with an average of 66% across all districts).[1]

Our method involves 10 distinct digit analysis tests that exploit different dimensions of the data. These tests include comparisons to the appropriate theoretical distributions of digits, Benford's Law, but take the novel approach of simultaneously analyzing multiple digit places to provide increased statistical power. We augment this result with a comparison between geographic regions in our dataset, which can be expected to have similar results. Furthermore, we compare distributions by characteristics of the spending, such as expenditure type, year, and value of the manipulated digits. The consistency of these results with the forensic audit suggests that this method can be used to mine data for patterns consistent with fraudulent behavior. Such applications may contribute to real time project monitoring and the targeting of costly forensic audits.

This work relates to a large literature on the monitoring of corruption, including the literature on audits. Ferraz and Finan (2008) show that randomized audits affect the electability of corrupt officials in Brazil; Avis et al. (2016) show that these same audits deter subsequent corrupt behavior. Olken (2007) demonstrates that a rise in Indonesian government audits from 4 percent to 100 percent decreases missing expenditures in road projects by 8.5 percent. Despite their effectiveness, field verified audits are expensive and difficult to implement, as they are labor intensive, use highly trained personnel, and require cooperation from individuals who may be implicated in the corruption, including auditors themselves (Duflo et al., 2013). Similarly, monitoring efforts that track expenditures (e.g. Reinikka and Svensson, 2006) suffer from issues of scale and the need for cooperation from those who may be complicit in the fraud.

A number of creative studies have pioneered methods for identifying and measuring corruption and fraud that require no cooperation from the subjects being monitored. Jacob and Levitt (2003) use aberrant patterns in test scores to uncover cheating in Chicago public schools. Satellite data have been used to track illegal logging (Burgess et al., 2012); stock market fluctuations have been used to measure financial returns to political influence in Indonesia (Fisman et al., 2006, Fisman, 2001). Like these studies, digit analysis does not require the cooperation of potential subjects. It has the added advantages that it scales well and holds potential for application across many domains.

Digit analysis as a method is not unique to this paper, nor are its applications restricted to auditing, though it has had considerable impact there (Amiram et al., 2015, Durtschi et al., 2004, Nigrini, 2012, Nigrini and Mittermaier, 1997). Digit analysis has also been employed to detect and monitor election fraud (Alvarez et al., 2008, Beber and Scacco, 2012), IMF data manipulation (Michalski and Stoltz, 2013), campaign finance fraud (Cho and Gaines, 2012), scientific data fabrication (Diekmann, 2007), and enumerator integrity during survey research (Bredl et al., 2012, Judge and Schechter, 2009, Schräpler, 2011). Our methods build upon this existing work and sharpen the method, providing both new tests and improved statistical power.

The remainder of this paper is organized as follows. Section II describes our dataset and the context of the World Bank project. Section III motivates our digit analysis tests with a discussion of the economics of data manipulation and an overview of the mathematical principles that govern digit distributions. Section IV presents our statistical tests and results. Section V validates our results by comparison to the World Bank forensic audit of the same project, and section VI concludes.

II. Data and Institutional Context

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 28 arid and semi-arid districts, encompassing over 75 percent of Kenya's land area. The project spent $224 million USD, targeting the most impoverished people in the heavily drought-prone regions of Kenya. It funded small infrastructure (schools, dispensaries, and water systems), income generating activities, drought and natural resource initiatives, and training exercises for villagers.

The data used in these analyses are from the original 11 arid districts that received funds from the project; they cover the years 2003 to 2009. These districts share many similar characteristics. Their economies depend primarily upon livestock and are among the poorest in

Kenya—remote from centers of power, poorly educated, and sparsely supplied with infrastructure (roads, schools, health services, access to clean water, and electricity).

The expenditure and participant data used in these analyses were culled from electronic project reports produced in each of the 11 districts. These reports break out the expenditures and numbers of male and female participants associated with each activity undertaken by the project that year. Ensminger's interview data with project staff indicate that usually only 1 or 2 individuals were involved in entering the data for a project component (natural resources and drought management, community driven development, and support for local development). Officers in the districts had considerable latitude over the magnitude of expenditures within budget categories.[2]

The district staff was subject to oversight both from project headquarters in Nairobi and to a lesser extent from the World Bank. Government auditors also routinely audited the project at both the district and the national levels, but Ensminger's interviewees consistently reported that the auditors at both the district level and headquarters were bought off. The World Bank's supervisory missions and financial management oversight were also largely ineffective in their monitoring. Numerous missions rated project financial management "satisfactory" across many years of operations right up to the point of the forensic audit, and the project was labeled "exemplary" in the project renewal proposal (World Bank, 2003: 84). Further, its financial management system was lauded and used as a model for another project, indicating that the World Bank's standard monitoring did not pick up problems (World Bank, 2007).

In 2009 the World Bank's Integrity Vice Presidency (INT) began a forensic audit of the project that lasted 2 years and culminated in a public report (World Bank, 2011). 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 and bank statements. They also travelled to the districts to conduct interviews with suppliers to verify the legitimacy of suspicious transactions. We conduct digit analysis on the reported total expenditures for each of these transactions, such as the total cost of a training exercise, while the forensic auditors investigated the underlying individual receipts for the same transactions.

To understand the decision-making of project staff contemplating embezzlement, it is important to establish their perceptions of the probability of getting caught, and the likely consequences should that happen. Despite the fact that this project was eventually the subject of an intensive World Bank audit, there is reason to believe that staffers would not have considered that to be a likely outcome. To the best of our knowledge, no other field-verified, transaction-based, forensic audit of this scope had taken place at the World Bank before this one, nor has one occurred since (Stefanovic, 2018).[3] More likely, staff engaging in embezzlement feared that their superiors or the Kenyan auditors would expect kickbacks if their activities were exposed.[4] Given this environment, the costs and consequences of being caught embezzling consisted mainly of paying a portion of one's takings, rather than risks of career consequences or prosecution.

We make use of 2 characteristics in the structure of our dataset. First, we have data from 11 arid districts with similar demographics, livelihoods, and ecological conditions, reporting on similar activities, and operating under the same project rules. We proceed from the null hypotheses that digit distributions will be similar across districts. Second, we have data on the number of participants in hundreds of training exercises, which is a count of people who responded to an open invitation for a training exercise. When the same pattern of deviations from theoretical distributions appears in both the expenditure and the participant datasets, it is strongly indicative of human tampering.

III. Theory

We begin by examining the problem of a bureaucrat's decision to accurately report or to fabricate data when tasked with producing expenditure reports. 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, described later, as they accurately reflect the data without human interference. Across different geographic regions of this World Bank project, we would expect similar patterns in the financial data when reporting is conducted honestly.

Bureaucrats have an incentive to falsify expenditure data and embezzle both for personal gain as well as 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 rational crime theory (Becker, 1968). Other costs may include payoffs to auditors or others who detect their fraud, as we discuss in the previous section.

When a bureaucrat determines that the benefits of data manipulation outweigh the costs, we can expect that they will manipulate the data to maximize payout and minimize the probability of detection. This can consist of a number of behaviors. A manipulator may change digits to maximize payout, or may invent new line items to increase the total reported expenditure. In line with a rational decision to commit fraud, we can expect that reporters would increase data tampering in response to greater incentives to steal, and attempt to produce data that appear random to subvert detection. We would furthermore expect that the bureaucrat would expend lower effort in subverting detection for data that are less likely to be monitored.

We analyze each dimension of the data and provide a set of non-overlapping tests that capture different ways in which data can be manipulated. Our tests fall into 3 categories: tests of digit conformance to expected distributions, tests of covariate characteristics of the data (e.g. district, year, and sector), and tests of strategic intent to deceive.

Benford's Law describes the distribution of digits in many naturally occurring circumstances, including 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 three 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 and 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 and Rouen, 2015, Diekmann, 2007, Hill, 1995, Nigrini and 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 and Mittermaier, 1997). Furthermore, Benford's Law arises from data drawn as random samples from random distributions (Hill, 1995). Because numbers 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 and Mittermaier, 1997).

