The Fingerprints of Fraud:

An In-depth Study of Election Forensics with Digit Tests

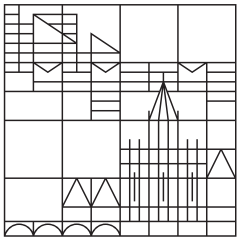
Dissertation zur Erlangung des akademischen Grades
eines Doktors der Sozialwissenschaften (Dr. rer. soc.)

vorgelegt von

Verena Mack

an der

Universität Konstanz



Sektion Politik - Recht - Wirtschaft
Fachbereich Politik & Verwaltungswissenschaft

Konstanz, 2016

Konstanzer Online-Publikations-System (KOPS)
URL: http://nbn-resolving.de/urn:nbn:de:bsz:352-2-163apgv2dvc5m5

Tag der mündlichen Prüfung: 20. Juli 2016

1. Referent: Prof. Dr. Susumu Shikano, Universität Konstanz
2. Referent: Prof. Dr. Nils Weidmann, Universität Konstanz
3. Referent: Prof. Dr. Thomas Hinz, Universität Konstanz

Acknowledgments

I could have not completed this dissertation without the help and support of many people. I'm grateful for all the ideas, feedback, motivations and inspirations I received during this time.

A great deal of gratitude goes to my supervisor Susumu Shikano, who encouraged me to do this dissertation and supported me all the way through it. Susumu Shikano is also the coauthor of the first dissertation paper, where he contributed his knowledge, ideas and experiences that greatly improved the paper and lead it to a successful outcome. However, his support and contribution go way beyond the coauthored paper, as all of them received valuable feedback and his critical perspective sharpened each project.

The generous funding of the Graduate School of Decision Sciences (GSDS) which I received for more than three years enabled me to conduct this project. The support of the GSDS is not limited to the scholarship, as they also generously funded my laboratory experiment for the third dissertation paper, provided a constructive and interdisciplinary environment to work in and supported me financially to participate on national and international conferences. Here, thanks go to participants of the panel at the annual meeting of the Midwest Political Science Association 2013 and 2014 for helpful comments to earlier versions of the first and second paper and to participants of the panel at the European Political Science Association 2015 that gave helpful feedback on an earlier version of the third paper.

I am indebted to my colloquies at the University of Konstanz and especially those of the chair of political methodology. I always enjoyed our gatherings and lunch talks and I am especially thankful for the great input I received in our colloquiums. Ideas, comments and thoughts that are traceable in the dissertation come also from Niklas Harder, Michael Herrmann, Peter Meißner, Konstantin Käppner and Florian Kern.

My third dissertation paper, the laboratory experiment, evolved a lot through the contribution of others. The input of Jan Hausfeld, Konstantin von Hesler, Fabian Dvorak and Dominik Bauer as well as all participants of the informal colloquium of Area A of the GSDS greatly improved the experimental design, so that it meets the high standards required by economic laboratory experiments. Here, I also want to thank Michael Stoffel, who gave valuable comments to the design and programming of the experiment.

I thank Karin Becker for all the good times at the chair and her first-class administrative support. I am also grateful to Jutta Obenland for her inspirational attitude and handling all administrative and non-scientific difficulties that emerged during the establishment of the GSDS.

I am blessed with a great group of friends and family. My parents, who did not just do a great job in parenting, but also did last minute spelling checks of for example the experimental introductions. I received a lot of inspiration and motivation and

support from people inside and outside the University. Thanks to Helen, Helena, Steffi, Simon, Kristin, Peter, Daniela, Espen, Sabine, Vera, Martin and many more. In memory to Hellen and Janosch, who I very much miss.

Finally, I want to thank Marco, who motivated and supported me during this time. Thanks for all the happy hours we spent together, exploring remote spots with ski, bike, a rope or by foot, the sad moments we shared, and for all the encouragement to reach out and see what is possible.

Zusammenfassung

Die Wahlforensik ist ein Analyseverfahren, bei dem mittels statistischer Methoden Wahlfälschung aufgespürt wird, um die Legitimität der Wahl zu beurteilen. In dieser Dissertation liegt der Fokus auf der Forensik der Ziffer-Tests, die in den letzten Jahren einen regelrechten Boom erlebten und deren Bedeutung und Einfluss weiterhin zunehmen. Bei den Ziffer-Tests wird die theoretische Verteilung einer bestimmten Ziffer mit der empirischen Häufigkeit dieser Ziffern in den gezählten Stimmen der Wahlergebnisse verglichen, wobei überzufällige Abweichung Anomalie oder Wählfälschung anzeigen. Trotz der allgemeinen Beliebtheit dieses Analyseverfahrens bleibt unklar, wie diese genau funktionieren, welche Fälschungsarten sie (nicht) aufdecken können und wie vertrauenswürdig die Fälschungsindikatoren sind. Da sowohl Wahlfälschung als auch fälschliche Anschuldigungen ein sensibles Thema sind, die unseren demokratischen Prozess beeinflussen können, ist die Genauigkeit solcher Aussagen von hoher Relevanz. Die Dissertation trägt dazu bei Wahlfälschung genauer zu identifizieren, in dem sie viele der Rätsel, welche die Ziffer-Tests umgeben, anpackt und löst.

In dem ersten Artikel untersuchen wir die Präsidentschaftswahlen in Frankreich, Finnland und Russland, die sich mit Hinblick auf die wahrgenommene Legitimität sehr stark unterscheiden. Wir evaluieren und validieren in diesem Rahmen den Test nach Benfords Gesetz und den Test der letzten Ziffer sowie die Verbindung, die zwischen beiden besteht. Dies geschieht mittels der Fälschungsmechanismen, die mit dem jeweiligen Test aufdeckbar sein sollten sowie mit einem Kreuzvalidierungsverfahren durch Indikatoren extremer Fälschung in Russland. Auf Basis der Ergebnisse kommen wir zum Schluss, dass signifikante Testergebnisse nicht zwingend Wahlfälschung signalisieren.

Im zweiten Artikel unterscheide ich Fälschungsmechanismen hinsichtlich der Koordination sowie die wichtigsten Fälschungshandlungen. Diese Mechanismen werden mit den Kanadischen Wahlergebnissen von 2008 kalibriert, für die es keine Anschuldigung der Wahlmanipulation gibt, um anschließend mittels Simulation künstlich manipulierte Wahlergebnisse zu generieren. Hierdurch lässt sich die Sensitivität des Tests nach Benfords Gesetz für unterschiedliche Fälschungsmechanismen schätzen sowie durch unverfälschte Wahlergebnisse dessen Spezifität. Den Ergebnissen nach zu urteilen können nur sehr wenige Fälschungsarten und dies nur unter bestimmten Rahmenbedingungen von dem Test aufgedeckt werden.

Im dritten Artikel verwende ich ein Laborexperiment, um zu untersuchen, wie Menschen Wahlen von Hand manipulieren. Die Analyse der Daten zeigt „Fingerabdrücke“ der Fälscher, die ansonsten nicht sichtbar wären und verbessert somit unser Verständnis darüber wie Wahlmanipulationen durchgeführt werden. Des Weiteren wird in dem Artikel untersucht, ob die Fälschungsstrategien durch die Rahmenbedingungen der Manipulation beeinflusst werden und ob Rahmenbedingung und Strategien die Aufdeckbarkeit mit dem letzten Ziffer-Test beeinflussen. Die Analyse identifiziert eine Reihe von Strategien, die in Kombination mit dem Fälschungsausmaß die Aufdeckwahrscheinlichkeit beeinflussen.

Abstract

Election forensics is an approach to assess the legitimacy of an election using statistical tools to detect election fraud. The dissertation focuses on the forensics of digit tests, for which applications and the impact is constantly growing. Digit tests compare a theoretical distribution of a specific digit with the empirical frequencies of that digit in vote counts of election results. A significant deviation indicates anomalies or fraud. Despite their popularity it remains uncertain how digit tests work, what kind of election fraud they can and cannot detect and if they are trustworthy as fraud indicators. The precision in detection and the accurateness of accusations is important, as they are sensitive issues that can affect the democratic process. The dissertation contributes to the accurateness of fraud detection through solving most of the riddle surrounding digit tests.

In the first paper we investigate presidential elections in France, Finland and Russia, which differ strongly in their perceived legitimacy, and evaluate the validity of the Benford's Law test, the last digit test as well as their linkage. In particular, we consider specific fraud mechanisms that each digit test can (and cannot) capture and cross validate the results in Russia with extreme fraud indicator. We conclude that significant digit tests do not necessarily indicate manipulation.

In the second paper I distinguish between different levels of fraud coordination and actions that capture relevant fraudulent activities. The mechanisms are calibrated with the 2008 Canadian election results, that are considered to be fraud-free and then used to simulate artificially manipulated election data. This gives a unique setting to estimate the sensitivity and specificity of the Benford's Law test. The findings narrow the fraudulent activities that are detectable by Benford's Law to very specific manipulation settings.

In the third paper I use a laboratory experiment to investigate how humans manipulate election returns by hand. The analysis of the experimental data can capture "fingerprints" of manipulation that are often not visible to the eye and improves our knowledge about fraudulent activities. Additionally, I investigate if the applied strategies are affected by the manipulation setting and if both, setting and strategies, affect last digit tests' detectability. I find many different fraud strategies, which in combination with the manipulation extent in deed impact the detectability.

Contents

Acknowledgments **v**

Zusammenfassung **vii**

Abstract **ix**

1 Introduction **1**

1.1 Election fraud 2

1.1.1 The definition of election fraud 5

1.1.2 Fraud mechanisms and fraudulent activities 6

1.1.3 Explanations of election fraud 7

1.2 Detection methods of fraud 9

1.2.1 Prediction models 10

1.2.2 Natural and field experiments 11

1.2.3 Vote and turnout distributions 13

1.2.4 Digit tests 14

1.3 Contribution 16

1.3.1 Analyzing the Performance and Linkage of Digit-Tests via Elections in France, Finland and Russia: 16

1.3.2 Fraud Mechanisms and Types that Benford's Law can and cannot Detect: Defining its Sensitivity and Specificity: 17

1.3.3 Election Fraud, Digit Tests and How Humans Fabricate Vote Counts: 18

1.4 Implications and Outlook 19

2 Analyzing the Performance and Linkage of Digit-Tests via Elections in France, Finland and Russia **23**

2.1 Introduction 24

2.2 Digit Tests 25

2.2.1 Mechanisms behind the 2BL and LD tests 27

2.3 Data and methods 29

2.4 Application of 2BL and LD to France, Finland and Russia 32
2.5 Cross validation of the 2BL and LD methods 34
2.6 Conclusion 38

3 Fraud Mechanisms and Types that Benford's Law can and cannot Detect: Defining its Sensitivity and Specificity 41
3.1 Introduction 42
3.2 The detection of election fraud and Benford's Law 43
3.3 Fraud coordination and actions 46
3.4 Data and test statistic 48
3.5 Artificially manipulated data 49
3.5.1 Calibrating different fraud mechanisms and simulating manipulated data 49
3.5.2 Manipulated election results 54
3.6 Detectability of election fraud 56
3.6.1 2BLs sensitivity 56
3.6.2 2BLs specificity 58
3.7 Conclusion 59

4 Election Fraud, Digit Tests and How Humans Fabricate Vote Counts 63
4.1 Introduction 64
4.2 The LD test and how humans manipulate electoral returns 65
4.3 Experimental design 69
4.4 Expected manipulative behavior 71
4.5 Results 73
4.6 Conclusion 82

A Supplementary Information for Chapter 2 87

B Supplementary Information for Chapter 3 91

C Supplementary Information for Chapter 4 97
C.1 Experimental design and procedures 98
C.2 Methods 102
C.3 Results 103

Bibliography 106

1 Introduction

Elections are a central part of the democratic process. They are "the formal process of selecting a person for public office or of accepting or rejecting a political proposition by voting" (Encyclopædia Britannica Online, 2016). Unfortunately, in the past as well as the present, there are actors who try to manipulate the local, regional or even the national election in order to gain political power, even at the cost of democratic legitimacy. Concurrently, if an election is perceived as fraudulent, citizens can lose their trust in the whole democratic process. Both aspects show how important it is that elections are not manipulated and that they are conducted as accurately as possible.

Election observers play an important roll to improve the quality of the elections. They do not only detect an report election fraud, but also reduce fraudulent activities solely through their presence (Hyde, 2007; Ichino and Schündeln, 2012; Enikolopov et al., 2013). However, election observers are confronted with various kinds of real-time information, which requires them to respond to it in a quick and flexible way. Due to the limited resources available, election observers can only examine a sample of polling stations within a country. Therefore, election observation missions can only give an approximate picture of fraudulent activities, but not the exact extent. Since election manipulators usually try to manipulate inconspicuously, it is generally very difficult to determine the exact extent of the manipulation. For these reasons, it is equally important to carry out a systematic analysis of past elections on the basis of different information, approaches and methods. Researchers are trying to determine when fraud occurred and how many votes were affected by it. This dissertation

focuses on digit tests, a statistical detection method of fraud that requires minimal information (only election results). The method has been increasingly utilized and gained popularity in recent years. Simultaneously, this fraud detection technique has become very controversial. Digit tests are surrounded by a high degree of uncertainty as to whether, why and which fraudulent activities they can detect or not. The aim of this dissertation is to solve this uncertainty to a larger extent.

This dissertation is structured as follows: The first chapter gives a short introduction to electoral fraud research and the three dissertation papers. It contains an overview on the various aspects of electoral fraud, the definition of fraud and various methods of fraud detection, including a brief description on the digit tests. The introductory chapter concludes with a report on the contribution of the individual dissertation papers as well as the outlook and conclusion of the dissertation. The second chapter presents the first dissertation paper. It examines digit tests in the comparative context of the presidential elections in France, Finland and Russia, which differ greatly in the expected election legitimacy. It evaluates the validity of the two digit tests in relation to specific fraud mechanisms and assesses linking test results with regard to theoretical expectation. The third chapter presents the second dissertation paper. This paper focuses on the specific manipulation mechanisms that the 2BL test should be able to detect. The manipulation mechanisms and types of fraud are simulated by artificially generated data. Moreover, the correct detection rate is calculated. This procedure improves the understanding of how the test statistic is affected through manipulation. The third dissertation paper (chapter 4) experimentally investigates how people would falsify election results by hand. This is the particular mechanism the last digit test can detect. The main question is whether these strategies are consistent with the assumption of how people manipulate election results and whether the falsified results can be identified as such by the last-digit test.

1.1 Election fraud

Elections should reflect the will of people, which is considered fulfilled by elections that are conducted free, fair and competitively (Alvarez, Hall and Hyde, 2008, 1). While this interpretation is straight forward, the term election fraud on the other hand is a diverse and colorful picture of different aspects. It comprises the definition of fraud, purpose of manipulation, fraud mechanism, particular fraudulent activities, its occurrence, fraud prevention as well as detection methods of fraud (compare Figure 1.1). Using statistical method to detect election fraud requires a profound knowledge of these aspects and the interactions between them.[1]

Figure 1.1 shows, for example, that election fraud can be interpreted as different things depending on the definition of a manipulated election. Therefore, the first

[1]The words election fraud and election manipulation are used interchangeably.