The appropriateness of Benford's Law for analysis of our data set is confirmed by the conformance of the first digits to the Benford distribution, as we show later. The nature of our expenditure data, which are based upon sums of numerous receipts that in turn include sums and multiplication of price times quantity, provides a theoretical basis for why we can expect Benford's Law to be the appropriate distribution. In our analysis, we consistently performed robustness checks by comparing our observed distributions to both the Benford and the uniform distributions. The statistical significance under the uniform distribution is even greater than those reported here. Finally, regardless of Benford's Law, tests of later digit places, particularly last digits, should be uniformly distributed under most conditions.[5]

When experimental subjects are asked to produce random numbers, studies consistently show divergent patterns of human digit preferences. When students were asked to make up strings of 25 digits, their results followed neither the Benford distribution nor the uniform distribution (Boland and 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 overview of African census data, where statisticians discuss a phenomenon known as age heaping, wherein self-reported demographic records 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 et al., 1973, UN Economic and Social Council Economic Comission for Africa, 1986). Throughout our analysis, we omit 0 and 5, which are heavily overrepresented, and analyze digits 1-4 and 6-9; we report rounding levels as measured by 0 and 5 separately.

IV. Digit Tests And Results

A. Tests of Digit Conformance to Expected Distributions: All Digit Places Beyond the First

Our first test is a simultaneous analysis of all digit places beyond the first digit for conformance to Benford's Law. We do not include the first digit because individuals tampering with data may not have complete control over the leading digit, or may avoid changing it to subvert detection. Compared with single digit place tests, which are common in the existing literature, a simultaneous analysis of multiple digit places increases sample size for statistical testing and therefore vastly increases statistical power.[6] 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*.

We use a two-way chi square test to compare the contingency table of all digit places beyond the first against the Benford distribution. As discussed before, we omit 0 and 5 from this analysis, which are handled separately in a discussion of rounding, below. For each digit place (1st digit, 2nd digit, etc), the frequency of each digit (1, 2, 3, 4, 6, 7, 8, 9) is compared with the expected frequencies given in Table 1. This is in contrast to existing studies, which analyze a single digit place with a single chi square test. Because the Benford distribution gives different frequencies by digit place, the two-way chi square test is the appropriate test rather than testing individual digit places. Furthermore, it corrects for multiple hypothesis testing issues that arise from individual digit place analysis.

Figures 1 and 2 present the data of all digit places beyond the first for expenditure (1) and participant data (2). The data are projected onto one axis for visualization. Among the expenditure data for all districts in Figure 1, 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 (Figure 2), 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.

To account for multiple non-overlapping tests, we 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. 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 1 here]

[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.

We do not include a test of the last digit place among our 10 tests because it is technically subsumed under this test, and we wish to avoid non-independence across our tests. Benford's Law predicts a uniform distribution in digit places beyond the fourth; that is, there is no reason that more data should end with a 4 instead of a 3. For comparison to other studies, we include the results of last digit analysis in Appendix A. Both the expenditure and the participant data diverge significantly from the predicted distributions, and both are consistent with our other tests, though we do not include them in the final tally of tests.

B. Tests of Digit Conformance to Expected Distributions: 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):

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

Figure 3A 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 indicate that the data are legitimate, as pooled data may cancel out different individual signatures of manipulation and replicate Benford's Law (Diekmann, 2007). This becomes evident when we look at the data from individual districts where the reports were constructed. Figure 3B 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.

[Figure 3AB here]

C. Tests of Digit Conformance to Expected Distributions: Digit Pairs

Underuse of digit pairs, e.g. 11, 22…99, is a common feature of humanly produced data (Boland and Hutchinson, 2000, Chapanis, 1995). Other applications of digit analysis examine the last two digits (Nigrini, 2012), or explicitly test for digit pairs (Beber and 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 of 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 4. Six of the 11 districts significantly underuse final digits pairs in the participant data at $p < 0.005$ significance, as does the combined sample of all districts ($p = 1.4 \times 10^{-9}$). However, Isiolo District significantly overuses repeated pairs, with $p = 5.6 \times 10^{-5}$ in the binomial test.

[Figure 4 here]

Due to the low value of the Kenyan shilling, rounding at the 1 shilling level may be legitimate among 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 1-shilling 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.

Our next 4 tests exploit the attribute data available in our dataset: districts, years, and sectors (civil works, goods and equipment, training, and transport). These tests do not rely upon Benford's Law.

D. Tests of Digit Covariate Characteristics: Comparisons of District Patterns in Rounding and Repeating

Our next two tests uncover patterns consistent with human tampering, as evidenced by substantial variation across districts without a plausible naturally occurring explanation. 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, 2012, Nigrini and Mittermaier, 1997). In the absence of theoretically acceptable levels of rounding and repeating, we compare districts to each other, as there is no reason to expect differences among them.

The Kenyan shilling was 66 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 places. 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 rounded digits rather than rounded line items, 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 total digits in the district dataset. For example: the number 30,000 has 4 rounded digits; the number 12,350 has 2 rounded digits; 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). We compute the percentage of rounded digits for each district.

Figure 5 shows the percentage of rounded digits by district, with the crosshatched districts in the top quartile of rounding. While we don't know the empirically correct level of rounding that one should observe in the dataset, there is good reason to expect that the same type of retailers, servicing the same type of contracts for similar districts, practiced the same rates of rounding. In the absence of an expected level of rounding, we flag those three districts, roughly the top quartile, that round most heavily, which is more than twice that of the lowest rounding district.

[Figure 5 Here]

Exactly repeated numbers are also a red flag for auditors (Nigrini and 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 three 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 training for 172 trainees should have exceeded all of the amounts listed.

Note that duplicate entries for the same project were removed from the dataset. In our calculations, repeating numbers refer to the use of identical expenditure amounts for completely 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 6 shows the results for the percentage of line items that repeat exactly. As we did with rounding, we indicate the top three districts that most heavily repeat numbers; for example, Baringo approaches 50 percent, while Turkana has about 5 percent. Although the empirically appropriate level of repeating is unknown, we rely on the fact that there is no reason for patterns across districts to differ. Figures 5 and 6 flag different districts in rounding and repeating behavior, indicating that these two tests pick up different signals.

[Figure 6 here]

E. Tests of Digit Covariate Characteristics: Year Effects and the 2007 Kenyan Election

We take advantage of the extra power afforded by use of our new test for analyzing multiple digit places simultaneously to partition our data by project year. This test is designed to detect potential fraud in a presidential election year (2007), which increased incentives to embezzle money for political campaigns. We look for padding of high digit numbers by project year by analyzing the proportion of high to low digits (6, 7, 8, and 9 versus 1, 2, 3, and 4) in all digit places beyond the first. We conduct a chi-square test on the contingency table of high versus low digits. We expect that the probabilities of high and low digits should follow the total probability of those digits from Benford's Law in each digit place. As before, we project this contingency table onto one axis for visualization.

As we see in Figure 7, while all other years slightly underused high digits on average, in 2007 (the only election year) there was a statistically significant overuse of high digits ($p = 6.5 \times 10^{-6}$). This is consistent with a greater incentive to embezzle during a presidential election year to support political campaigns. This phenomenon is supported by Ensminger's interview data and the well-known general pattern of large corruption scandals just prior to national elections in Kenya.

[Figure 7 here]

F. Tests of Digit Covariate Characteristics: Sector Effects

Economic theory (Becker, 1968) and empirical work (e.g. Olken, 2007) indicate that individuals are more likely to cheat when there is a lower risk of detection. Training and transport (travel, fuel, and vehicle maintenance) provide greater opportunities for individuals to conduct fraud when compared to civil works projects or the purchase of goods and equipment, because the latter leave physical evidence of spending, while the former do not. For example, tracking down nomads who were reported as present for a training exercise in a remote village two years prior to an audit is all but impossible. Similarly, fuel can be diverted to private vehicles while leaving no trace. Therefore, we predict that individuals fabricating data for these sectors may do so with less effort expended on deception. We look for evidence of a greater incidence of repeated numbers in these sectors. We plot the percentage of repeated line items that match year, district, sector, and amount, for each of the districts by sector. Figure 8 shows this result.

[Figure 8 here]

We crosshatch those districts that have three times the number of repeats in training and transport as compared to the average number of repeats in civil works and goods and equipment. Six of 11 districts and the all district test fail, but Turkana District provides evidence that there is no structural reason for there to be more repeated data in training and transport.