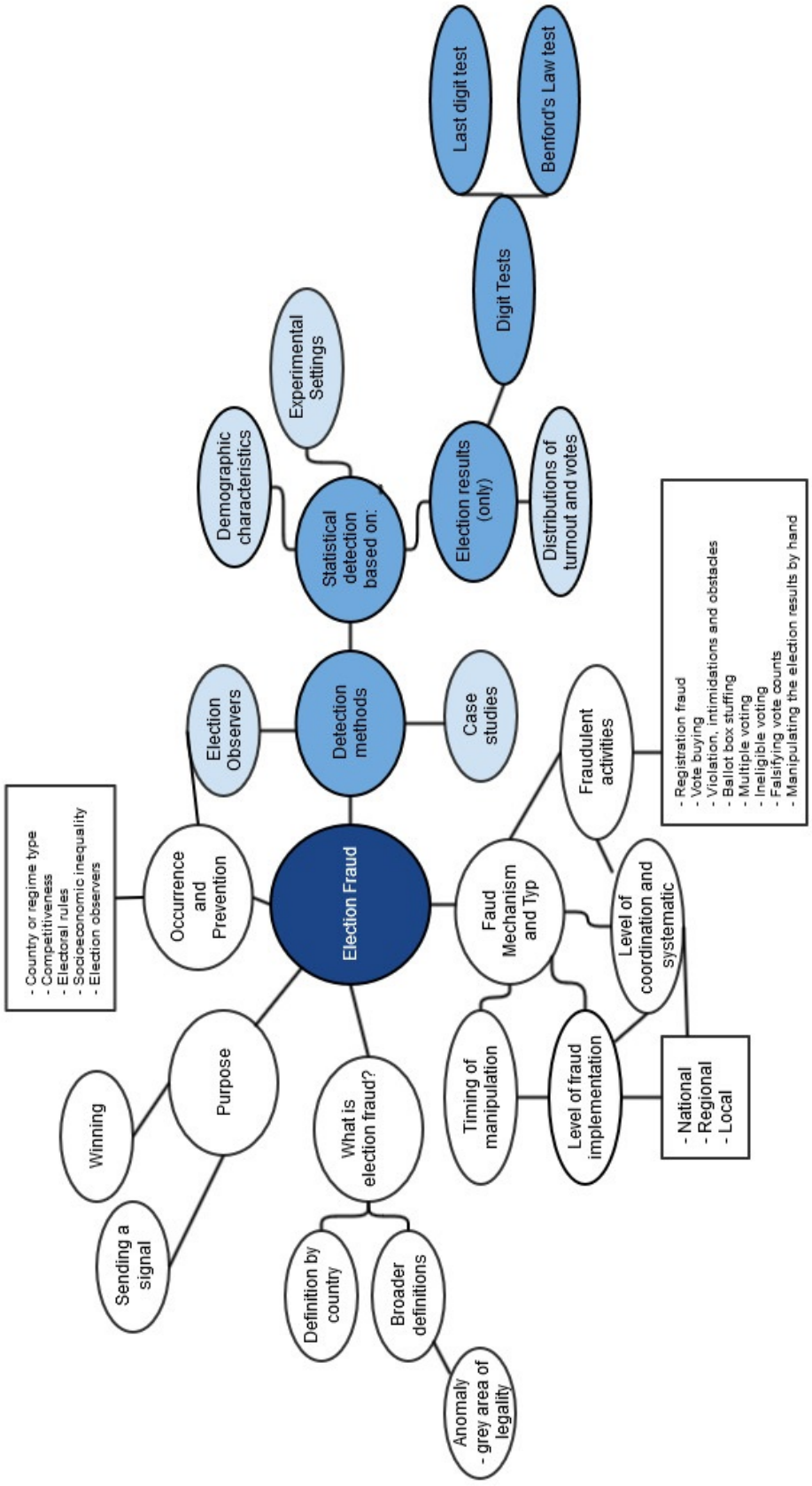


Figure 1.1: Election fraud

essential step of fraud detection is to clarify when an election is considered manipulated.

Closely connected to this aspect are the mechanisms of fraud and types of manipulation, as it can take different forms, which can impact the election outcome. Within this category it can be further distinguished between the fraudulent activities (e.g., vote buying, ballot box stuffing etc.), the timing of the manipulation, (e.g., if the fraudulent activity is carried out at the election day), at which level the manipulation is conducted, (e.g., at the polling stations or a higher level of vote aggregation) and how the fraud is coordinated. For instance fraudulent activities can be conducted independent from each other or they can be coordinated across a district, region or even across the whole nation. In the case of uncoordinated manipulation, there is little or no systematic approach across polling stations, which is why there are hardly any patterns that can be identified, making it difficult to detect this manipulation. Coordinated manipulation, on the other hand, significantly increases the likelihood of an election victory, but the coordination of electoral fraud across many polling stations has a systematic component and should therefore produce patterns of falsification that should be easier to detect by means of statistical methods. The identification of such patterns is an important basis for digit tests.

Another aspect of election fraud, displayed in Figure 1.1, is the purpose of the manipulation. This must be further distinguished into the aim of winning and sending a signal of power. If the aim of manipulation is to secure the election victory, it can be assumed that the falsification is as inconspicuous as possible and that it is difficult to identify this manipulation as such. However, if the power demonstration is in the foreground, manipulation is often not inconspicuously carried out and can therefore be identified using simpler statistical methods.

Studies that focus on circumstances and causes as well as the prevention of election fraud can be grouped into another category. The central questions here are when and why manipulation occurs more frequently and which strategies can prevent it. The knowledge about the causes of fraud and fraud detection complement each other, as it requires some knowledge about causes of fraud to design an adequate procedure of fraud detection. However, to investigate such causes there has to be some measure to capture the existence of election fraud.

The information of the subcategories has to be carefully considered as methods of fraud detection can only capture a certain range of each subfield. For example, the last digit test can only capture manipulation when people manually falsify the vote counts of polling stations. A manipulator might use this specific fraudulent activity after the votes are counted, since it can only be detected within the results of polling stations, but not after they are aggregated into the regional or national level. Moreover, a certain amount of polling stations have to be falsified otherwise the vote count

manipulation cannot cause a deviating pattern from the expected digit distribution. The individual categories are discussed in more detail below.

1.1.1 The definition of election fraud

The ideal version of free, fair and competitive elections is not feasible since even advanced democracies cannot perfectly realize democratic norms. For instance the entrance to the electoral arena is costly, the influence and jurisdiction of electoral offices is limited and the political content is not fully sovereign as it is determined by social and legal norms (Schedler, 2002, 38). If this is the case, what is considered election fraud? Is an election manipulated if only one vote is consciously or unconsciously miscounted? If not, how many miscounted votes are enough to label an election as manipulated? These kinds of questions make a unique definition of the term election fraud very difficult. In fact, there is no standard definition of the term election fraud. Alvarez, Hall and Hyde (2008) highlight that what is considered as fraud varies across countries and time. An action such as door-to-door campaigning is common practice for voter mobilization and persuasion in the USA, but it qualifies as fraudulent activity in Mexico as it apparently pressures voters (Alvarez, Hall and Hyde, 2008, 5-7).

However, there is a general understanding about important norms of elections which should not be violated. For example, the Copenhagen Document by the Organization for Security and Cooperation in Europe's (OSCE) Office for Democratic Institutions and Human Rights (ODIHR) articulate international standards for democratic elections, and election observer missions try to comprehensively cover these norms (Hall and Wang, 2008). In contrast, researchers usually define election fraud by focusing on the essential purpose of an election, which is to reflect voters' will in the election outcome. Correspondingly, manipulation is defined as "any activity that has the effect or intent of subverting the rights of voters to cast ballots free of intimidation or improper influence and to have their votes accurately counted without dilution by illegal ballots" (Goldberg, 1987, 180). Scholars define the difference between *accurate* and *inaccurate* election often through the election outcome. An election is inaccurate if manipulation is large enough to determine the outcome and does not reflect voters' intentions (Mebane, 2007; Carriquiry, 2011).

Such conceptualization of fraud obviously includes illegal activities of manipulation, but it can also cover some legal activities or circumstances. An exemplary case is the butterfly ballot used in Palm Beach Country, Florida 2000 presidential election. Wand et al. (2001) and Mebane and Sekhon (2004) show that this voting machine confused many democratic voters to mistakenly voted for Pat Buchanan instead of Al Gore. The estimated incorrectly casted ballots exceed the certified margin of victory of George W. Bush in Florida. Moreover, Florida was decisive for the outcome of the

entire presidential election.

1.1.2 Fraud mechanisms and fraudulent activities

The presented concept of election fraud includes a wide range of fraudulent activities that can occur at different phases of the election, which can be coordinated for instance across polling stations or districts to different degrees. OSCE/ ODIHR reports by election observers and election fraud literature almost always focus on fraudulent activities, its timing and the level of implementation, but mostly ignore the coordination. Coordination of fraudulent actions is to some extent visible in cases of blatant fraud like Russia where vote counts of Putin, the certain winner, were boosted in some republics by local elites (Myagkov, Ordeshook and Shaikin, 2008; Myagkov, Ordeshook and Shakin, 2009; Simpser, 2013). The common belief is that there has been systematic ballot box stuffing in these republics. Beyond such blatant cases, systematic coordination of fraud and its level of implementation are often implicitly assumed. There is a strong belief among scholars that manipulators will use fraudulent activities to reach a small winning margin, because anything exceeding victory would unnecessarily increase the cost of the manipulator (Alvarez and Katz, 2008; Bailey, 2008; Nyblade and Reed, 2008). This assumption is connected to a popular idea that the only reason to manipulate an election is to win it (Alvarez, Hall and Hyde, 2008). To reach small winning margins and a certain win requires some degree of coordination. Scholars further suggested that coordination is most cost-effective if the number of manipulated votes, necessary to swing the election, is small. Here too, information about the unit of interest is required and manipulation has to be coordinated across this unit in order to achieve the optimal outcome for the manipulator. Based on the assumption of fraud coordination, manipulators know more or less how many votes have to be manipulated and how to distribute them across the unit(s) in which the manipulations take place. The critical number of manipulations can be achieved, for example by systematic ballot box stuffing or tempering with vote counts. Manipulation can also take place at polling stations more or less independent from each other. However, this carries the risk that the election victory will remain uncertain even when there are different fraudulent activities at multiple polling stations. If manipulation is uncoordinated, a manipulator could intimidate voters, vote multiple times or change the vote count of the polling station without knowledge of fraudulent activities in other polling stations. As mentioned, these manipulations will not necessarily affect the election results. Table 3.1 of the second dissertation paper contains the fraudulent activities that could be implemented as coordinated or uncoordinated manipulation.

As already indicated, coordinated or uncoordinated fraud can comprise different fraudulent activities. OSCE election observer reports name a wide variety of observed

types of fraud as well as the list of common tactics by Simpser (2013, 35). Before and at the time of casting ballots the following kinds of activities are used to possibly manipulate election results:

- creating obstacles to candidate registration and/or voter registration
- tampering with voter registration lists
- intimidating candidates and/or voters
- vote buying
- voting multiple times
- voting by those who are ineligible, such as minors

The above activities aim to influence candidates and/or voters directly (e.g. vote buying) or indirectly (e.g. obstacles to registration). In contrast, fraudulent activities can also occur after casting ballots. At this stage, the activities are aimed at the casted ballots and/or election results themselves:

- stuffing or destroying ballot boxes
- tampering with the vote count
- falsifying results

Fraudulent activities always aim to affect the absolute and relative vote share of individual candidates/parties. At the same time, fraudulent activities can also change another aspect of the election result: voter turnout. In most cases a variety of frauds that affect both turnout and the vote counts of candidates/parties are used to manipulate an election (Simpser, 2013).

Election fraud changes the election results, which should be identifiable. Given the current state of the art in methods for fraud detecting, only specific fraudulent activities (e.g. falsifying results) and fraud mechanisms (e.g. coordinated fraud) can generally be identified.

1.1.3 Explanations of election fraud

A popular idea among scholars is that the only reason to manipulate an election is to win (Alvarez, Hall and Hyde, 2008). Based on this assumption small winning margins should be mostly observed in manipulated elections, because anything exceeding victory would unnecessarily increase the cost of manipulation (Alvarez and Katz, 2008; Bailey, 2008; Nyblade and Reed, 2008). Simpser (2013) finds that small

winning margins are not the norm and large winning margins occur more frequently than expected. Large winning margins are especially prominent in authoritarian regimes and are often considered to be blatant cases of manipulation with the aim to send signals of power (Simpser, 2013). Gehlbach and Simpser (2015) build on this argument and suggest that such manipulation can solve the problem of bureaucratic compliance as it influences bureaucrats' beliefs about the power of the manipulator, which encourages bureaucrats to work on the manipulators behalf.

There are studies that intend to capture the causes and determinants of election fraud which requires some knowledge about the existence of fraud. Blatant cases of manipulation are usually the starting point for such investigations as obvious manipulations give at least some information about the existence of fraud. Scholars identified competitiveness (Lehoucq, 2003; Simpser, 2013), structural and economic changes (Lehoucq and Jiménez, 2002), the electoral rule (Lehoucq, 2003), socioeconomic inequality (Ziblatt, 2009), and the presence of international election observers (Hyde, 2007) as influential factors of election fraud.[2]

Election observers take on more than one important role in the process of an election. They observe the whole process of an election and note any irregularities. This gives an idea about the legitimacy of an election and is used by many scholars as a measure of fraud. Election observers also have a preventive function and its effectiveness have been demonstrated multiple times (Hyde, 2007; Ichino and Schündeln, 2012; Enikolopov et al., 2013). Consequently, a fraudulent measurement based on election observation cannot fully capture manipulation as it is limited in its scope due to a sample of observed polling stations and its impact on the observed manipulation.