G. Tests of Strategic Intent: Unpacking Rounded Numbers

Much of what auditors catch in their routine work falls into the category of sloppy bookkeeping. 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. 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. We turn now to the first of two new tests that reveal strategic data manipulation.

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, and second, training exercises in remote villages are notoriously difficult to verify. 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, which they converted to a round number of participants. This total number of participants was then split into males and females, as was required for reporting. Therefore, we expected greater indicators of data fabrication when the total number of participants was a round number (e.g. 300).

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 a chi-square test on the contingency table of digits in digit places beyond the first, versus Benford's Law. 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 9A). 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 9B ($p = 1.9 \times 10^{-6}$), but the statistical significance is considerably higher in the rounded data, Figure 9A ($p = 2.6 \times 10^{-64}$ in the sample of all districts). For 8 out of 11 districts, we reject the null hypothesis that the male and female participant data, when totaling to a round number, are Benford conforming ($p < 0.005$).

[Figure 9AB here]

H. Tests of Strategic Intent: Value of Digit Place with Monte Carlo Simulation

Our final new test reveals patterns consistent with data manipulation that is both profitable and consistent with attempts to conceal such manipulation. 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. Furthermore, we detect signs of behavior consistent with an attempt to make it more difficult to detect the padding by overusing low digits in less valuable digit places.

Benford's Law governs the distribution of digits by position from the left (1^{st} digit, 2^{nd} digit), but not by value, which depends on digit place from the right (e.g. 1s, 10s, 100s place). 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 the 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 datasets, and 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. The Appendix contains technical details of this process.

Figure 10 shows the project data by sector 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 project data in the 10,000s place exceeded 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 10 here]

An interesting finding in Figure 10, which corroborates the strategic placement of digits, 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 among the pooled sector data, represented by the black bars. 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 overuse of high numbers in any given table in the report.

I. Summary: Application of Digit Tests

These tests, taken together, comprise a set of non-overlapping analyses along different dimensions of potential data manipulation. Importantly, some of the tests are not a turnkey system for digit analysis under other circumstances. Some characteristics of this dataset, such as the comparison of expenditure to beneficiary tests, are particular to these data, but are likely to have analogies in many real world situations.

The exact battery 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. What we show is that analysis along all available dimensions of *our* data can be used to uncover suspicious patterns in an efficient and effective way. We expect that our new tests, especially the powerful test using all digit places, and the last test, which takes account of the value of the digit place, should prove useful in many contexts.

By facilitating the full use of our attribute data, this battery of tests helps reveal the magnitude of potential fraud in this project, as well as the important finding that aid funds were very likely being diverted to campaign coffers during an important presidential election year.

V. Comparing Digit Analysis to The World Bank Forensic Audit

Table 2 compiles the results of 10 tests for each district. To correct for type 1 error due to the number of tests we ran, we perform a Bonferroni correction. We divide our desired significance level (0.05) by the number of tests (10), and therefore choose a significance level of 0.005. For the rounding and repeating tests where districts are compared to each other in the absence of a theoretical measure (Figures 5 and 6), the top quartile of districts, 3 of 11, are flagged. For the sector effects (Figure 8), we mark those districts for which training and transport exhibit more than triple the level of repeating compared to the other sectors. These 10 tests avoid overlap and pinpoint different aspects of data tampering. In the bottom row, we sum the number of failed tests by district, which ranges from 3 to 8 out of 10.

[Table 2 here]

The existence of an extensive forensic audit for this project provides us with a measure of external validity for our digit analysis. In Table 3 we compare the results of our digit analyses by district to the results of the World Bank auditors (World Bank, 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 hypotheses for those same 4 districts in 6 to 8 of our 10 digit tests. The remaining district, Tana, had considerably lower levels of suspected fraud than the other districts (44 percent), and we rejected the null on 3 of our 10 digit 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 .939, and a 95% confidence interval of [.338, .996]. We reject the null hypothesis of no correlation at the 5% significance level, with p = 0.018. The World Bank's forensic audit confirms the findings from our digit analysis tests.

[Table 3 here]

We also found significant digit violations in all of the unaudited districts, 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, Baringo, Ijara) have some of the highest number of digit analysis violations (8, 6, and 6) in our sample. This underscores the potential gains of using digit analysis as a diagnostic for targeting costly auditing techniques.

VI. Conclusion

We present new methods to detect data tampering and demonstrate their use on data from a World Bank dataset in Kenya. Our tests reveal patterns consistent with strategic and profitable data fabrication. Notably, the presence of an independent forensic audit of the same project strongly correlates with our digit analysis, lending external validity to the method and the substantive findings.

One of our new tests, employing the generalized Benford's Law to analyze multiple digit places, 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. By partitioning by project years, we are able to 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 new test of overuse of high digits in valuable digit places uncovers patterns consistent with profitable deviations as well as attempts to evade detection. This is the first test we know of that relates aberrant digit patterns to the monetary value of the digit place. This is consistent with intentional behavior, rather than sloppy bookkeeping, and to the best of our knowledge is something heretofore not demonstrated in Benford analyses.

The substantive findings of this project attest to the need for, and importance of, better measures and identification of corruption. The forensic auditors determined that 66% of the district transactions they examined were suspected fraudulent or questionable. On average, the districts we examined failed 60% of the all district digit tests.

Our new tests provide a particularly powerful toolkit for monitoring budget expenditures and uncovering suspected fraud. This method works even when field monitoring is challenging, as is often the case in remote and insecure parts of the developing world. In addition, it requires minimal cooperation from those inside the organization or government, who may have an incentive to impede an investigation. In developing countries, where one faces strong corruption cartels, and weak rule of law with which to force compliance, independence is a major benefit.

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 also 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 still face the same incentive to self-deal, which would undercut efforts to produce aggregate results consistent with Benford's Law. Such coordination would also expose leadership at high risk of detection.

Our methods are complementary to newly developed machine learning methods for fraud detection. Machine learning would not be appropriate on this data set, due to the small number of features (columns of data), the relatively small sample size, and the need for a training set of known outcomes, which these and similar datasets lack. However, digit analysis can be used to further machine learning techniques in other contexts. Fundamentally, machine learning relies on pattern detection. The more dimensions of analysis available, the more powerful machine learning becomes. Digit analysis is another dimension along which machine learning can be trained, and the patterns we have detailed in this study can be useful for even more sophisticated fraud-screening techniques.

By addressing external validity and expanding the capabilities of digit analysis, we hope to facilitate its broader use, especially where sample size is an issue, and data partitioning is desirable. The areas that might benefit, and where digit analysis has already been used, are in auditing, election fraud, scientific data fabrication, and the monitoring of enumerator integrity during survey research. The fact that digit analysis can be deployed without the cooperation of potential offenders is a significant advantage for many monitoring efforts. For example, our method could have been used in real time monitoring of this project to reduce potential fraud, or in the forensic audit of this project to identify and target the worst offending districts, three of which were missed in the World Bank audit sample. Such applications can potentially provide substantial savings. We also foresee use in a variety of new applications, for example, to check the authenticity of data supplied by governments in compliance with international economic, ecological and environmental agreements, or pollution and labor data supplied for treaty compliance. In the modern environment where big data proliferates, stronger tools to analyze these data for strategic and profitable manipulation are necessary.

REFERENCES

Alvarez, R. Michael; Thad E. Hall and Susan D. Hyde eds**.** 2008. *Election Fraud: Detecting and Deterring Electoral Manipulation*. Brookings Institution Press.

Amiram, Dan; Zahn Bozanic and Ethan Rouen. 2015. "Financial Statement Errors: Evidence from the Distributional Properties of Financial Statement Numbers." *Review of Accounting Studies*, 20(4), 1540-93.

Avis, Eric; Claudio Ferraz and Frederico Finan. 2016. "Do Government Audits Reduce Corruption? Estimating the Impacts of Exposing Corrupt Politicians." *National Bureau of Economic Research Working Paper Series*, No. 22443.

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

Becker, Gary S. 1968. "Crime and Punishment: An Economic Approach." *Journal of Political Economy*, 76(2), 169-217.

Boland, Philip and Kevin Hutchinson. 2000. "Student Selection of Random Digits." *The Statistician*, 49(4), 519-29.