On the one side, detailed knowledge why fraud occurs and which factors impact manipulation can improve methods of fraud detection. For instance, some studies use specific circumstances such as the random assignment of election observers to estimate the intensity of fraud and explore some mechanisms of fraud in that particular setting (Ichino and Schündeln, 2012; Enikolopov et al., 2013). The studies incorporate the affect election observers have on reducing fraud in the design of their field experiment, which makes the more precise estimation of fraud possible. On the other hand, scholars implicitly assume that there is a big discrepancy between the actual happening of election fraud and our knowledge about it. Manipulators are assumed to "typically wish to hide these illicit activities" (Ichino and Schündeln, 2012, 292), at least when the manipulation is not aimed as a signal of power. As accounts for fraud are often based on obvious cases of manipulation, the explanations are likely based on a skewed sample of manipulation incidences. Therefore, meth-

[2]Specific examples are the study by Lehoucq (2003) who analyzed a period of almost 50 years in Costa Rica and identified that "[p]lurality districts generate more accusations of fraud than proportional-representation ones" and that of Hyde (2007) who shows via a natural experiment in Armenia that election observers significantly reduce manipulation.

ods of fraud detection also complement the explanations of fraud as they extend the knowledge about the existence of the manipulation beyond the blatant cases. If digit tests become valid methods of election fraud detection, it could enrich research about causes and consequences of election fraud. Digit tests are theoretically not limited to the detection of extreme fraud. Instead, a critical number of vote counts have to be manipulated, e. g. by exchanging them manually with made up numbers. A small, but relatively widespread changes in vote counts should be sufficient for fraud detection.

1.2 Detection methods of fraud

The common factor in fraud detection methods is the idea that election fraud is a significant deviation from fraud-free "normal votes". The term "normal votes" refers to a hypothetical election result, which should be obtained if there were no fraudulent activities in the electoral process. However, it is difficult to set up such fraud-free normal votes since the exact processes generating such normal votes is unknown. More specifically, one can never be sure whether all relevant factors behind such normal votes (e.g. composition of citizens, issues at stake, candidates' characteristics) are known. Further, even if one can assume to have information about all relevant factors, individual citizens' behavior can have inherent randomness (e.g. citizens' cognitive errors). Therefore, normal votes cannot be specified as deterministic. Instead, the normal votes are conceived to be the outcomes of certain random processes. Consequently, the significant deviation of empirical data from the normal votes should signal the existence of fraud. In other words, a certain fraudulent activity is conceived as an intervention of certain random processes. By using such an approach, one can also treat different kinds of context-specific factors which are irrelevant to fraud as random factors. This is simultaneously an important assumption which should hold to distinguish fraudulent elections from those being fraud-free.

This section presents the assumptions made in various statistical methods of detecting electoral fraud. The first set of approaches constructs the fraud-free random process in a more theory-guided way, which requires some predictors of normal votes. The next set utilized natural- or field-experimental situations, which realize a random process. The third set tries to model more directly the random process in a single election result. Among them, the so-called digit tests are discussed separately. Such test focus, as mentioned before, on the distribution of a specific digit in the election result.[3]

[3]Classification of statistical methods of fraud detection varies between scholars. For example Montgomery et al. (2015) roughly classifies the approaches of those that compare results with an empirical baseline and those that compare them with a theoretical baseline.

1.2.1 Prediction models

The very first attempt of fraud detection in political science research is based on the idea that one can predict normal votes by using sociodemographic information. After predicting the normal votes by using e.g. regression analysis, one can conduct the outlier analysis to identify the fraudulent cases. One of the first known analysis was done by Powell (1989) who used OLS regression to model the aggregated voting results about the ratification of Mississippi's constitution in 1868. As a predictor, he used race, party membership and the registration rates. The outlier analysis identified 15 districts whose results significantly deviate from the predicted vote share. Based on the results, he concluded that there was likely to be fraud in the 15 districts. Baum and Hailey (1994) conducted a similar analysis on the 1948 Texas Senate race and found some suspicious cases.

The rather naive model based on the OLS regression with socio-demographic information has been extended by further model specification and robuster models. Wand et al. (2001) modeled the 2000 presidential election in Florida by combining a series of demographic characteristics with previous election results. The authors speculated that many voters who actually intended to vote for the democratic candidate Al Gore mistakenly voted for the rather unknown candidate Buchanan. The confusion is likely due to the Butterfly Ballot, used in Palm Beach County. The results supported this speculation and provided the minimal proportion of votes that were accidently given to Buchanan, which would have been enough to turn the election results in Florida. Mebane and Sekhon (2004) show a further possibility to improve the analysis of Wand et al. By including more time and society specific factors and additionally using more robust statistical models, they could improve the estimation of the normal votes, and in turn the precision concerning confused voting.

The approaches above assumed that the significant deviation from the normal votes signals election fraud. This can be the other way around: the nearly perfect congruence of the empirical result with certain expected value may be interpreted as fraud. As such a case, Delfino and Salas (2011) investigated the 2004 Venezuela referendum to recall the president. Besides the official result of the referendum, the proportion of registered voters who signed the petition for the referendum in each polling station was available. One may conceive it as normal votes. However, the authors found a suspiciously high correlation between the number of signatures and YES votes (both below 50%) at the computerized polling stations. In other words, it means that almost all citizens who did not sign the petition voted NO at the computerized polling stations. Given the suspiciously high correlation and also difference between computerized and manual polling stations, the authors conclude that the results in the computerized polling stations were manipulated in favor of the incumbent.

This 2004 Venezuela referendum was further investigated based on the idea of the significant deviation from the normal votes. Prado and Sansó (2011) predicted the normal vote based on exit poll data from two independent surveys. The estimated differences between the exit polls and the official results are significantly large and corresponds to the conclusion of Delfino and Salas (2011). Hausmann and Rigobón (2011) consider both aspects of the above analyses and draw the same conclusion. Martín (2011) relies on a different and innovative data source of the 2004 Venezuela referendum: the number of bytes of incoming and outgoing data to the servers of the Consejo Nacional Electoral (CNE), the connection time between the servers and the polling stations and the data packets in the transmissions. While the data traffic should not differ across polling stations, there existed unexplainable differences in the information volume among polling stations. Given the results, the author is doubtful about the integrity of the referendum, but also raised possible technological explanations.

All studies reviewed above suffer from some drawbacks. First, researchers need additional information to predict the normal votes. Some of them (e.g. the traffic data from the electronic voting data transmissions of Martín, 2011) are hardly obtainable. Second, and related to the first point, there exists the risk of an incorrectly and/or incompletely specifying the model of normal votes. Political science has a long tradition of voting behavior research. Recently there are also a series of attempts to forecast election results. There are numerous models of voting behavior and election results, which vary strongly in respect to the statistical model as well as model specification. It is challenging to find the correct model for the election at stake.

One possible solution may consist in the fact that socio-demographic information is often temporarily stable and has some spatially clustered distribution. If it is the case, one can find normal votes in the past or neighboring election data, which share the similar socio-demographic information with the election results at stake. For example, Alvarez and Katz (2008) used the aggregated votes in geographically close districts to estimate if manipulation occurred in the 2002 general election in Georgia, USA. The same idea is also behind the model of Wand et al. (2001) who used previous election results. Moreover, Myagkov, Ordeshook and Shakin (2009) follow this assumption: By estimating the "flow of votes" between succeeding elections they identify whether a candidate receives an oversized share of support from previous nonvoters, which should signal election fraud.

1.2.2 Natural and field experiments

Natural- and field-experimental approaches utilize a specific natural setting where some fraud-relevant factors can be explicitly set up and/or measured as treatment in only some units. A treatment can suppress (e.g. by assigning election observers) or

trigger (e.g. due to elections per se, see below for more detail) election fraud. In both cases, comparison of the units with and without such treatment enables inference of the existence and magnitude of election fraud. To avoid the influence of potential confounders, random or as-if random assignment of treatment to individual units is required (Dunning, 2008, 2012). For the natural experiment with as-if random assignment, it is also crucial that treated units are compared with controlled units. Here, researchers usually made the same assumption as above, which is spatially neighboring units are similar in their normal votes. Therefore, it is preferred to compare such neighboring units with and without treatment.

As fraud-relevant treatment, it is straightforward to utilize presence of election observers, which is expected to suppress fraudulent activities. Enikolopov et al. (2013) utilized this treatment in Moscow during the 2011 Russian parliamentary elections to estimate the intensity of election fraud. Election observers were randomly assigned to 156 out of 3,164 polling stations. The assignment was secret until the last minute so that potential manipulators could not strategically react to the assignment. The analysis compares polling stations with and without observers within each of the 125 electoral districts in Moscow. Therefore, the authors only compare people who live very close to each other. Their result shows that the presence of independent observers decreased the vote share of United Russia by almost 11 percentage points.

Callen and Long (2015) aimed to investigate a specific kind of election fraud, which they call aggregation fraud. This kind of fraud takes place in the aggregating process after ballots were cast and counted at polling stations through manipulating vote totals. To investigate it in the context of the 2010 parliamentary election in Afghanistan, they utilized photographs of vote tally sheets at sampled polling stations before aggregating them at a higher level. By comparing the photographed sheet with the official and published national-level election results, they measured the intensity of election fraud in individual polling stations. This measure is very innovative and quite important since it revealed a significant amount of election fraud and provided detailed information about which specific kind of aggregation fraud took place. Beyond the measurements, the authors conducted an experiment mainly to assess the impact of announcing the use of photographs on fraudulent activities, which also provides information about the intensity of election fraud. For this purpose, they announced the use of photographing vote tally sheets in the randomly selected 238 polling stations out of the 471 sample stations. By comparing the results, they confirmed that the announcement of taking photographs reduced fraudulent activities in form of damaging election materials and changing vote counts in the aggregation process. They resulted in the reduced number of votes for candidates connected with the local election officials in charge of the aggregation process.

Researchers can also exploit certain situations to investigate fraudulent activities in the voter registration process. Fukumoto and Horiuchi (2011) speculated that a

significant number of Japanese citizens changed their registered address to a municipality with an election in near future so that they can be eligible to vote for their preferred party. This is fraudulent behavior if the address only changed on paper but their actual residence has not changed. Therefore, the authors estimated the difference in residential registration rates between municipalities with a municipality election (treated) and those without election (control) in 2003. The estimated effect of the existence of an election was substantial enough to affect competitive municipality elections.

Ichino and Schündeln (2012) use a similar approach to investigate registration fraud in the 2008 Ghanaian general election. They apply a two-level randomized field experiment to estimate the effect of domestic election observers on irregularities in voter registration. Their treatment is the presence of domestic election observers. To assign the treatment, they first block constituencies based on the vote share of the previous election in 2004. Within each block, two constituencies are randomly selected as control, while one constituency is randomly selected for the treatment. The treatment (domestic election observers) is however not assigned to all polling stations of the treated constituency, but randomly assigned only to some polling stations. Therefore, there are treated and control polling stations in the treated constituency. This two-step procedure enables inference about differences (1) between treated and controlled constituencies, (2) between treated and controlled polling stations of a treated constituency and (3) spillovers to neighboring polling stations. The findings suggest that the registration rates are lower in polling stations with domestic observers. There is a local spillover effect that indicates fewer manipulations in districts that are geographically closer to the observed ones. Therefore, it seems plausible that manipulation efforts are reallocated to other districts. Moreover, the overall registrations rates are considerably higher in non-treated constituencies.

As demonstrated above, experimental approaches are particularly useful for learning about the magnitude of specific fraudulent activities. However, the design strongly depends on the setting, which enables the explicit assignment or measurement of treatment. If the treatment is randomly assigned to units, it is often the case that one can investigate only a limited number of units due to various resource limitations. For these reasons, it is difficult to generalize the results beyond the analyzed sample.

1.2.3 Vote and turnout distributions

In contrast to the approaches above, the methods below solely rely on the official results of the election at stake. They are based on the assumption that election results generated in fraud-free elections should follow a certain probabilistic distribution. Thus, certain theoretical probabilistic distributions serve as normal votes.

Researchers start with an important assumption that turnout and/or vote counts in fraud-free elections are more or less normally distributed. This can be disturbed

by ballot box stuffing, which for instance artificially increases the turnout in the manipulated districts, resulting in a skewed distribution of turnout rate. Based on this idea, Myagkov, Ordeshook and Shakin (2009) establish a set of indicators that are useful for detecting manipulation in former Soviet countries. In addition to such indicators, Kobak, Shpilkin and Pshenichnikov (2012) visualize extreme manipulation in Russia, which is defined as extremely high correlation between (nearly) complete turnout and the vote share of Putin or his party.

Klimek et al. (2012) build on both studies to establish a parametric model quantifying the extent to which fraudulent activities influenced the observed election results. The deviations of the statistical features of voting results are investigated in a cross-national setting and seem relatively independent from the aggregated level of the election data or the size of the sample. This method is mostly designed to capture the fraudulent activity of ballot box stuffing. The two different parameters indicate (1) whether a given rate of ballots are added in favor of a particular party and/or votes from other parties are taken away, and (2) whether turnout is almost 100% and almost all votes are cast for one party. The estimated parameters signal anomalies that are likely to be attributed to fraud.

The strength of this approach is that it requires only the official election results and the logic behind the approach is quite intuitive: the manipulating party takes the ballot from the other parties and inflates the voter turnout in its own favor. However, there are still several drawbacks. First, the extraordinary high turnout rate and dominance of a single candidate/party may be due to the regional distinctiveness and/or successful voter mobilization. Second, this approach strongly focuses on a specific fraudulent activity: ballot box stuffing. For the other possible fraudulent activities, it is not clear which kind of deviation from the normal votes is expected. Further, the application of this approach is so far limited to the national level, while election fraud is likely to have certain geographic difference in its magnitude.

1.2.4 Digit tests

Digit tests also only require the current official election results like the approaches investigating vote/turnout distribution. In recent literature, digit test received a lot of attention and the number of applications is high. The digit tests are also based on the assumption that fraud-free election results should follow a certain theoretical distribution. The difference from the approach in the last section is that the distribution of vote counts or turnout rate is not at stake, but rather the distribution of a certain digit. More specifically, most studies using a digit-based test focus on either the second digit or the last digit of the election results. The tests focusing on the second digit is often called the second digit Benford's Law (2BL) test. The so-called last digit test examines the distribution of the last digit of vote counts. Both are briefly introduced.

The 2BL test relies on Benford's Law, which gives the frequency distribution of the leading digit of diverse kinds of existing numerical data such as the surface areas of rivers, or the number contained in an issue of Reader's Digest (see for more examples Benford, 1938). The frequency distribution is parameter-free and each number has a certain frequency independently of the data at stake (1's appear the most with 30.1%, 2's with 17.6% and 9's appear the least with 4.6%). This distribution has been generalized by Hill (1995) to the frequency distribution of the further digits. It was no wonder that some researchers came to the idea to use Benford's Law as normal votes to detect election fraud (Pericchi and Torres, 2004; Mebane, 2006*a*). That is, one collects frequencies of a certain digit of election results at the polling station and compares them with Benford's Law. A significant deviation from Benford's Law should signal that election fraud intervened the "natural process" that generates election results. In most applications of Benford's Law, researchers prefer to use the second digit instead of the first digit. This is because a significant deviation of the first digit distribution can appear although election fraud does not exist. Given that one country has a constant size of polling station (e.g. 1000) and the support level of a candidate is also constant across the polling stations (e.g. 25%). In this case, we expect similar vote counts for the candidate (about 250) and an equal leading digit number ("2"). This kind of circumstance is much less likely for the second digit.

The analogous procedure applies to the last digit test. Here, researchers assume a uniform distribution instead of Benford's Law as the normal vote and compare it with the empirical frequency of the last digit in the election results. Scholars considered the human inability to fabricate random numbers when manually manipulating result sheets as an adequate mechanism causing the last digit to distribute differently (Beber and Scacco, 2012; Weidmann and Callen, 2013).

The digit-based tests have been increasingly utilized in recent fraud detection literature (see Table 2.1). The reasons are twofold. First, the application of digit tests is relatively simple. They require little information for the analysis, which is the vote counts of the election results at the lowest level of aggregations. Many scholars hope or expect that method can be used as a standard tool to screen elections for indications of fraud. Second, digit tests and especially the 2BL test are highly controversial in the literature. Many scholars raised valid critique about the 2BL test (Mebane, 2010*a*; Deckert, Myagkov and Ordeshook, 2011; Mebane, 2011; Shikano and Mack, 2011; Mebane, 2012). This is partly due to mixed and sometimes contradicting test results, but more importantly due to the fact that it is still unclear why fraud-free normal election processes can generate results following Benford's Law and why only election fraud can disturb that process resulting in deviation from Benford's Law. Shikano and Mack (2011) even demonstrated some circumstances in which the second-digit's frequency distribution does not correspond to Benford's Law without election fraud, just like the first-digit's distribution.

Although the validity of the last digit tests has not yet been so controversially debated as that of the 2BL tests, the same criticisms against the 2BL test can hold for the last digit tests. For both digit tests, it needs a better understanding if, how and why election fraud can disturb the digit generating process.

1.3 Contribution

In the broader framework, this dissertation contributes to the detection of election fraud and thereby elevating the legitimacy of an election. In particular, it contributes to digit tests. Their use and impact is constantly growing. The accepted view is that digit tests can be applied in situations where no other information other than the vote counts of election results are available. Therefore, in such situations most alternative fraud indicators cannot be applied. Digit tests can be applied to any election as soon as the results are published. The tests are useful to investigate data of districts as well as the national level (Mebane, 2006*b*; Beber and Scacco, 2012; Medzihorsky, 2015), and they are relatively easy to implement. This distinguishes them from approaches such as that of Klimek et al. (2012) which are difficult to implement and only applicable at the national level. Last but not least, the application of digit tests are apparently not specific to the election, country, or culture (Montgomery et al., 2015). However, digit test are sometimes described as a "magic black box"[4] which reflects well the confusion among scholars about how the digit test works. The three papers of the dissertation elaborate different aspects of the digit test and gives a more holistic understand about them, their potential as a fraud indicator and the impact of fraud actions and mechanisms on vote counts and the digit distribution. It contradicts the view that digit tests are easily applicable without additional information to all kinds of elections. Moreover, it shows that superficial understanding and applications without detailed knowledge can lead to misinterpretation of test results.

1.3.1 Analyzing the Performance and Linkage of Digit-Tests via Elections in France, Finland and Russia:

Many scholars once claimed that there is little systematic research on the detection of election fraud and that applications of the Benford's Law test is often not even peer-reviewed work (Deckert, Myagkov and Ordeshook, 2011). This criticism is outdated, since there are a large number of applications outside and inside science. The first dissertation paper summarizes the literature based on important aspects such as the context of the application, expectations, findings and contradictions.

In the literature, the Benford's Law test has especially been criticized as being vague about what a significant test statistic signals and its theoretical foundation. To fill the gap concerning the theoretical foundation, the first paper elaborates on the linkage of fraudulent activities and the test statistic, and also derives explicit hypoth-

[4]Deckert, Myagkov and Ordeshook (2011) use that term in respect to the Benford's Law test.

esis about them. Moreover, this makes it possible to establish a link between the last digit test and the Benford's Law test and derive expatiations about their connection. This long neglected combined perspective should improve the understanding of digit tests.

This work is closely connected to findings of Shikano and Mack (2011) who showed narrowly dispersed vote count distributions violate the assumption of Benford's Law, which in turn result in a fraud signal without substantive election fraud. We use institutional settings which should a priori comply with the distributional assumptions of digits tests. Furthermore, previous studies mostly used single countries for their analysis of which the specific characteristics can produce deviating results despite fraudulent activities. We address this issue with a comparative analysis for elections of Russia, France and Finland.

Significant deviations of digits tests are only expected in Russian elections for which there is some certainty that fraud occurred (Myagkov, Ordeshook and Shakin, 2009; Kobak, Shpilkin and Pshenichnikov, 2012; Klimek et al., 2012; Enikolopov et al., 2013). We build on this literature and use measures of extreme election fraud for cross validation of digit tests. Moreover, this is the ideal setting to investigate the linkage between digit tests.

We find few indications of fraud or anomalies in France and Finland, and some indications of fraud or anomalies in Russia. The linkage of digit tests is not as expected and we could not link significant deviations of the Benford's Law test to the measures of extreme fraud. These findings are further supported by the non-significant results of digit tests when they are applied to intentionally manipulated data. Therefore, we conclude that significant digit test statistics do not necessarily indicate manipulation at all.

1.3.2 Fraud Mechanisms and Types that Benford's Law can and cannot Detect: Defining its Sensitivity and Specificity:

The second dissertation paper builds on the findings of the first paper and goes into a detailed investigation of the 2BL test to improve the understanding of its function. In respect to Benford's Law, the first paper establishes the theoretical link of fraudulent activities, which should be detectable by the 2BL test. The fraudulent activities are very likely present to different degrees in Russia, but not in France and Finland. The test cannot differentiate between these cases to a satisfactory level and additionally fails the cross validation with other fraud parameters. However, as there is uncertainty about the specific fraudulent activities and the extent and intensity of manipulation in the different region, the paper cannot fully capture the performance of the 2BL test in respect to its sensitivity (manipulations that are correctly identified as such) and specificity (fraud-free data that is correctly identified as such). This

important information is needed to evaluate how useful the test is to identify election fraud. The question has been partly addressed by Deckert, Myagkov and Ordeshook (2011) who use a simulation study to define the sensitivity and specificity. The simulation is based on the normal distribution, which cannot adequately represent real election results and violates the assumption of the 2BL test (Mebane, 2011). Moreover, the findings by Deckert, Myagkov and Ordeshook (2011) might not be valid as 2BL has been shown to be sensitive to election characteristics like strategic voting and narrowly dispersed vote count distributions (Mebane, 2010*a*; Shikano and Mack, 2011) that are neglected by the simulation approach.

The research carried out so far has dealt only a little or broadly with the assumptions of the 2BL test and the resulting consequences. In this study, common fraud mechanisms and activities are associated with theoretically expected deviations from Benford's Law. Mechanisms and activities that 2BL but not LD test should be able to detect are quantified, like adding votes through ballot box stuffing, modification of the computer software etc. To this purpose the study uses the 2008 Canadian election which is highly likely to be free of fraud. The election results are used to calibrate simulated data, which are manipulated by different fraud mechanisms. The procedure enables a quantification of the sensitivity in respect to different fraud intensities and different levels of spread across polling stations. Within this framework, special attention is paid to the systematics of manipulation, as it plays a central role for possible deviations from 2BL. In order to complete the detailed image of 2BL, the specificity of 2BL is also determined. This is based on a calibrated fraud-free election model.

The paper finds that the detectability by 2BL test depends on how systematic the manipulation is implemented. If many polling stations and many votes are manipulated with some system, then the 2BL test can detect about 70% of the manipulation correctly. In any other scenario it detects less than 50% of the manipulation correctly. The results of the specificity are not as striking, but they support the conclusion to not trust it as a technique for detecting election fraud.

1.3.3 Election Fraud, Digit Tests and How Humans Fabricate Vote Counts:

Paper three uses a laboratory experiment approach to study how humans manipulate election results. This contributes to the election fraud literature in general as it gives valuable insights about how a person or people might manipulate election fraud. Until now, there exist only an assumption about how this kind of manipulation is conducted, but it has never specifically been investigated. Within this broader framework, detailed knowledge about the applied strategies can help to develop actions of fraud preventions.

The experimental approach makes it possible to test the distributional assumption if a person or people replace the vote counts by made up numbers. It investigates which strategies they use and how extensively they manipulate in respect to the number of polling stations. Additionally, it captures whether or not the behavior is affected by the intensity of the manipulation and the setup in which the manipulation is conducted. Moreover, it can capture whether the strategies applied to conduct the manipulation affect the last digit test's detectability and whether humans are bad random number generators in the context of election manipulation under the control of non-random strategies.

The findings show that the link between the human vote count manipulation and last digit deviations is not as simple as suggested. Subjects used many different strategies to manually manipulate election result sheets. Among applied strategies there is one in which subjects replace vote counts with new numbers, but they also swap votes between candidates and manipulate a different digit than the last one. Finally, the prediction of the regression analysis shows it would not trigger the LD test to signal election fraud, even if 100% of LD are replaced by man-made numbers.

1.4 Implications and Outlook

The findings of the dissertation improve the understanding of digit tests as election fraud detection tools. Unfortunately the findings of all three dissertation papers coherently show that digit tests are not valid tools to detect election fraud. As digit tests became "standard tools" (Pericchi and Torres, 2011; Cantú, 2014) in this field, the findings impact the whole discipline as they strongly suggest to focus on different methods of fraud detection.

In the first dissertation paper, the empirical basis for the strong mistrust of digit test was created. This shows that obvious and strong election fraud is not accompanied by significant test results and that the expected relations between 2BL and the last digit test are not correct. In the second dissertation paper it was examined in detail which of the manipulation mechanisms and fraudulent activities can actually be identified by 2BL. As a result, it is found that the systematic changes in the number of vote counts due to all manipulation mechanism or fraudulent activity are not sufficient to use 2BL as a reliable fraud indicator. This means that there is no specific manipulation mechanism or fraudulent activity that 2BL can detect. In the third dissertation paper it was shown that even the assumption of how people falsify election results by hand falls short because people manipulate very strategically the electoral results. Even if they are manipulating according to the assumption, on average this does not lead to significant deviations in the last digit test.

Based on these results, recent developments in this area of research can be regarded as very critical. Very influential and recently published work combine digit

tests with other fraud detection methods: a machine learning algorithms or using an alternative statistical approach based on latent classes (compare Cantú and Saiegh, 2011; Montgomery et al., 2015; Medzihorsky, 2015). The identified problems of the Benford's Law test and the last digit are very unlikely to be solved by adding another "black box" through its combination with a machine learning algorithm or a different statistical approach for estimation. Moreover, there is only a cumulative value when combining different fraud detection methods, as the election forensic toolkit (compare Mebane, 2015), if the individual method can contribute to fraud detection. If the individual methods of the toolkit or other combined approaches to uncover election fraud cannot make an individual contribution, they will probably lead to the aggregation of false information. This problem underlines how important it is to examine the strengthening and weakening of individual methods and to question the respective assumptions as it was done in this work.

Digit tests cannot contribute to the evaluation of the legitimacy of an election. Therefore, further research should concentrate on elaborating other exciting procedures of fraud detection. A promising starting point is the estimation techniques by Klimek et al. (2012) and its improvements by Mebane and Wall (2015). However, there is potential in improving the estimation of fraud parameters, extending the model to the subnational level, which can improve the precision of fraud detection and allegation. Additionally, demonstrating the validity of the method is important to establish trust and reduce uncertainty about the sensitive issue of fraud allegation.

Using a machine learning algorithms that integrates structural information and forensic indicators is another promising approach for establishing a detection tool that is applicable to a wide range of elections and does not depend as much on the presents of one particular fraudulent activity. This requires work in line with Montgomery et al. (2015), but such approaches should mostly refrain from integrating digits test. Combining information about election fraud within such a framework is likely to be the future of the statistical detection techniques.

2

Analyzing the Performance and Linkage of Digit-Tests via Elections in France, Finland and Russia

Verena Mack and Susumu Shikano

Abstract

Expectations surrounding the legitimacy of elections vary widely among countries. While some elections are perceived to be manipulated, others are believed to be fraud free. Scholars have developed techniques such as digit tests to detect election fraud systematically. Such tests have many different applications, but also a variety of different findings and interpretations. We investigate presidential elections in France, Finland and Russia, which differ strongly in their perceived legitimacy, and evaluate the validity of the second digit Benford's Law test, the last digit test as well as their linkage. In particular, we consider specific fraud mechanisms that each digit test can (and cannot) capture and validate digit test results in Russia. We show that significant deviations from the second digit Benford's Law test do not correspond convincingly to extreme fraud indicators. We contradict the interpretation that both detect different types of manipulation, and conclude instead that significant test statistics do not necessarily indicate manipulation.

2.1 Introduction

The expected and perceived legitimacy of elections varies among countries. While, for example there were multiple indications of strong election fraud in the 2012 Russian presidential election (compare Bader and Schmeets, 2013; Kobak, Shpilkin and Pshenichnikov, 2012; Klimek et al., 2012), this is not the case for presidential elections in most established European democracies. To distinguish between manipulated and fraud-free elections, many scholars rely on systematic "election forensic" techniques such as digit tests (Pericchi and Torres, 2011; Cantú, 2014). Digit tests flag anomalies in the distribution of a specific digit of electoral vote counts. They have gained popularity in recent years because they are simple to use and interpret, and require minimal information (vote counts of elections). Two known classes of such tests are the Benford's Law test (BL) (Roukema, 2009), mostly applied as the second digit Benford's Law test (2BL) (Mebane, 2006*b*), and the last digit test (LD) (Beber and Scacco, 2012).

While the popularity of digit tests has risen, their increasing use has also revealed shortcomings. First, their results are often inconclusive, and sometimes contradict other fraud indications. This has raised concerns about their validity and usefulness, especially with respect to different applications of BL (Mebane, 2010*a*; Deckert, Myagkov and Ordeshook, 2011; Mebane, 2011; Shikano and Mack, 2011; Mebane, 2012). Second, the majority of the empirical analyses are based on single-country applications; however, the specific vote count distributions in a particular country do not necessarily fulfill the assumptions of digit tests (Shikano and Mack, 2011). Third, the underlying manipulation mechanisms for fraud detection have rarely been considered, yet they are relevant for understanding which digit test can detect which kind of election fraud. Furthermore, the linkage between digit tests has been completely ignored.

Given these shortcomings, this paper contributes to the literature in four main ways. First, we summarize the previous (and sometimes contradicting) applications of digit tests. Second, we identify the specific types of fraud that each test should be able to detect. By doing so, we establish the linkage between 2BL and LD. Third, we conduct a comparative analysis using election data from Russia, France and Finland, the institutional settings of which should comply with the assumptions of 2BL and LD. Fourth, we use extreme election fraud indicators from Russia to cross validate the digit tests. According to our results, significant deviations from 2BL and LD do not correspond convincingly to extreme fraud indicators. Therefore we conclude that the application and interpretation of digit tests should be done with caution.