Boyle, Jeff. 1994. "An Application of Fourier Series to the Most Significant Digit Problem." *The American Mathematical Monthly*, 101(9), 879-86.

Bredl, Sebastian; Peter Winker and Kerstin Kötschau. 2012. "A Statistical Approach to Detect Interviewer Falsification of Survey Data." *Survey Methodology*, 38(1), 1-10.

Burgess, Robin; Matthew Hansen; Benjamin A Olken; Peter Potapov and Stefanie Sieber. 2012. "The Political Economy of Deforestation in the Tropics." *The Quarterly Journal of Economics*, 127(4), 1707-54.

Chapanis, Alphonse. 1995. "Human Production of "Random" Numbers." *Perceptual and Motor Skills*, 81, 1347-63.

Cho, Wendy K Tam and Brian J Gaines. 2012. "Breaking the (Benford) Law: Statistical Fraud Detection in Campaign Finance." *The American Statistician*, 61(3), 218-23.

Debowski, Lukasz. 2003. "Benford's Law Number Generator," Polish Academy of Sciences, Institute of Computer Sciences,

Deckert, Joseph; Mikhail Myagkov and Peter Ordeshook. 2011. "Benford's Law and the Detection of Election Fraud." *Political Analysis*, 19, 245-68.

Diekmann, Andreas. 2007. "Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data." *Journal of Applied Statistics*, 34(3), 321-29.

Duflo, Esther; Michael Greenstone; Rohini Pande and Nicholas Ryan. 2013. "Truth-Telling by Third-Party Auditors and the Response of Polluting Firms: Experimental Evidence from India." *The Quarterly Journal of Economics*, 128(4), 1499-545.

Durtschi, Cindy; William Hillison and Carl Pacini. 2004. "The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data." *Journal of Forensic Accounting*, V, 17-34.

Ferraz, Claudio and Frederico Finan. 2008. "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes." *The Quarterly Journal of Economics*, 123(2), 703-45.

Fisman, David; Ray Fisman; Julia Galef; Rakesh Khurana and Yongxiang Wang. 2006. "Estimating the Value of Connections to Vice-President Cheney." *The B.E. Journal of Economic Analysis & Policy*, 12(3).

Fisman, Raymond. 2001. "Estimating the Value of Political Connections." *The American Economic Review*, 91(4), 1095-102.

Hill, Theodore P. 1995. "A Statistical Derivation of the Significant-Digit Law." *Statistical Science*, 10(4), 354-63.

Integrity Vice Presidency of the World Bank 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 2 (Alrmp 2) for Fy07 & Fy08," Washington, DC:

Jacob, Brian A. and Steven D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *The Quarterly Journal of Economics*, 118(3), 843-77.

Judge, George and Laura Schechter. 2009. "Detecting Problems in Survey Data Using Benford's Law." *Journal of Human Resources*, 44(1), 1-24.

Michalski, Tomasz and Gilles Stoltz. 2013. "Do Countries Falsify Economic Data Strategically? Some Evidence That They Might." *The Review of Economics and Statistics*, 95(2), 591-616.

Nagi, M.H; E.G. Stockwell and L.M. Snavley. 1973. "Digit Preference and Avoidance in the Age Statistics of Some Recent African Censuses: Some Patterns and Correlates." *International Statistical Review*, 41(2), 165-74.

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

Nigrini, M and L Mittermaier. 1997. "The Use of Benford's Law as an Aid in Analytic Procedures." *Auditing: A Journal of Practice and Theory*, 16(2).

Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy*, 115(2), 200-49.

Rath, Gustave J. 1966. "Randomization by Humans." *The American Journal of Psychology*, 79(1), 97-103.

Reinikka, Ritva and Jakob Svensson. 2006. "Using Micro-Surveys to Measure and Explain Corruption." *World Development*, 34(2), 359-70.

Republic of Kenya. 2006. " Arid Lands Resource Management Project (Phase Ii) Tana River
District Progress Report 2003-2006," 1-61.

Schräpler, Jörg-Peter. 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(5/6), 685-718.

Stefanovic, Michael. Former head of Investigations at INT (World Bank Integrity Vice Presidency). 2018. Interview with Jean Ensminger, Email.

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:

Walter R. Mebane, Jr. 2011. "Comment on "Benford's Law and the Detection of Election Fraud"." *Political Analysis*, 19, 269-72.

World Bank. 2011. "Forensic Audit Report: Arid Lands Resource Management Project - Phase Ii," Redacted: World Bank,

____. 2003. "Project Appraisal Document on a Proposed Credit in the Amount of Sdr 43.6 Million (Us$ 60m. Equivalent) to the Republic of Kenya for the Arid Lands Resource Management Project Phase Two," C. D. Eastern and Southern African Rural Development Operations, Africa Region, 31.

____. 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 Integrity Vice Presidency. 2018. "Redacted Investigation Reports," http://www.worldbank.org/en/about/unit/integrity-vice-presidency/redacted-investigation-reports. Accessed: 5/25/2018

Tables and Figures

TABLE 1: EXPECTED DIGIT FREQUENCIES UNDER BENFORD'S LAW

		Digit Place				
		1	2	3	4	5
	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
Digit	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

Source is Nigrini and Mittermaier (1997:54).

TABLE 2. SIGNIFICANCE OF DIGIT TESTS BY DISTRICT

Fig	Digit Test	Mandera	Isiolo	Baringo	Ijara	Wajir	Garissa	Samburu	Marsabit	Moyale	Turkana	Tana	All Districts
1	All Digit Places beyond the First: Expenditure	**3.6E-14** **846**	0.0082 437	**7.2E-17** **1352**	**2.6E-05** **769**	**1.9E-06** **1248**	**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**
2	All Digit Places beyond the First: Participant	**9.0E-18** **886**	**6.1E-11** **478**	**2.1E-04** **674**	**5.9E-11** **765**	**6.5E-15** **731**	**6.2E-18** **858**	**2.3E-05** **639**	0.25 527	0.033 736	**0.0037** **591**	0.013 500	**5.7E-51** **7385**
3	First Digit: Expenditure Data	**1.4E-08** **489**	**5.5E-06** **308**	**1.4E-09** **488**	**2.3E-13** **386**	0.37 578	0.029 430	**5.7E-05** **359**	0.011 293	**1.9E-12** **319**	0.071 357	**0.0037** **332**	0.089 4339
4	Digit Pairs: Participant	0.0091 238	**0.0044** **125**	**8.2E-04** **251**	**0.0037** **176**	**7.1E-05** **255**	**1.8E-04** **293**	0.38 166	**0.0031** **126**	0.41 173	0.49 119	0.035 137	**1.4E-09** **2059**
5	Rounding Digits: Expenditure	**Top Quartile**	**Top Quartile**						**Top Quartile**				DNA
6	Repeating Numbers: Expenditure	**Top Quartile**	**Top Quartile**	**Top Quartile**									DNA
7	Year Effects (2007): Expenditure	0.18 117	0.22 98	0.0073 182	0.045 222	0.088 273	0.016 139	0.15 231	0.032 88	0.033 238	**0.0045** **192**	0.75 165	**6.5E-06** **1945**
8	Sector Effects: Expenditure	**> 3x**	**> 3x**	**> 3x**	**> 3x**	**>3x**		**> 3x**					DNA
9	Unpacking Rounded Numbers: Participant	**6.1E-21** **453**	**4.4E-13** **157**	0.0085 248	**1.2E-10** **298**	**7.6E-11** **433**	**5.9E-24** **459**	**3.9E-05** **179**	0.014 222	**0.0030** **179**	**3.1E-05** **205**	0.057 142	**2.6E-64** **2975**
10	Deviation from Mean in 10,000s Digit Place	**1.0E-05**	0.131	0.024	0.0054	**1.0E-05**	**0.0015**	**1.0E-05**	**1.0E-05**	**1.0E-05**	**1.0E-05**	**1.0E-05**	**1.0E-05**
	Number of Significant Tests p < 0 .005 (Out of 10)	8	7	6	6	6	5	5	4	4	4	3	6