The paper is structured as follows. The next section briefly summarizes the concept of digit tests, including their applications and the main criticisms, and discusses the types of fraud that each digit test should be able to detect. The third section in-

troduces the data from France, Finland and Russia, while the fourth section presents the application of 2BL and LD distribution. Subsequently, we conduct cross validations with extreme fraud indicators and use simple fraud scenarios to assess the performance of 2BL and LD tests. The last section concludes.

2.2 Digit Tests

Digit tests are based on the assumption that the digits of fraud-free vote counts follow a particular theoretical distribution. Thus, deviation from this distribution should indicate fraud or anomalies. Depending on the investigated digit, one can use different theoretical distributions, the most common of which are the 2BL and LD tests.

Digit tests have been increasingly applied over the last few years, mostly to different countries but also to simulated elections. For those investigations, scholars either suspected fraud, did not suspect fraud or used digit tests as a screening method without *a priori* expectations. Table 2.1 summarizes these studies,[1] demonstrating the wide variety of applications across countries, years and election backgrounds. While the 2BL test dominates earlier studies, more recent publications mostly apply the LD method to detect human vote count manipulation.

Mixed, and sometimes contradicting, results initiated a strong dispute over whether the 2BL test is a valid and adequate tool for detecting election fraud. Deckert, Myagkov and Ordeshook (2011, 260) conclude that it is not a "universally applicable magic box into which we plug election statistics and out of which comes and assessment of an election's legitimacy." Mebane (2011) contradicts their conclusion, which is, as he argues, based on inadequate simulation and insufficient data aggregation for the analysis, as well as ambiguous results. Shikano and Mack (2011) provide a technical explanation of why 2BL signals anomalies in the absence of fraud: the inflated frequency of specific digits can provide information about a specific range in which the vote count distribution has a high density. Therefore, some criticism of the 2BL test can be attributed to violations of its assumptions. While relevant concerns about the test have been raised, there are always counterarguments; thus its ability to detect election fraud has not been definitively proven or disproven.

[1] Table 2.1 contains only information concerning digit tests, and does not provide an exhaustive review of each study.

Table 2.1: Summary of digit test literature

Country	Election	Year	Expected fraud	Fraud action	Test	Indication	Author
Venezuela	Referendum	2004	-	-	2BL	Yes	Pericchi & Torres 2004
USA, Florida counties	Presidential	2004	No	-	2BL	A few	Mebane 2006a
Mexico	General	2006	Suspicion	-	2BL	Yes	
-	Simulation	-	Yes	Additions or subtractions of candidates vote counts	2BL	Yes	
USA, Florida counties	Presidential	2000	Other anomalies	Over and under votes	2BL	No	Mebane 2006b
USA, Florida counties	Presidential	2004	Some	-	2BL	No	
USA, Ohio	Presidential	2004	Controversial	-		Yes	
USA	Presidential	2000/2004	-	-	2BL	A few	
Iran	Presidential	2009	Suspicion	-	1BL	Yes	Roukema 2009
Iran	Presidential	2009	Some indications	Multiple	2BL	Yes for 3 of 4 candidates	Mebane 2009
Russia	Duma	2003/2007	Some/ many	Turnout inflation	Mean 2BL	More anomalies in 2007/2008 and beyond turnout inflation	Mebane & Kalinin 2009
Russia	Presidential	2004/2008					
Germany	Parliamentary	1990 - 2005	East and west differences	-	2BL	Yes (many)	Breunig & Goerres 2011
Germany	Parliamentary	2009	No	-	2BL	Yes	Shikano & Mack 2011
-	Simulation	-	Yes/no	-	2BL	Bad performance	Deckert et al. 2011
Ukraine	Presidential; round 2	2004	Yes	Falsification of votes			
Ukraine	Presidential; round 3	2004	No	-	2BL	Opposed expectations	
Ukraine	Presidential	2007	In some regions	Switching votes			
Russia	National	2004/2008	In some regions	Turnout inflation and vote switches			
Buenos Aires, Argentina	Provincial	1932	No	Switching votes	Machine learning 1BL	No	Cantu & Saiegh 2011
Buenos Aires, Argentina	Gubernatorial	1935	Suspicion			Yes	
Buenos Aires, Argentina	Legislative	1940	No			No	
Buenos Aires, Argentina	Provincial	1941	Yes			Yes	
Sweden	Parliamentary	2002	No	Made up vote counts	LD	No	Beber & Scacco 2012
Nigeria	Presidential	2003	Suspicion			Yes	
Senegal	Presidential	2003	Unlikely			No	
Senegal	Presidential	2007	Likely			Yes	
Afghanistan/ 3 provinces	Presidential	2009	Suspicion	Made up vote counts	LD	Yes	Weidmann & Callen 2013
Azerbaijan	Parliamentary	2008	Likely	Made up vote counts	LD	Yes	Sjoberg 2013
Switzerland	Referendum	2011	Uncertain	Made up vote counts	LD	Opposed expectation	Leemann & Bochsler 2014
Mexico	Gubernatorial	2010	Uncertain	"Fraude hormiga" (ant fraud)	2BL/LD	Differ between statistics	Cantu 2014

The LD test seems to address the biggest drawbacks of 2BL, which concern its validity and the missing mechanism linking detectable fraud and the underlying fraudulent activity. Nevertheless, some questions remain. Leemann and Bochsler (2014) find significant deviations from the LD digit test where they did not expect any, last digits were found to be uniformly distributed when deviations would have been plausible.[2] The result could be due to chance (5% error rate) or to the fact that the LD test has a high likelihood of type I error,[3] as identified for 2BL (Deckert, Myagkov and Ordeshook, 2011; Shikano and Mack, 2011). Concerning the LD test, the only validations are based on the investigation of the 2002 Swedish parliamentary election, a simulation (Beber and Scacco, 2012) and a recount of the votes in three districts of the Afghanistan election in 2009 (Weidmann and Callen, 2013). 2BL has been shown to vary in its applicability to elections across countries due to election characteristics other than fraud, which could also be relevant for the LD test.

In summary, the literature includes a number of different (and sometimes inconclusive) 2BL and LD results across countries. In most studies they are applied to a single country. Exceptions for 2BL applications are Mebane (2006*a*) and Deckert, Myagkov and Ordeshook (2011), who apply it to two countries each.[4] For the LD test, the Beber and Scacco (2012) study investigates the elections of three countries. Additionally, none of the studies discusses the linkage between 2BL and LD. Only one study applies both simultaneously, but with a very different purpose: to cross validate a new approach (Cantú, 2014). While fraud indications of digit tests are not only different than suggested by Cantú (2014), 2BL and LD differ in their indication of anomalies.

2.2.1 Mechanisms behind the 2BL and LD tests

Given the contradicting results in previous studies, and to establish a relationship between 2BL and LD, in this section we discuss the fraud mechanisms that 2BL and LD can capture.

The 2BL and LD tests share a common assumption: that certain fraud mechanisms will interrupt the "natural" digit-generating mechanism. One obvious manipulation type that should interrupt this process is a person or people replacing vote counts with numbers made up by humans. It is well established in the literature that humans are bad random number generators (Nickerson, 2002; Boland and Hutchinson, 2000; Rath, 1966). The evidence is mostly based on experimental studies showing that

[2] More specifically, Leemann and Bochsler (2014) investigate whether the last digit deviates from the probability given by BL for the last digit. However, if we follow Beber and Scacco (2012), this should not be significantly different from the uniform distribution and therefore is equivalent to the LD test.

[3] A result that indicates that fraud is present when it actually is not present.

[4] It could be argued that only Mebane (2006*a*) applies 2BL in a comparative setting, as Deckert, Myagkov and Ordeshook (2011) do not apply it to an adequate data aggregation level in the Ukraine.

humans who are specifically asked to produce random sequences fail to do so, even if they have some statistical knowledge. Studies investigating the human capability to produce random digit sequences have found that humans have particular digit preferences. For example, they mostly avoid 0 and prefer especially 1 (Rath, 1966; Boland and Hutchinson, 2000).

The LD test is based on the assumption that last digits and digit pairs distribute uniformly (Beber and Scacco, 2012). Beber and Scacco (2012, 215) show that it "require[s] very particular distributional assumptions in order for last digits not to be distributed uniformly." Thus, we need strong assumptions about fraudulent activities and mechanisms that cause digits to distribute differently. Humans' inability to produce random numbers is one such mechanism that disturbs the uniform distribution of digits. According to Beber and Scacco (2012), this is the *only* known mechanism to do so.[5] This means that significant deviations of the last digits from the uniform distribution can be attributed to human vote count manipulation.

This logic is equivalently applicable to the 2BL test. The difference from the LD test is that we observe the second digits of vote counts, and calculate their deviation from the second digit Benford's Law distribution. Most scholars have analyzed the distribution of the second digit instead of the first digit because of an inherent characteristic of many electoral systems that more or less constant precinct sizes produce a certain frequent digit in the first digit of vote counts in the absence of election fraud (compare Brady, 2005; Mebane, 2006*b*). Many scholars have empirically observed that digits of vote counts follow a Benford-like distribution (compare applications of Table 2.1). Mebane (2006*b*) also offers a theoretical argument why digits distribute according to Benford's Law. Compared to the LD test, the mechanism that disturbs Benford's Law is not limited to human vote count manipulation. Mebane (2006*b*) suggests that Benford's Law should be sensitive to types of manipulation that involve falsifying ballots: "On the whole the simulations suggest that the 2BL test can be highly sensitive to additions or subtractions from candidates' precinct-level vote totals" (Mebane, 2006*b*, 17). The findings also suggest that in tied (simulated) election outcomes, small changes (such as an increase of 2–3%) can trigger the test statistic, but 2BL will not indicate fraud if the manipulation is sufficiently small.

Based on the discussions above, we can derive expectations about the linkage of 2BL and LD. As argued above, the most obvious fraudulent activity that interrupts the digit-generating process is a person replacing vote counts with made-up numbers. Both digit tests should be sensitive to this possibility. This is the only situation in which LD might not be uniformly distributed. By contrast, Benford's Law has been shown to signal the addition or subtraction of numbers in finance data as well as in

[5] One other mechanism mentioned by the authors is the rounding of votes during counting, which might reflect laziness instead of an intention to manipulate. If rounding is likely, the authors advise against applying LD.

Table 2.2: Detectable manipulation by 2BL and LD

		2BL: yes	2BL: no
LD	yes	Vote counts generated by humans	-
	no	Additions or subtractions of candidates' vote counts (e.g., ballot box stuffing, multiple voting, falsifying ballots	Sufficiently small manipulation

elections (Nigrini, 2012; Mebane, 2006*b*). Therefore, 2BL should also detect additions or subtractions of candidates' vote counts (e.g., ballot box stuffing, multiple voting or falsifying ballots) to a certain extent. However, neither digit test can capture vote count manipulation if it is sufficiently small, nor can they detect manipulation that does not systematically add to or subtract from vote counts. The corresponding linkage between 2BL and LD is summarized in Table 2.2.

2.3 Data and methods

We analyze 2012 presidential election data from France, Finland and Russia. To evaluate the validity and the linkage of both digit tests, we selected these cases for multiple reasons. First and most importantly, France and Finland are assumed to be fraud free, while the Russian election was most likely manipulated. There have been no accusations of electoral fraud or anomalies in France for quite some time (compare Klimek et al., 2012; OSCE/ODIHR, 2012*a*). In the Finish presidential election, some anomalies were found in the correlation between overall turnout and True Finns party's vote share, which can be attributed to successful voter mobilization by the controversial party (Klimek et al., 2012). Apart from this naturally caused anomaly, the OSCE characterized the election as clean (OSCE/ODIHR, 2012*b*). By contrast, Russian elections do not appear to be controversial over the fact that there have been strong indications of fraud since 2003 (see Myagkov, Ordeshook and Shakin, 2009; Mebane and Kalinin, 2009; Deckert, Myagkov and Ordeshook, 2011; Kalinin and Mebane, 2011; Kobak, Shpilkin and Pshenichnikov, 2012; Klimek et al., 2012). The final election OSCE observer report assessed voting on election day as good and very good in 95% of the observed polling stations. However, due to procedural irregularities during the counting process, almost one-third of observed polling stations were assessed as bad or very bad. The irregularities cited included cases of

group, proxy and multiple voting; improperly sealed ballot boxes and indications of buses transporting groups of voters to vote at multiple polling stations. Concerns regarding the counting procedures were mainly listed as insufficient transparency during the counting process and some ballot box stuffing (OSCE/ODIHR, 2012*c*, 10f). In Russia it has been shown that ballot box stuffing changes the shape of the turnout distribution and generates a high correlation between vote share and turnout (Myagkov, Ordeshook and Shakin, 2009; Kobak, Shpilkin and Pshenichnikov, 2012; Klimek et al., 2012).

The second reason for the choice is that these countries have dispersed vote count distributions. It is known that narrow vote count distributions with limited range cause a strong deviation from the expected 2BL distribution (Shikano and Mack, 2011). Similar limitations have been addressed for LD distributions, as they "require that vote counts do not cluster within a very small range of numbers" (Beber and Scacco, 2012, 217). Electoral units in France, Finland and Russia vary more than, for example, in Germany, which should increase the fit with 2BL and LD in vote counts. Table 2.3 provides information about the main quantities of the data. It shows that mean (eligible) voters in France and Russia are similar, while those from Finland are slightly higher.[6] The variation in numbers of voters across districts is substantial for all three countries, but France has a particularly large range of voters. This is a potential drawback of the French election data, which are aggregated at the municipality level – the lowest level available in France. Vote counts of smaller municipalities are provided at the level of polling stations, while large municipalities (which contain different polling stations) report aggregated data. In general, this increases the scope of vote counts and therefore should contribute to the fit of 2BL and LD. To avoid inconsistency in our analysis, we include a control sample for the first and second rounds of the French election, excluding all municipalities that exceed the 90% quintile of voters. This excludes all aggregated data points and decreases the range and mean of the number of voters substantially;[7] 50% of the vote counts vary between about 100 and 500 voters, which reveals the limited scope of the control data and makes deviations more likely. This is mainly problematic if districts have very homogeneous distributions, which French elections are not particularly known for.

In the following analysis, we rely in principle on the most popular test statistic, Pearson's χ^2. For 2BL we calculate the test statistics as:

$$\chi^2_{2BL} = \sum_{i=0}^{9} \frac{(d_i - dq_i)^2}{dq_i}, \tag{2.1}$$

[6] In contrast to many electoral districts in many countries, those in Finland are not designed to contain a similar number of eligible voters.

[7] There is no other clear distinction possible between the aggregated data and vote counts from polling stations.

Table 2.3: Descriptive statistics of the presidential elections in 2012

Pres. election 2012	N obs.	N units	Mean unit size	Mean eligible voter	Mean voters	Range voters	25% of voters	75% of voters
France: first round	36,441	92	396	1,085	887	4 - 378,905	135	667
France: control	32,766	91	360	437	371	4 - 1,673	124	493
France: second round	36,441	92	396	1,085	891	6 - 382,975	136	664
France: control	32,767	91	360	437	371	6 - 1,669	125	493
Finland: first round	2,301	14	164	1,902	1,329	72 - 8,095	542	1,897
Finland: second round	2,302	14	164	1,901	1,257	37 - 8,231	497	1,824
Russia	95,566	84	1,138	1,152	752	0 - 19,711	247	1,195

where q_i denotes the expected relative frequency of i in the second digit, d_i is the empirical frequency of second digit i in a district and d is the sum of polling stations in the district. The statistic is assumed to be distributed according to a χ^2 distribution with 9 degrees of freedom. Therefore, we can evaluate the significance of the deviation of empirical data from Benford's Law using a critical value of 16.92 at a significance level of 5%. This method is also applicable when testing the deviation of the last digit; the only difference is that the frequency of the last digit is compared with the expected frequency of the uniform distribution.[8]

We conduct separate test statistics for each of a country's electoral districts, which makes it necessary to adjust the assessment of statistical significance for multiple testing. As suggested by Mebane (2006*b*), we use the false discovery rate (FDR) method (Benjamini and Yekutieli, 2005; Benjamini and Hochberg, 1995). When assuming independence across tests, then let $t = \{1, \ldots, T\}$ index the T-tested district and let S_t be the significant probability of the test statistic for each district. S_t is sorted from the smallest to the largest value, starting with S_1. The FDR is controlled using the following procedure: for a chosen test level α, let d be the smallest value such that $S_{(d+1)} > (d+1)\alpha/T$. d gives the number of tests rejected by the FDR criterion, which should be $d = 0$ if digits in vote counts distribute according to 2BL or LD.

Testing 2BL and LD with χ^2 means that the power of the tests depends on the sample size – i.e., the number of polling stations in a district. Therefore, we exclude all electoral districts that contain 50 or fewer polling stations. For the French election data, this reduces the number of districts from 107 to 92 (91 for the control sample) – and from 15 to 14 districts for Finland, and 85 to 84 for Russia. If we adjust the test statistic for each candidate based on the number of tests we compute in each election according to the FDR criteria (compare N units in Table 2.3), we get an adjusted χ^2 of 29.45 in France (29.42 for the control data), 24.50 in Finland and 29.22 in Russia.[9]

[8] This formula and detailed information are listed in Mebane (2008*b*, 179) and Shikano and Mack (2011), and on the last digit in Beber and Scacco (2012).

[9] Data from France and Finland are available at `http://www.interieur.gouv.fr/Elections` and `http://pxweb2.stat.fi/Database/statfin/vaa/pvaa/pvaa_2012/pvaa_2012_en.asp`. Rus-

2.4 Application of 2BL and LD to France, Finland and Russia

Table 2.4 presents the summary results of the 2BL and LD statistic as well as the FDR controlled statistic. Using a significance level of 5%, we compare 2BL and LD statistics with the critical value of 16.92. Since we assume a 5% error rate, the proportion of significant statistics relative to the total number of districts should not exceed 5% (if it does, this is emphasized in Table 2.4). The procedure is reasonable for 2BL and LD tests in France and Russia, which have a relatively large number of districts (92 and 84, respectively), but not for Finland, which has only 14 districts. We also compute the FDR-adjusted statistics (France 29.45 and 29.42 for the control data, Finland 24.50 and Russia 29.22), which solves the multiple testing problems by adjusting the expected proportion of type I error. If vote counts in France and Finland follow the 2BL and/or LD distribution, we do not expect any significant FDR-adjusted statistics.

Table 2.4: Summary of 2BL and LD statistics

		2BL	2BL with FDR	LD	LD with FDR
France					
First round	Hollande	0/92	0/92	4/92	0/92
	Sarkozy	2/92	0/92	**5/92**	0/92
Control	Hollande	2/91	0/91	4/91	0/91
	Sarkozy	2/91	0/91	2/91	0/91
Second round	Hollande	1/92	0/92	2/92	0/92
	Sarkozy	**6/92**	0/92	**5/92**	0/92
Control	Hollande	2/91	0/91	4/91	0/91
	Sarkozy	**7/91**	0/91	4/91	0/91
Finland					
First round	Niinistö	**1/14**	0/14	**1/14**	0/14
	Haavisto	**1/14**	**1/14**	**1/14**	0/14
	Väyrynen	0/14	0/14	0/14	0/14
Second round	Niinistö	**2/14**	0/14	0/14	0/14
	Haavisto	**1/14**	0/14	0/14	0/14
Russia					
	Putin	**16/84**	**3/84**	**5/84**	0/84
	Zyuganov	**13/84**	**1/84**	**15/84**	**6/84**

Note: indicates the number of significant statistics relative to the total number of electoral districts.

Both approaches (the proportion of significantly large statistics and FDR-adjusted

sian election data were made available by Klimek et al. (2012) at `http://www.complex-systems.meduniwien.ac.at/elections/election.html`.

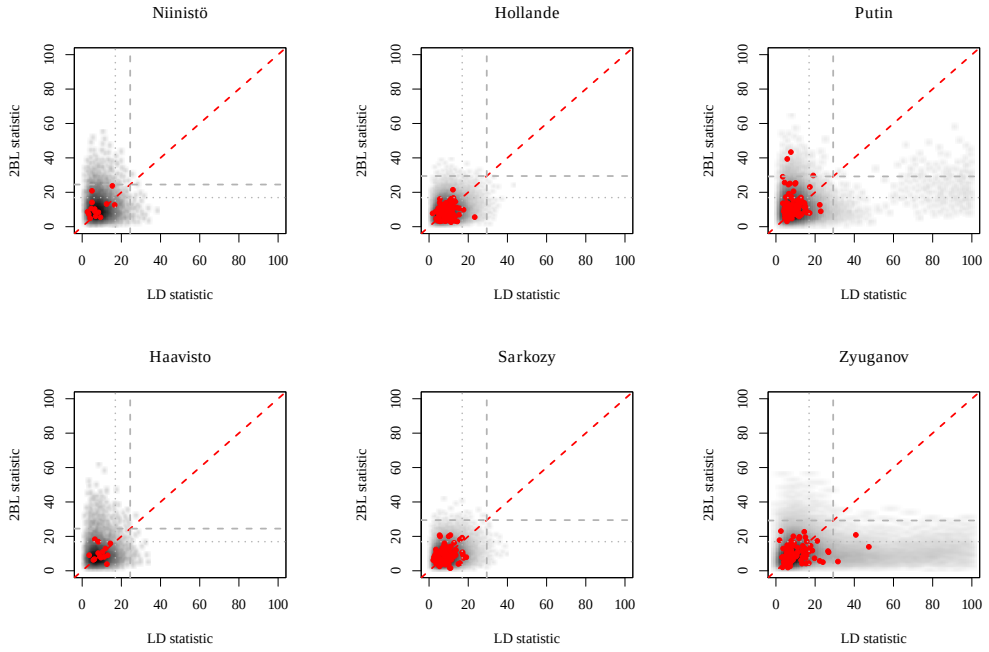
statistics) should be consistent in indicating that no fraudulent activity took place.[10] Table 2.4 shows that both approaches differ in their fraud signaling in some cases. For example, while more than 5% of the 2BL and LD statistics for Sarkozy (France, second-round vote counts) exceed the critical value (and therefore signal fraud or anomalies), the FDR-adjusted 2BL and LD statistics do not. There are indications of fraud or anomalies for Russia based on the proportion of significant 2BL and FDR-adjusted 2BL statistics. However, there is an exception for the LD statistic: there are more than 5% significant statistics for Putin, but the FDR-adjusted statistic indicates no fraudulent activity. In Finland, we find no indications of fraud or anomalies, with one exception – the FDR-adjusted 2BL statistic for Haavisto. In summary, we find few indications of fraud or anomalies in France and Finland, and some indications of fraud or anomalies in Russia. The tendency of the findings is in line with what is widely believed, but the findings are also diverse. Indications of fraud are not always consistent for the two approaches we use to deal with multiple testing: they sometimes differ across statistics (2BL and LD signal fraud in different districts) and across countries.

We now focus on the relationship between 2BL and LD. As shown in Table 2.2, there should be a specific link between 2BL and LD, if both techniques are able to detect human vote count manipulation. 2BL should also capture other kinds of fraud, which means that a high LD statistic should go along with a high 2BL statistic. Simultaneously, we would expect significant 2BL statistics to be at least as high as (or higher than) significant LD statistics. While Cantú (2014) did not assume any linkage between 2BL and LD, his application showed different fraud indications for all three statistics: his own , 2BL and LD.

We investigate the linkage between 2BL and LD not only in empirical terms, but also using simulated data, since the empirical election outcome can be viewed as one of many possible realizations. Using calibrated data can help expose patterns of fraud or anomalies. We use the multi-party election model by Katz and King (1999), which is based on the multivariate t-distribution. Following Shikano and Mack (2011), we estimate the multivariate distribution for each district based on empirical vote counts. Taking 1,000 random draws from the estimated distribution gives 1,000 different sets of election results for each district.

Figure 2.1 displays the bivariate relationship between the 2BL and LD statistics. According to our theoretical expectation (compare Table 2.2), we expect data points to be located above the red diagonal. This means that we expect significant 2BL statistics to be at least as high as (or higher than) significant LD statistics. This does not hold in France or Russia unless we further consider the FDR-adjusted critical value. The only statistics that exceed the adjusted significance levels are Putin's

[10] As mentioned, due to the small number of districts, we should only interpret the FDR-adjusted statistics in Finland.



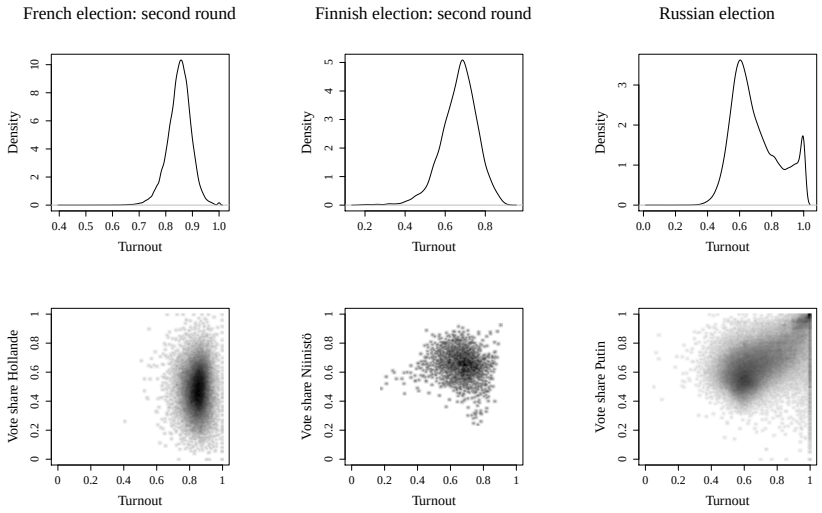
(a) (b) (c)

Figure 2.1: The graphics display the bivariate relationship between the 2BL and LD statistics for Finland (second round), France (second round) and Russia. The red dots mark empirical data points, and the grey shaded area denotes the simulated data. The red dashed line marks the diagonal, the light grey dotted line indicates the critical value of 5% significance and the dark grey dashed line the FDR-adjusted critical value.

2BL statistics, which are higher than the corresponding LD statistics.

2.5 Cross validation of the 2BL and LD methods

The results from the previous section more or less confirm our expectations of fraud-free elections in France and Finland and manipulation in Russia. However, we found exceptions in each country. In this section we cross validate the results with another indicator of election fraud: extreme election fraud (Myagkov, Ordeshook and Shakin, 2009; Kobak, Shpilkin and Pshenichnikov, 2012; Klimek et al., 2012). Accordingly, inflated turnout and a strong correlation between turnout and the vote share of a certain candidate should be interpreted as a sign of election fraud. Figure 2.2 displays the turnout distribution for France, Finland and Russia in the upper sub-graphs, while the lower sub-graphs illustrate turnout and the winning candidate's vote share. Here, Putin's very high vote share and high turnout can be interpreted as sign of extreme election fraud. According to Klimek et al. (2012), extreme fraud should be indicated in a peculiar shaped and bimodal turnout distribution and/or



(a) (b) (c)

Figure 2.2: The upper sub-graphs display the turnout distributions for France, Finland and Russia. The lower sub-graphs depict a scatter plot of turnout and vote share of the winning candidate. The intensity of the gray indicates the density of the data points.

extremely high turnout in combination with an extremely high vote share for the winning candidate. Such remarkable findings for Russia are displayed in Figure 2.2(c): the turnout rate has a bimodal distribution, where a mode is at 100%. In addition, the scatterplot of the turnout rate and vote share for Putin shows a concentration of units in the upper-right corner. Similar structures cannot be identified for the French or Finnish presidential elections in the same year.

While Klimek et al. (2012) quantify such extreme election fraud at the national level, its implementation at the district level is not feasible. Therefore, we compute an alternative approximation of extreme fraud at the district level that captures both aspects. Since an above-95% electoral turnout is widely considered to be an obvious indication of fraud (Myagkov, Ordeshook and Shakin, 2009; Klimek et al., 2012; Simpser, 2013), we compute the share of polling stations with a turnout over 95% relative to the total number of polling stations in a district. This share should indicate the extent of the extreme turnout fraud. Additionally, we compute the relative share of polling stations that have a correlation above 0.7 between the winning party and turnout as an indicator of high correlation.[11]

We display the results for Putin in Figure 2.3, scattering the 2BL statistic and

[11] One could argue that information from election observers should be used for cross validation. However, studies have found that the presence of election observers reduces election fraud (Hyde, 2007; Enikolopov et al., 2013). Further, since election observers are only sent to a sample of polling stations across the country, information would not be available for most polling stations across Russia; nor would data be available on the district-level intensity of fraud.