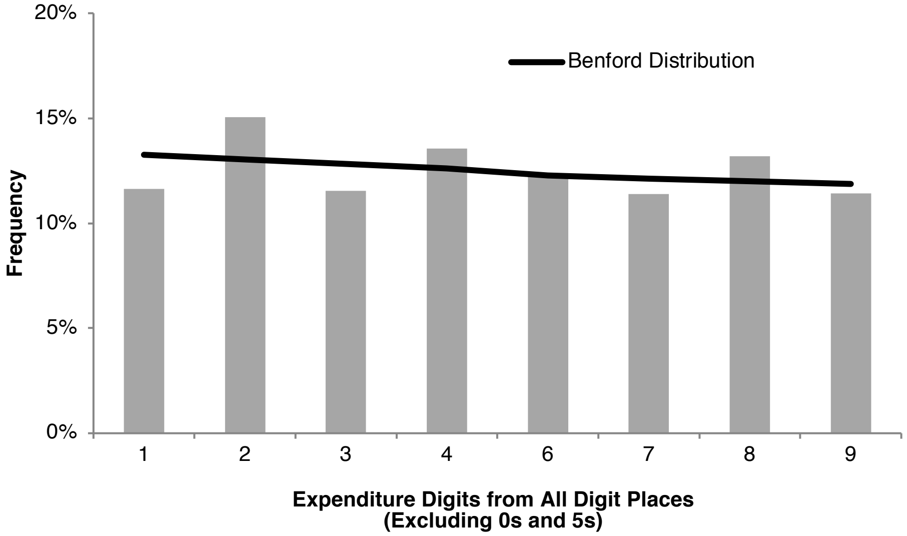
We ran 10 digit tests on each of 11 districts. The tests were chosen to analyze different, non-overlapping aspects of the data. Given the large number of tests, a Bonferroni correction was used to establish 0.005 as the acceptable *p* – value for our tests. Failed tests at the 0.005 level are indicated in bold. For rounding and repeating (Figures 5 and 6), there is no theoretical means to establish the expected level and we work from the null hypothesis that there should be no significant difference between the districts. We flag the districts that are outliers in the upper quartile. Similarly, for the sector analysis (Figure 8), we flag the districts for which repeated numbers in sectors with higher risk of fraud (training and transport) are more than triple the level of other sectors (civil works and goods and equipment). We tabulate the number of significant tests for each district in the bottom row.

TABLE 3. 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)
Isiolo	7	74
Wajir	6	75
Samburu	5	68
Garissa	5	62
Tana	3	44
Mandera	8	Not Audited
Baringo	6	Not Audited
Ijara	6	Not Audited
Moyale	4	Not Audited
Marsabit	4	Not Audited
Turkana	4	Not Audited

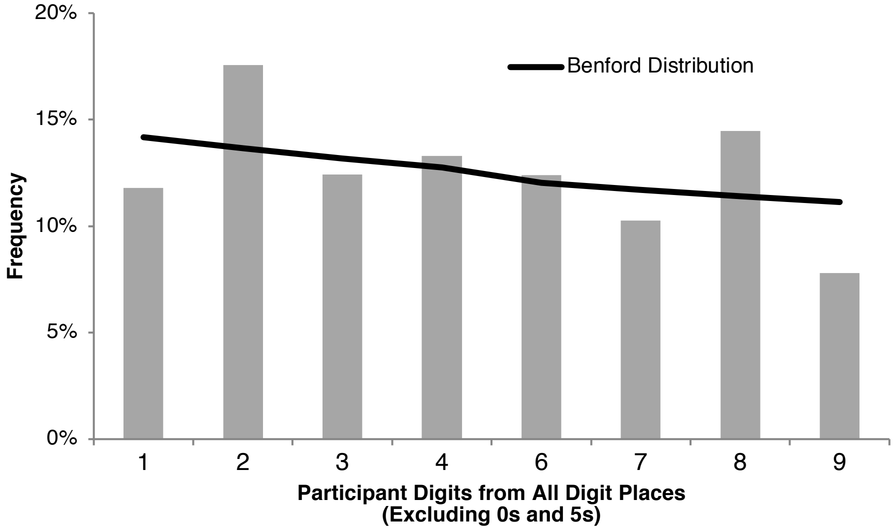
Source for the INT forensic audit data is World Bank (2011).

FIGURE 1: ALL DIGIT PLACES BEYOND THE FIRST AGAINST BENFORD'S LAW FOR EXPENDITURE DATA



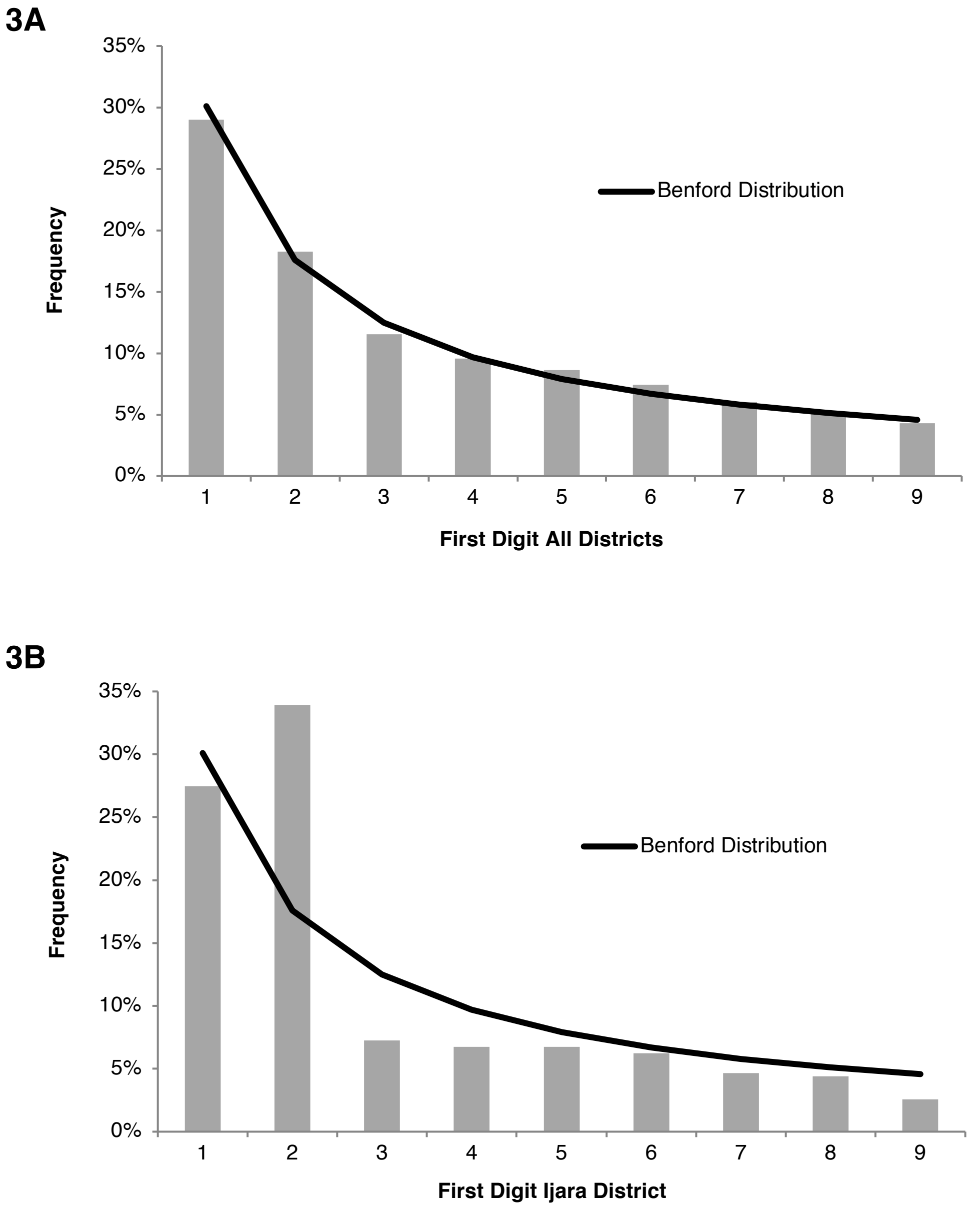
All districts combined (p = 3.9 × 10^{-15}; n = 9371).