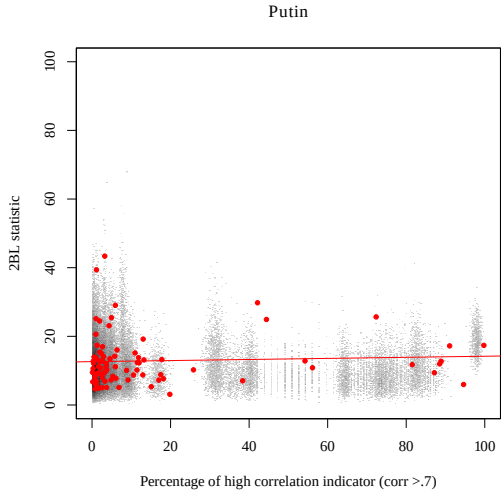


Figure 2.3: The red dots mark the bivariate relationship between 1) the relative share of the high correlation between vote share and turnout and 2) the 2BL statistic of empirical observation. The gray shades depict the relationship between both based on the simulated data (darker colors show higher density).

the relative share of polling stations with a high correlation between vote share and turnout. An increasing percentage of polling stations that correlates higher than 0.7 does not produce higher 2BL statistics. We also tested whether there was a linear relationship between the share of high turnout and 2BL, or between the share of high correlation/turnout and LD for Putin (and Zyuganov). In the visual illustration, we cannot identify any systematic pattern or significant correlation between 2BL statistics and high turnout/correlation. This is a striking result, since extreme fraud – likely caused by ballot box stuffing – cannot be detected by 2BL. The results for LD statistics and extreme fraud indicators are similar and non-informative. In contrast to 2BL, this result was expected, because this type of manipulation should not inflate the LD statistic.[12]

Based on these findings, we have to conclude that 2BL is incapable of detecting massive ballot box stuffing and miscounting the votes for other parties. If this is the case, the question remains what high 2BL statistics indicate in Russia. We analyze data from the highest 2BL statistic in Russia, which comes from unit 58 (Moscow region). If we assess the second significant digits of the simulated vote counts, we find strong deviations for digits 1, 2 and 3, and slightly weaker deviations for digits 0, 4 and 5 – as we would expect, according to 2BL. If we focus on the vote count distribution,

[12]The results for 2BL statistics and high turnout, as well as for LD and high turnout/correlation, are available upon request. For Zyuganov, since there are no high values of the correlation indicator, we cannot identify any pattern or significant correlation between his 2BL statistics and the indicator for high turnout. Further, we cannot conduct similar analyses for Finland or France, as neither indicator is relevant in these countries.

we find a high density for vote counts between 1,000 and 1,500. In this range, only an additional 100 votes changes the second digit. A distribution concentrated in such a critical range of numbers violates the assumption of vote count distribution for Benford's Law. This is an extention of the previous findings of Shikano and Mack (2011), which emphasize the importance of dispersion and the central tendency of the vote count distribution for the 2BL statistics.[13] We have to conclude that, even in countries with a more dispersed vote count distribution, such violations of the Benford's Law assumption can occur in specific situations. In the Russian case, it is not possible to tell whether the high density of vote counts in a specific range is caused "naturally" by voter preferences or constituency characteristics, or whether it is the result of election fraud.

Last but not least, we test the 2BL and LD statistics using four hypothetical fraud mechanisms that are plausible with respect to the Russian election. First, the manipulator knows the number of voters and pretends that all of the votes have been cast for a candidate. Second, all eligible voters cast votes for a candidate. Third, the manipulator is ordered that the vote share for a candidate should be between 85% and 100%. Fourth, a random number of polling stations is sampled, and the vote counts of the winning candidate are replaced by a) the number of eligible voters and b) voters of the manipulated polling stations. This corresponds to 100% turnout and vote counts for the former scenario, and to 100% vote counts for a candidate in the manipulated polling station in the latter. The procedures of the fourth mechanism were repeated randomly 100 times. We implemented all four hypothetical fraud mechanisms in the fourth district of Finland, France and Russia, which did not indicate electoral manipulation or anomalies according to 2BL and LD. Applying the first three manipulation mechanisms does not trigger the 2BL or LD statistics, with the exception of 2BL in Russia, when every vote was counted for Putin (Mechanism 1) and when Putin received between 85% and 100% of the votes (Mechanism 3). Results for the fourth mechanism are displayed in the Appendix in Figure A.2. We cannot find a linear relationship between the relative share of manipulated polling stations and the digit statistics. Hence, even if more polling stations in a district are massively manipulated, the 2BL and LD statistics do not increase significantly.

To sum up, 2BL and LD do not detect extreme vote count fraud. This was expected concerning LD, but it is an unsatisfying result with respect to 2BL. Further, it is possible that the highest 2BL statistic is the result of a violation of the 2BL assumption rather than an indication of electoral fraud. These findings are supported by the non-significant results of digit tests when they are applied to manipulated data. Digit tests did not signal anomalies after we implemented simple manipulations to an unsuspicious district of each country.

[13] The finding is illustrated in the appendix in Figure A.1.

2.6 Conclusion

This paper investigates the 2BL and LD tests by applying them to presidential election data in France, Finland and Russia. We use a comparative design, since a wide range of previous studies has revealed diverse (and sometimes contradicting) test results in their single-country analyses. The contradictions might be the result of country-specific characteristics. As expected, we find few indications of fraud or anomalies in France and Finland, and some indications of fraud or anomalies in Russia. However, there are exceptions, which are statistically unlikely to be due to chance. In particular, the discordance between the two approaches to dealing with multiple testing is remarkable. 2BL and LD statistics for Putin do not correlate, and significant FDR-adjusted 2BL statistics signal anomalies while the corresponding LD statistics do not. Given the fraud detected by 2BL and LD, this hints at some additional votes in favor of Putin, but does not indicate human vote count manipulation. To cross validate this result for the Russian election, we computed another indicator to capture extreme election fraud. We could not link significant deviations of 2BL to this indicator. These findings are further supported by the non-significant results of digit tests when they are applied to intentionally manipulated data. Based on our findings, we contradict the interpretation of Cantú (2014) that the different statistics detect different types of manipulation and therefore complement each other. Instead, we conclude that significant 2BL and LD statistics do not necessarily indicate manipulation at all.

We concede that evaluating the performance of LD within this framework is not straightforward. While there are indications in OSCE reports of people falsifying vote counts in Russia, they are rare compared to other fraudulent activities that are not detectable by LD. We do, however, find a few deviations in France and Finland that may indicate that the LD test also has a too-high type I error, which could falsely signal an effect. We need to quantify LD's type I error and investigate how strong human vote count manipulation has to be in order to be detectable if other fraudulent activities interfere.

3

Fraud Mechanisms and Types that Benford's Law can and cannot Detect: Defining its Sensitivity and Specificity

Verena Mack

Abstract

To detect election fraud, some techniques have been developed that are based only on vote counts. Of those, Benford's Law is one of the most popular techniques, but also the most controversial. This paper distinguishes between different levels of fraud coordination and actions that capture relevant fraudulent activities. Such mechanisms are calibrated with the 2008 Canadian election results that are considered to be fraud-free and then used to simulate artificially manipulated election data. This gives a unique setting to estimate its sensitivity for different fraud coordination, actions, intensity and spread, which has not been done yet. Moreover, a fraud-free and calibrated election model is applied to estimate the specificity of Benford's Law. The findings show that Benford's Law can detect about 70% of the manipulation correctly if the election was extremely systematic, many polling stations were affected and a lot of votes were manipulated. Almost all other manipulation settings reveal that the probability of correct detection is less than 50%. This shows that not even those fraudulent activities, which should be very well suited for 2BL detection, can be identified sufficiently precisely by the 2BL test.

3.1 Introduction

The legitimacy of elections has been and remains an important issue of politics. Scholars are challenged by the question whether elections can be trusted or the outcome is fraudulent. Election fraud is difficult to identify with certainty if the manipulation is not blatant. Detection methods require some prototype representing the non-manipulated election outcome in order to identify deviation from it. Therefore, establishing a good and valid model of non-manipulated election outcomes is essential. This usually requires additional information, for example socioeconomic factors, which are not always available in environments of electoral fraud. More recent approaches only require vote counts to investigate deviations of its digits or deviations of the vote/turnout distribution (Mebane, 2006*b*; Beber and Scacco, 2012; Klimek et al., 2012). Of those, the second digit Benford's Law (2BL) test is the most applied but also the most controversial approach. In contrast to the last digital test, 2BL is not assigned a certain fraudulent activity which is an important point of criticism. According to Mebane (2006*b*), systematic interventions (e.g. systematic applied fraudulent activities) should interfere with the digit-generating process and thus cause significant deviations from 2BL. Instead of investigating 2BLs capability to detect different fraudulent activities applied with varying systematic, recent applications combine 2BL with other fraud detection methods or even with a machine learning algorithm (compare Cantú and Saiegh, 2011; Mebane, 2015; Montgomery et al., 2015). These approaches combine different sets of information, but they do not increase our understanding about 2BLs capability. Instead, the accumulation of false or misleading information could lead to a wrong evaluation of an election. This paper investigates the sensitivity of the 2BL test (manipulations that are correctly identified as such) for different fraud mechanisms that impact the levels of the systematics as well as varying intensity and varying spread of election fraud. Moreover, it investigates the specificity of the 2BL test (fraud-free data that is correctly identified as such) in respect to non-manipulated election results. This improves our knowledge of what fraudulent activities 2BL can detect and how they would have to be used in an election to be traceable.

The fraud coordination and fraud actions describe different aspects of an election manipulation process. The first aspect captures fraud if it is coordinated, the level of coordination and how systematic it is carried out. The fraud action refers to a specific fraudulent activity such as ballot box stuffing. Different levels of coordination combined with manipulations that capture a range of fraudulent actions are calibrated with the 2008 Canadian election results and then used to simulate artificially manipulated election data. It is important to use calibrated instead of (pure) simulated data because election characteristics have been shown to influence the 2BL statistic (Shikano and Mack, 2011). Moreover, simulations are always based on specific distri-

butions that hardly resemble the underlying digit generating process of vote counts (see also the argument between Deckert, Myagkov and Ordeshook (2011) and Mebane (2011)). The calibrated data allow to define the sensitivity of 2BL for different fraud intensities, different levels of spread across polling stations and also if it matters if fraud is carried out systematically (and how systematic). The estimation of 2BLs specificity is based on a calibrated fraud-free election model. The findings limit the range of fraud types that are detectable by 2BL more than a priori expected. The main findings show that 2BL can detect about 70% of the manipulation correctly if the manipulation was done extremely systematic, many polling stations were affected and a lot of votes were manipulated. If the manipulation contained some kind of random component as it is the case in uncoordinated fraud and less systematic manipulation, then the probability of correct detection is less than 50%. This shows that not even those fraudulent activities, which should be very well suited for 2BL detection, can be identified sufficiently precisely by the 2BL test. Compared to 2BL's sensitivity, its specificity is relatively high, but it classifies more fraud-free election results as manipulated than the 5% error rate that is considered acceptable.

This paper is structured as follows: the next section discusses the advantages, drawbacks and controversies of different detection methods, followed by a section on fraud coordination, actions and their expected detectability by 2BL. The fourth section focuses on the empirical data and the test statistics. Subsequently, the calibration and simulation of the manipulated data is presented. The sixth section investigates the sensitivity and specificity of 2BL. Finally, the last section concludes and discusses the findings.

3.2 The detection of election fraud and Benford's Law

Detecting election fraud and quantifying election manipulation is challenging whenever fraudulent activities are not blatant. Many scholars address the issue with different statistical methods, but most of these approaches have a common basis: scholars estimate a model for fraud free election results, and then estimate the deviation from the model as an indication (and quantification) of fraudulent activities. Some of the forensic methods focus on spatial deviations of votes from polling stations, vote flows between candidates and elections, or they analyze the turnout rates.[1] Such studies require additional information as previous election results, socioeconomic information or specific knowledge about district characteristics. In an environment of manipulated elections, it can be difficult or even impossible to gather such information. Therefore, a small number of developed techniques require only the absolute number of the vote count. We can classify such forensic approaches by their focus on a) digits

[1]Some interesting approaches are those from Wand et al. (2001); Mebane and Sekhon (2004); Alvarez, Hall and Hyde (2008) and Myagkov, Ordeshook and Shakin (2009).

of vote counts and b) the distribution of vote counts. In the digit tests, deviations of a specific theoretical distribution in a specific digit of vote counts are searched. The statistics that are applied are the last digit test and Benford's Law. The second class contains applications that search for deviations and anomalies within the vote count distribution.

This paper focuses on Benford's Law, the most applied digit-based method. It has been found in the empirical observation that leading digits of numerical data are often not uniformly distributed (Newcomb, 1881; Benford, 1938). Data is Benford distributed "if probability distributions are selected at random and random samples are then taken from each of these distributions in any way so that the overall process is scale (or base) neutral" (Hill, 1995, 360). If digits of vote counts in normal or more specific in fraud free elections follow a Benford-like distribution, the deviation of such a distribution should indicate election fraud, as the digit generating random process is interrupted by a certain artificial non-random process. Many scholars showed that digits of vote count indeed seem to follow a Benford-like distribution. The investigated digit, and therefore the adequate probability distribution given by Benford's Law, varies across applications. Roukema (2009) for example estimates the deviation of the first digit of the presidential election in Iran from the first digit in Benford's Law distribution in order to detect election fraud. This can be problematic due to an inherent characteristic of many electoral systems that are designed to contain more or less the same number of eligible voters in each polling station. If turnout and voter preferences are similar across the polling stations of a district, then it can produce a frequent digit as a first digit that is not connected to election fraud (compare Brady, 2005; Mebane, 2006*b*). Consequently, most scholars analyze the distribution of the second digit instead of the first digit, which is known as the second digit Benford's Law (2BL) test.

Applications of 2BL are numerous and vary across countries, electoral systems, and they differ also in their expectations concerning fraud. 2BL has been applied to elections in the US, Mexico, Indonesia, Russia, Iran, Germany, Ukraine, Puerto Rico, Venezuela, France and Finland (Pericchi and Torres, 2004; Mebane, 2006*b*, 2007, 2008*a*; Mebane and Kalinin, 2009; Mebane, 2010*b*; Breunig and Goerres, 2011; Shikano and Mack, 2011; Deckert, Myagkov and Ordeshook, 2011; Pericchi and Torres, 2011; Cantú, 2014).[2] Further applications cover some variations in the investigated digit: Leemann and Bochsler (2014) investigate the third or fourth digit in a Swiss referendum, while Cantú and Saiegh (2011) search for deviations in the first digit in different Argentinian election using Benford's Law in combination with a machine learning algorithm. Next to these scientific applications, there is a great number of Internet blog posts applying and discussing Benford's Law. Thus, one can-

[2]Chapter 2 gives a compact and structured overview about Benford's Law applications.

not deny its impact within and outside of science. However, applications, findings and interpretations of Benford's Law as well as its validity are not free of controversies, which makes a deeper understanding of Benford's Law as an election fraud detection technique crucial.