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



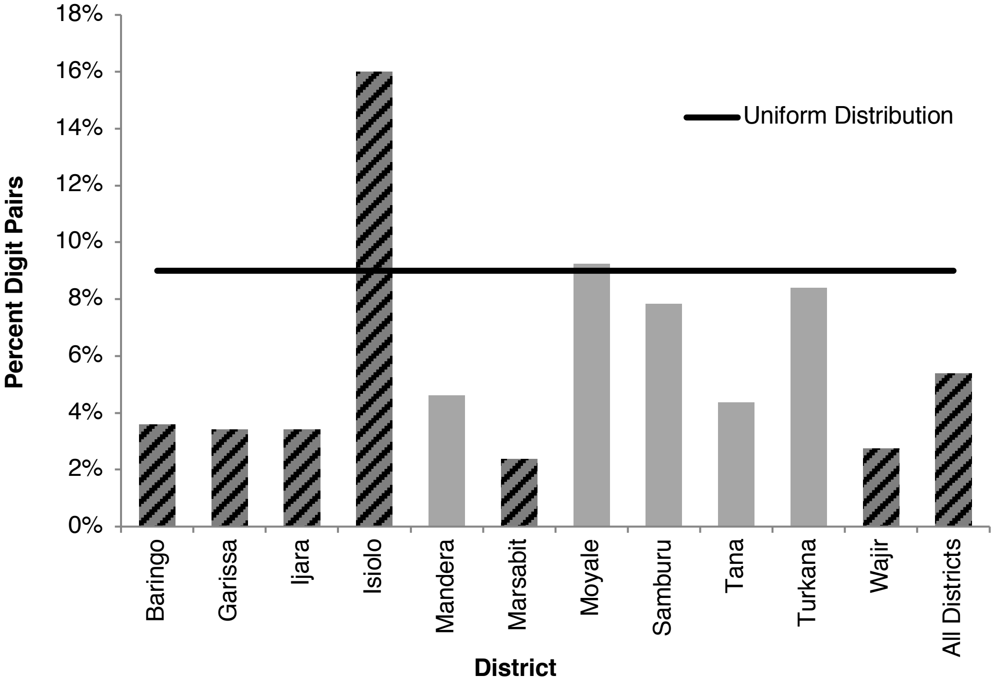
All districts combined (p = 5.7 × 10^{-51}; n = 7385).

FIGURE 3AB: FIRST DIGIT EXPENDITURE DATA AGAINST BENFORD'S LAW



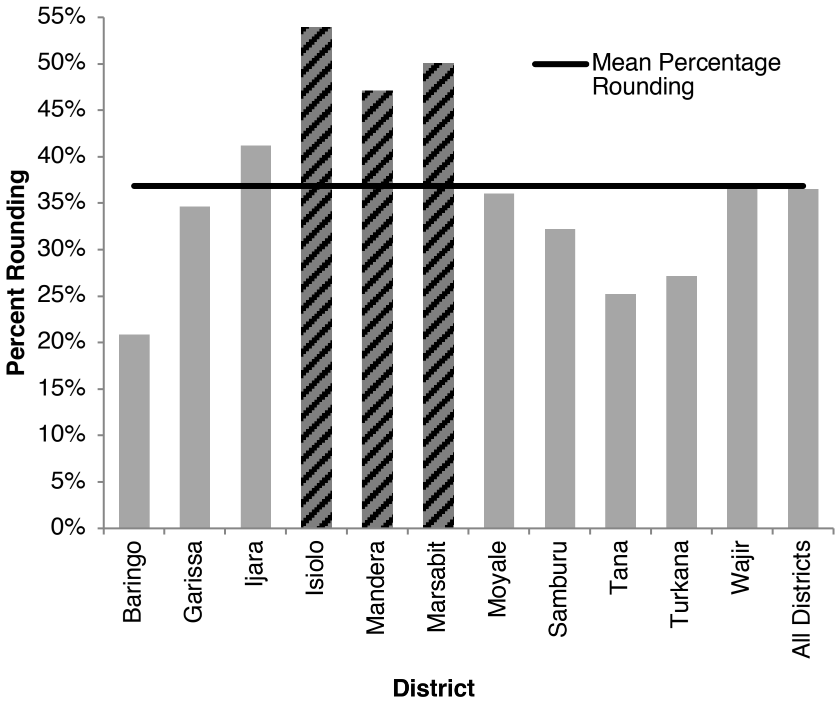
(A) All districts combined ($p = 0.089$; $n = 4339$). (B) Ijara District only ($p = 2.3 \times 10^{-13}$; $n = 386$).

FIGURE 4: DIGITS PAIRS IN THE LAST TWO DIGITS FOR PARTICIPANT DATA BY DISTRICT



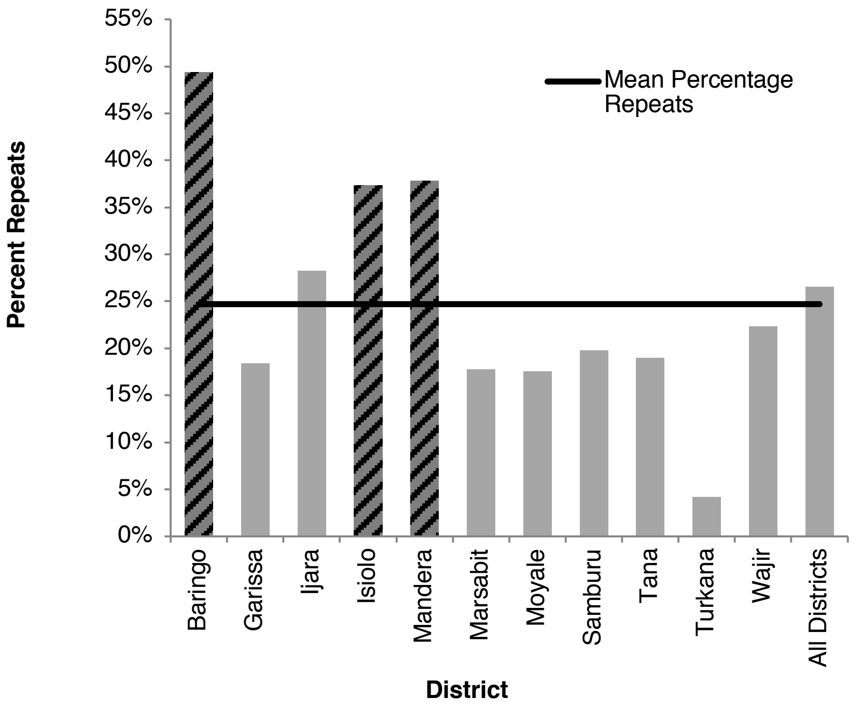
Crosshatched districts fail the binomial test at $p < 0.005$ by over or underutilizing digit pairs such as 11, 22, and 33 (Baringo p=8.2 × 10^{-4}, n = 251; Garissa p = 1.8 × 10^{-4}, n = 293; Ijara p = 0.0037, n = 176; Isiolo p = 0.0044, n = 125; Marsabit p = 0.0031, n = 126; Wajir p = 7.1 × 10^{-5}, n = 255; all districts p = 1.4 × 10^{-9}, n = 2059).

FIGURE 5: PERCENTAGE OF ROUNDED DIGITS IN EXPENDITURE DATA BY DISTRICT



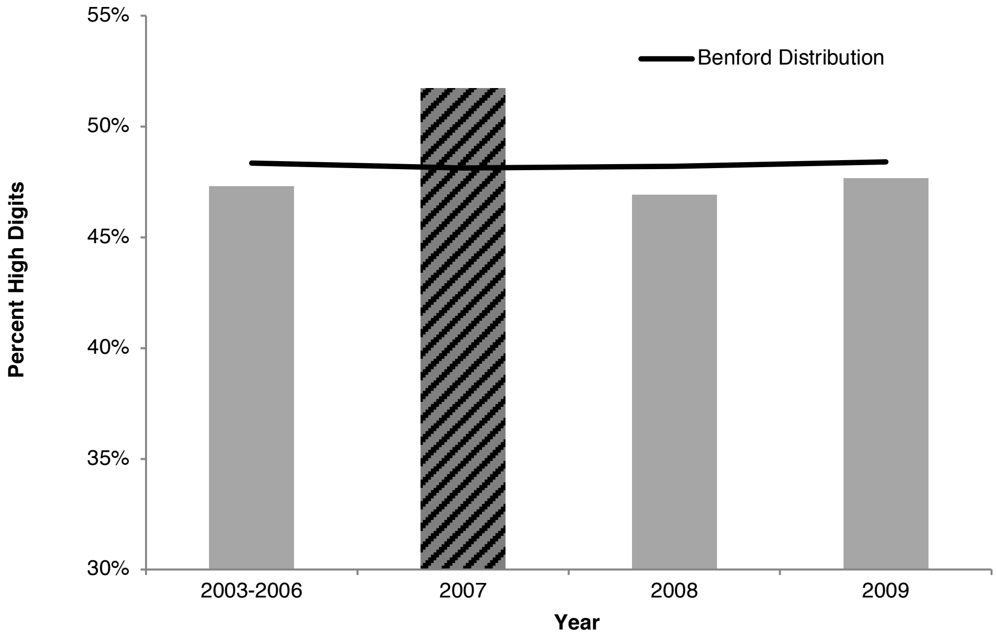
Percentage of digit places rounded in expenditure data by district. The three districts representing the top quartile are crosshatched.

FIGURE 6: PERCENTAGE OF REPEATED ENTRIES IN EXPENDITURE DATA BY DISTRICT



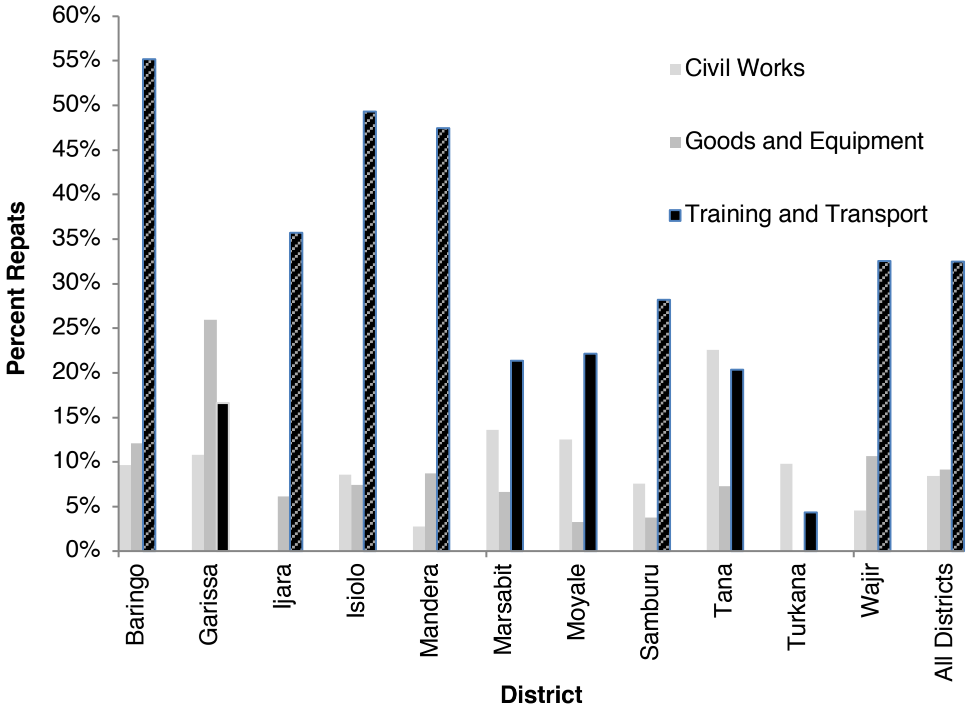
Percentage of exactly repeated expenditure entries by district for a given annual report. The three districts representing the top quartile are crosshatched.

FIGURE 7: ELECTION YEAR EFFECTS IN EXPENDITURE DATA



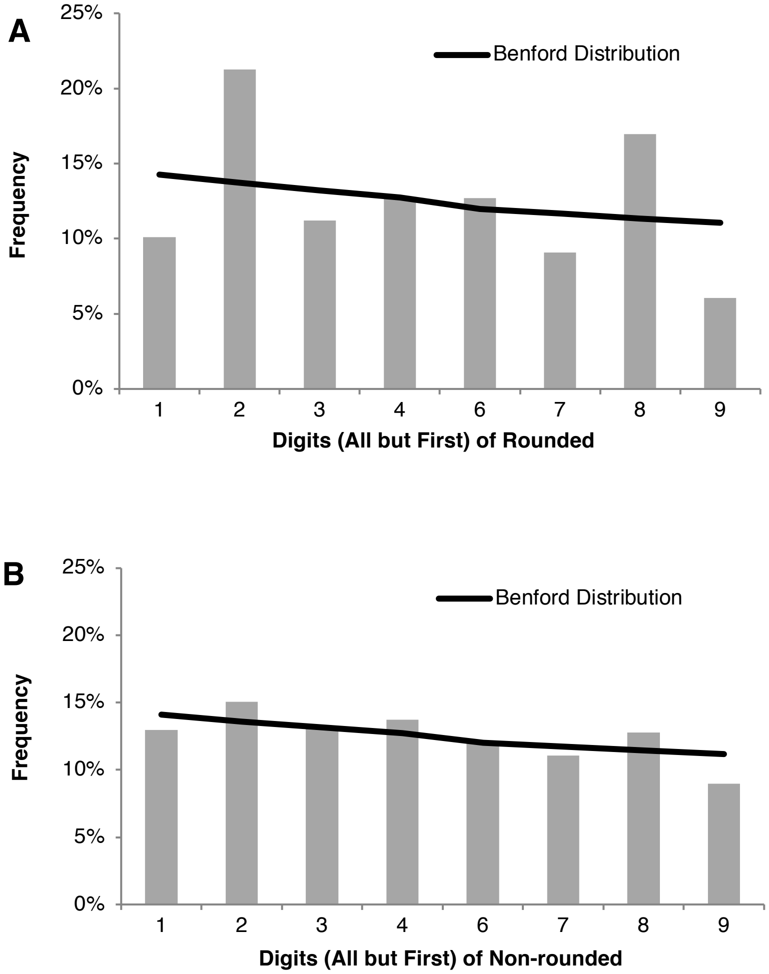
Percentage of high digits (6, 7, 8, 9 versus 1, 2, 3, 4) in all digit places but the first, for all districts, by year. 2007 was a Presidential election year. 2007 has a statistically significant presence of high digits in a ($p = 6.5 \times 10^{-6}$; $n = 1945$.)

FIGURE 8: SECTOR EFFECTS IN EXPENDITURES



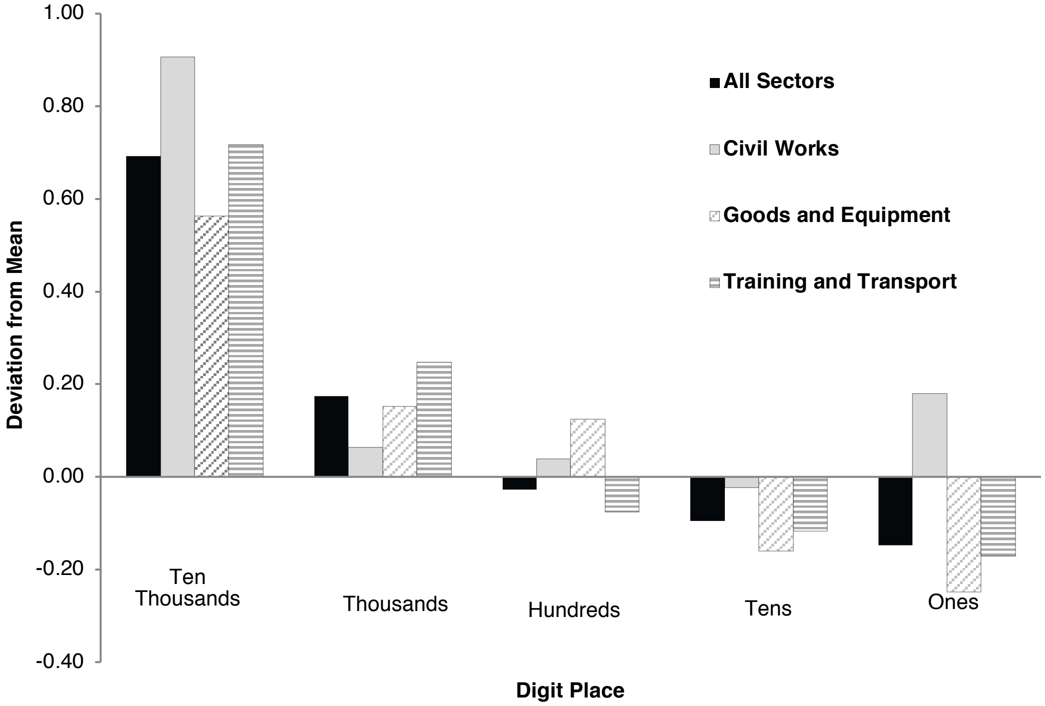
Percentage of line item expenditures repeated exactly, by district, year, and sector. Crosshatched districts report over three times as many repeated numbers in training and transport, versus other sectors, consistent with low-effort data manipulation.

FIGURE 9AB: UNPACKING ROUNDED AND UNROUNDED DIGITS IN PARTICIPANT DATA



(A) Digit breakout of all but the first digit (excluding 0s and 5s) when the total of male and female participants sums to a rounded number ($p = 2.6 \times 10^{-64}$; $n = 2975$). (B) Digit breakout of all but the first digit (excluding 0s and 5s) when the total of male and female participants sums to a non-rounded number ($p = 1.9 \times 10^{-6}$; $n = 4410$).

FIGURE 10: DEVIATION FROM BENFORD'S LAW MEAN IN EXPENDITURE DATA WITH MONTE CARLO SIMULATION



We compare the observed mean by digit place from the right to the Benford expected mean in each sector. Zero reflects conformance to the Benford expected mean. Positive values indicate the mean is higher than Benford's Law predicts. The observed pattern is consistent with a strategy of 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: 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 ($p = 0.061$).

Appendix

A. Last Digits

The 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 (Beber and Scacco, 2012, Nigrini and Mittermaier, 1997). Results on the terminal digit are presented in Appendix Figure 1AB and show exceptional statistical significance for both expenditure and participant data. However, Benford's Law limits to the uniform distribution for digit places beyond the fourth, and therefore, tests of the final digit place are subsumed by the test of conformance to Benford's Law in our test of all digit places beyond the first.

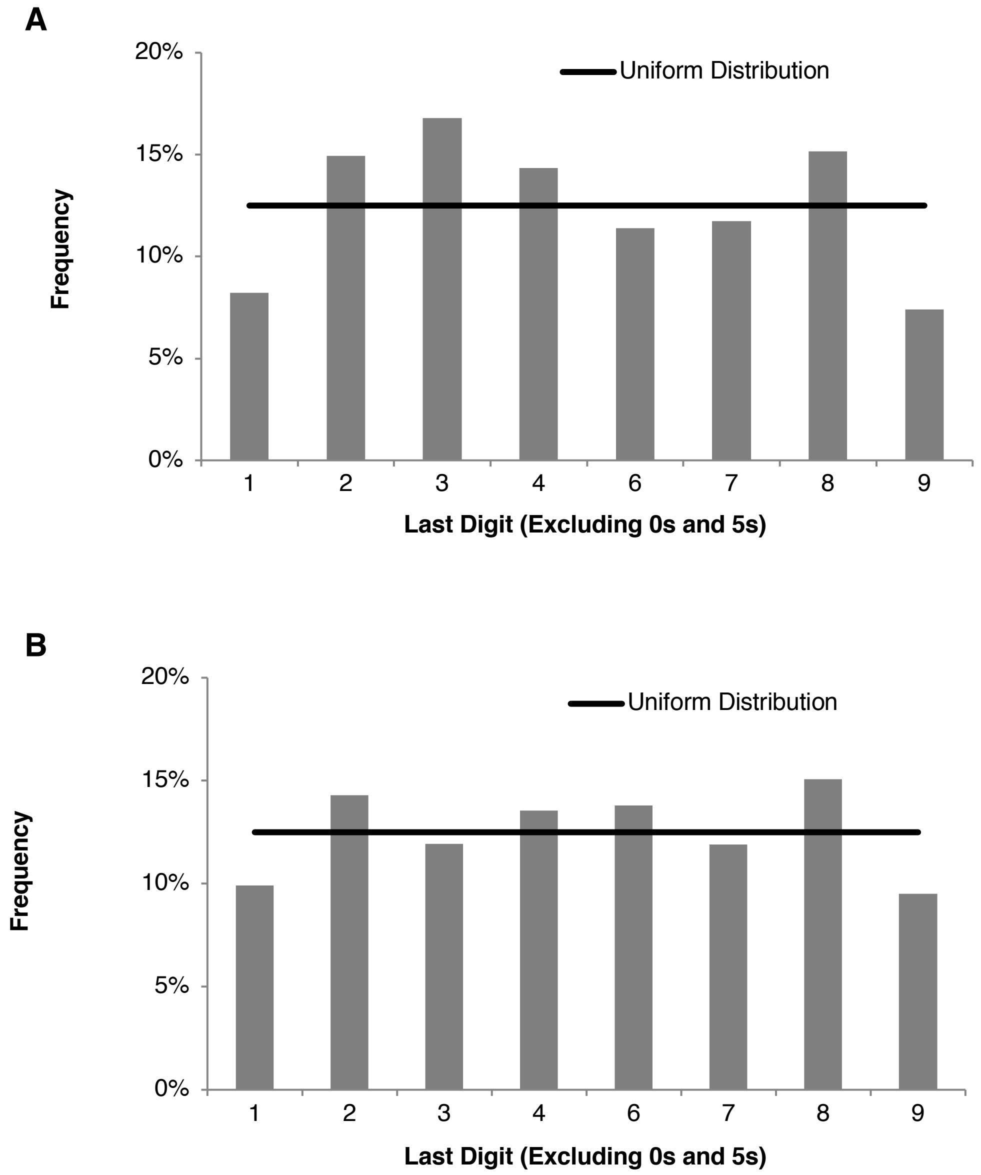
B. Difference of Means Statistics with Monte Carlo Simulations

We compute the mean by digit place in the last five digits from the right among 5, 6, and 7 digit numbers. We eliminate 0s and 5s from this computation and reweight 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 of 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, indicating padding. Negative values indicate a weighted mean that is below the weighted Benford's Law, indicating overuse of low digits.

In order to determine significance of each of our statistics, we perform a Monte Carlo simulation. We generate 100,000 datasets that are identical to the digit lengths observed in our dataset. Code for simulating Benford-distributed numbers was used with permission (Debowski, 2003). Code for matching Benford-conforming numbers with the lengths of our data was produced in Python. For each simulated dataset, 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 datasets, 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. Because there are 100,000 samples, there is a minimum *p*-value of 1 in 100,000.

APPENDIX FIGURE 1AB: LAST DIGIT EXPENDITURE AND PARTICIPANT DATA AGAINST THE UNIFORM DISTRIBUTION.



(A) Expenditure data ($p = 1.5 \times 10^{-9}$; $n = 851$). (B) Participant data ($p = 7.0 \times 10^{-26}$; $n = 5850$).

Footnotes

[1] The World Bank flagged 66% of the district transactions as suspicious; of these, 49% were classified as suspected fraudulent and 17% as questionable.

[2] We exclude community driven development projects from our analysis. These expenditures were grants that were subject to caps, and as such are not appropriate for digit analysis. However, we do use data (expenditures and numbers of participants) from the training exercises and transport costs that supported these activities, as they were not subject to fixed caps.

[3] For example, this is the only such audit on the World Bank INT website (World Bank Integrity Vice Presidency, 2018).

[4] 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 $4 million USD for disallowed charges (Integrity Vice Presidency of the World Bank and Internal Audit Department, 2011). It is noteworthy, therefore, that no one from this project was taken to court, and this speaks to the probability of consequences in the current Kenya context.

[5] 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 (Beber and Scacco, 2012, Deckert et al., 2011, Walter R. Mebane, 2011). 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 and Pacini, 2004).

[6] Individual digit place analyses beyond the first include second and last digit analysis. (Beber and Scacco, 2012, Diekmann, 2007, Nigrini and Mittermaier, 1997)