Benford's Law tests, in general, are not bounded to one particular type of fraud, but supposedly can capture different fraudulent activities. One fraud type that Benford's Law can detect is human vote count manipulation. It is well established that humans are bad random number generators with preferences for certain digits. If the vote counts of an election result are replaced by humanly generated once, then the digits of such numbers should deviate from Benford's Law. Mack and Shikano (Chapter 2) establish a linkage between the Benford's Law test and the last digit test based on this mutually detectable fraud type.[3] This fraudulent activity is examined in detail by Mack (Chapter 4), since the last digit test should only identify this fraud action. Contrary to the last digit test, Benford's Law is known to detect a variety of fraudulent activities. The idea is borrowed from forensic digital analysis that is in use by "accounting and auditing firms, as well as governmental tax authorities worldwide, as routine check on data for fraud" (Kossovsky, 2015). Digital analysis with Benford's Law can, for example, signal systematic addition or subtraction of numbers in finance data (Nigrini, 2012). Similar characteristics are identified in the context of elections, since 2BL seems to be highly sensitive to additions or subtractions of vote totals. Mebane (2006*b*) shows that 2BL is especially sensitive in tied election outcomes, but the test cannot detect manipulation if it is sufficiently small. Mack and Shikano (Chapter 2) also examine 2BL's ability to display the systematic addition of votes using ballot box stuffing in the Russian presidential election 2012. Unfortunately, significant deviations from 2BL are not consistent with other fraud indicators that confirm such fraudulent activities. However, the suggested logic of 2BL fraud detection should work well for the massive fraud with such high level of systematics.

Deckert, Myagkov and Ordeshook (2011) use simulation data with some implemented fraud types to quantify the sensitivity and specificity of 2BL, and reveal its poor performance. Whereas, Mebane (2011) considers the finding to be the result of an inadequate simulation and therefore unsuitable. Simulation attempts to assess the performance of 2BL have one common problem: they never adequately represent the variety of features in an election district that lead to a specific election result. Such naturally caused characteristics of vote count distribution can cause deviations that have nothing to do with election fraud (Shikano and Mack, 2011). The specific

[3]The last digit test investigates the deviation of the last digits of vote counts from the uniform distribution. It requires particular distributional assumptions for last digits to not be distributed uniformly (Beber and Scacco, 2012). Humans replacing vote counts with made up numbers qualifies as such a particular mechanism.

characteristics of districts and elections in general, seem to be an important difference between fraud detection in accounting data and election results. Deckert, Myagkov and Ordeshook (2011) as well as Mebane (2011) give both relevant pro and contra arguments for 2BL as fraud detection technique. More recent findings support the contra side, concluding that 2BL cannot necessarily detect fraudulent activities like ballot box stuffing (compare Chapter 2). The authors cross-validate if anomalies (indicated by 2BL) correspond with anomalies indicated by other established fraud indicators, which is not the case. While arguments against 2BL arise, none of these studies can precisely define fraud mechanisms (coordination and actions) that 2BL can (and cannot) detect. Furthermore, there is still no adequate measurement of 2BLs sensitivity and specificity in the context of election fraud detection.

3.3 Fraud coordination and actions

Manipulation can be viewed as a two stage process: the first stage is the fraud coordination that defines the level of implemented fraud and whether it is carried out systematically. The second stage contains the specific fraudulent activity that is carried out. OSCE reports by election observers and election fraud literature almost always focus on fraudulent activities, but mostly ignore the fraud coordination. This might be due to the obscure situation in that election fraud takes place, and coordination is therefore often difficult to detect with some certainty (compare Alvarez, Hall and Hyde, 2008). Fraud coordination seems to be clearer in cases of blatant fraud where it is observable across a country or districts. This is the case in Russia where vote counts of Putin, the certain winner, were boosted in some republics by local elites (Myagkov, Ordeshook and Shaikin, 2008; Myagkov, Ordeshook and Shakin, 2009; Simpser, 2013). Beyond such blatant cases, systematic coordination of fraud and its level of implementation are often implicitly assumed. There is a strong belief among scholars that manipulators will use fraudulent activities to reach a small winning margin, because anything exceeding victory would increase the cost of the manipulator unnecessarily (Alvarez and Katz, 2008; Bailey, 2008; Nyblade and Reed, 2008). The assumption is connected to a popular idea that the only reason to manipulate an election is to win it (Alvarez, Hall and Hyde, 2008). This implicitly assumes certain coordination across districts among a first past the post electoral system. Otherwise small winning margins are difficult to reach and winning is left to chance. Scholars further suggested that coordination is most cost-effective if the number of manipulated votes, necessary to swing the election, is small. In both cases information about the unit of interest is required and manipulation has to be coordinated across this unit in order to achieve the optimal outcome for the manipulator.[4]

[4]Simpser (2013) proofs with a formal model that there is an equilibrium if only winning matters (i.e. fraud does not have indirect effects beyond winning) and manipulation is costly.

Manipulation can also occur at polling stations more or less independent from each other. However, such a mechanism has an important drawback compared to coordination; the winner of an election remains uncertain even when there are fraudulent activities. Winning depends on the tightness of the race, how many polling stations are manipulated in a district as well as the number of votes that are manipulated in each polling station. Most scholars have neglected this manipulation mechanism as it is not cost effective. However, if possible detection bears higher cost than the benefits of a certain win, then coordination becomes unattractive. Moreover, in non-blatant cases and also in more democratic settings, institutional obstacles might make coordination of fraud extremely difficult. The absence of coordination across districts decreases detectability of the manipulation and therefore makes local and independent fraudulent activities more attractive.

Within different levels of fraud coordination, different types of fraudulent activities can be applied. In Russia it has been noted that substantial fraud occurred through ballot box stuffing and tampered election protocols (Myagkov, Ordeshook and Shaikin, 2008, 198). OSCE election observer reports name a wide variety of observed fraud types that also fit the list of common tactics by Simpser (2013, 35):

- stuffing ballot boxes (or destroying ballots)
- falsifying results or otherwise tampering with the vote count
- tampering with voter registration lists
- vote buying before/during the election
- creating obstacles to voter/candidate registration
- intimidating voters before/during the election
- intimidating candidates
- voting multiple times
- voting by those who are ineligible, such as minors

All fraudulent activities aim to increase the vote share of the manipulating party. The manipulation either affects the nonvoters and therefore the turnout of an election or, if a manipulator tampers with vote counts of the other parties, it shifts votes in favor of the manipulating party. In most cases, a variety of fraud types that affect both turnout and the vote counts of parties are used for manipulation (Simpser, 2013). One feature of fraudulent activities crucial for fraud detection with Benford's Law is the addition of votes to the manipulating party. Fraud types such as abstention buying that are discussed by some scholars (Gans-Morse, Mazzuca and Nichter, 2014; Cox and Kousser, 1981), increase the vote share of the manipulating party through

motivating people not to vote for an opposing party. This decreases turnout and increases the vote share as intended, but the actual vote count of the manipulating party does not change and therefore cannot be detected by 2BL. It is known that systematic tampering of data can be identified by Benford's Law (Kossovsky, 2015). Therefore, all fraud types that change the vote count should be detectable as long as the change is systematic. How systematic the manipulation has to be in order to be detectable is part of this investigation. Moreover, it reveals important and specific information about detectability of different fraud mechanisms.

3.4 Data and test statistic

The paper uses the 2008 Canadian election results, to implement manipulation artificially for the following reasons: First, to the best knowledge, there have not been any allegations of fraud during or after the election. Second, seats for parliament are contested in single-member districts making the manipulation of the districts the most effective manipulation that can be implemented. Such a setting enables a detached investigation of each district from the overall national vote share. This is because a party wins through winning most districts rather than winning the majority of votes across the country.

In October 2008 the Canadian federal election was carried out to elect the 40th Canadian Parliament. The winner of the election was the Conservative Party of Canada, winning 143 out of 308 districts. With elections in single-member districts this translates into 143 seats, thus, they missed the majority of 155 seats by 12 seats and had to rule as minority government. The Liberal Party won the second most seats as they won 77 districts, the Bloc Quebecois achieved 49 seats, the New Democratic Party 37 seats and also two independent candidates were elected to Parliament. If the winning margin is less than 0.1% of the votes casted, a judicial recount is automatically ordered according to the Canada Election Act. Furthermore, electors can request such a recount if the winning margin exceeds 0.1% but is still narrow. Small winning margins caused five judicial recounts which confirmed four times the election night result and on one occasion gave the seat to the Liberal candidates instead of the Bloc Quebecois. There have been no allegation of electoral manipulation in the 2008 election, neither in the official election report nor in the media.

The test statistic that is mostly applied is based on the Pearson's χ^2 statistic:

$$\chi^2_{2BL} = \sum_{i=0}^{9} \frac{(d_i - dq_i)^2}{dq_i}, \tag{3.1}$$

where q_i denotes the expected relative frequency of i at the second digit, d_i is the empirical frequency of second digit i in a district and d is the sum of polling sta-

tions in the district (compare Mebane, 2008*b*, 179). The statistic is assumed to be distributed according to a χ^2 distribution with 9 degrees of freedom. Therefore, we can evaluate significance of deviation of empirical data from Benford's Law using a critical value of 16.92 at a significance level of 5%.

3.5 Artificially manipulated data

In an optimal scenario, the manipulation of vote counts should be conducted where it is most effective and it is feasible to change the outcome of the election. Thus, it is most effective to manipulate those districts in which the candidate of the Conservative Party lost with less than 5% difference in the vote share compared to the winning party. If at least 12 out of 15 manipulated districts are turned in favor of the Conservative Party, the party would reach the absolute majority in the Canadian parliamentary election 2008. Moreover, the Conservative Party is the incumbent party and incumbents are the ones fraud is often attributed to as they have better resources and administrative access to fraud (compare findings of Alvarez, Hall and Hyde, 2008; Myagkov, Ordeshook and Shakin, 2009; Simpser, 2013). Such an initial situation is an adequate setting to implement artificial manipulation and it also corresponds to the most recent Canadian election of 2011. The incumbent Conservative Party won the majority of seats with 166 out of 308 seats, however, there have been allegations of election fraud against the Conservative Party in the media.[5]

3.5.1 Calibrating different fraud mechanisms and simulating manipulated data

The manipulation strategy for the 15 districts follows the logic of the outlined fraud coordination and actions. The first mechanism is fraud coordination across a district, which seems to be unspoken but common wisdom among scholars. In order to coordinate, it requires information about the (likely) election result of a district. This is important to calculate the number of votes necessary to swing the election outcome. Winning tightly is one possible outcome, but, as Simpser (2013) shows, there are plausible reasons why winning with high winning margin can be reasonable. Therefore, different fraud intensities are considered through increased winning margins across the simulations. Furthermore, it is relevant how many polling stations in a district are manipulated. The number of manipulated polling stations is labeled as the fraud spreads.

The second mechanism is uncoordinated and independent fraudulent activities

[5]To exemplify, here are some information `http://www.canada.com/news/Coyne+Robocon+scandal+with+clear+pattern/6230302/story.html`; `http://en.wikipedia.org/wiki/2011_Canadian_federal_election_voter_suppression_scandal`

Table 3.1: Summary of digit test literature

Manipulation	**Winning**	**Variance of added votes**	**Fraud types**	**Impact on**	**Simulated mechanism**	**Expected detectability by the literature**	
						2BL	**LD**
No	No	-	-	-	-	No	No
Coordinated manipulation with constant addition of votes	Certain	No	Ballot box stuffing	Nonvoters, CP	CC nonvoters	Yes	No
		No	Tempering with registration lists	Nonvoters, CP	CC nonvoters	Yes	No
		No	Voting multiple times	Nonvoters, CP	CC nonvoters	Yes	No
		No	Tempering with vote counts	Potentially all parties	CC mixed	Yes	No
			Manipulating all BUT CP	All, except CP	No	No	No
Coordinated manipulation with varying addition of votes	Certain	Some	Ballot box stuffing	Nonvoters, CP	CV nonvoters	Yes	No
		Some	Tempering with registration lists	Nonvoters, CP	CV nonvoters	Yes	No
		Some	Voting multiple times	Nonvoters, CP	CV nonvoters	Yes	No
		Some	Tempering with vote counts	Potentially all parties	CV mixed	Yes	No
			Manipulating all BUT CP	All, except CP	No	No	No
			Human vote count manipulation	Potentially all parties	No	Yes	Yes
			Vote buying	Individual voters	Likely Not	No	No
	Difficult to control		Intimidation	Individual voters	Likely Not	No	No
			Obstacles for registration	Individual voters	Likely Not	No	No
			Voting by those who are ineligible	Individual voters	Likely Not	No	No
Uncoordinated manipulation	**Uncertein,** likelihood of winning increases with fraud intensity	High	Ballot box stuffing	Nonvoters, CP	U nonvoters	Maybe	No
		High	Tempering with registration Lists	Nonvoters, CP	U nonvoters	Maybe	No
		High	Voting multiple times	Nonvoters, CP	U nonvoters	Maybe	No
		High	Tempering with vote counts	Potentially all parties	U mixed	Maybe	No
			Manipulating all BUT CP	All, except CP	No	No	No
			Human vote count manipulation	Potentially all parties	No	Maybe	Maybe
			Vote buying	Individual voters	Likely Not	No	No
			Intimidation	Individual voters	Likely Not	No	No
			Obstacles for registration	Individual voters	Likely Not	No	No
			Voting by those who are ineligible	Individual voters	Likely Not	No	No

across a district. This has been neglected by scholars due to uncertainty in winning and because it is not considered to be a cost effective manipulation effort. In this case, no detailed information is required at the district level about the election outcome, as local manipulators act independently. Like in the previous coordination mechanism, each polling station has a certain probability to be manipulated. The probability increases with the spread of fraud. Additionally, the intensity of fraud varies across specified fraud intervals. In the first interval, little manipulation is investigated, then the variation between little and stronger fraud is increased across two intervals and for the last two intervals, strong and very strong fraud, are investigated.

Independent from the coordination, a variety of fraudulent activities can be applied to increase the vote share of the manipulating party. Such activities either reduce nonvoters and add them to the manipulating party, subtract vote counts of other parties and add to the manipulating party, or a combination of both. Each activity increases the vote share of the manipulating party. Changing the number of nonvoters through, for example, ballot box stuffing is one of the most prominent fraudulent activities, but there are many other possibilities to temper with nonvoters and vote counts of other parties. OECD election observers and scholars state that almost always many different fraudulent activities are applied. Therefore, this study first investigates fraudulent activities that affect only nonvoters and secondly the changes in the number of nonvoters and other parties to capture a mix of different fraudulent activities. In the framework of this study and as previously discussed, this comprises only fraudulent activities that add votes to the manipulating party. It also excludes activities that decrease the turnout as it effects vote counts of other parties but not that of the manipulating party.

Table 3.1 presents the linkage of coordination mechanism, fraudulent activities and the simulated mechanism (which is explained below). The first row shows the baseline of no manipulation. The two dark gray blocks correspond to fraud coordination across a district (first mechanism) and the light gray block to uncoordinated and independent fraudulent activities (second mechanism). The column *Winning* indicates if the strategy gives a certain win to the manipulator. *Impact on* lists which vote counts are effected by the particular *Fraud types*. It is emphasized in section 3.3 that fraud types should be detectable by 2BL as long as they are systematic. The higher the *Variance of the added votes* is, the more difficult detection by 2BL should be. Within coordinated manipulation it is further distinguished between the extreme systematic manipulation (the exact same number of votes are added to manipulated polling stations) and a bit less systematic manipulation (a similar but slightly varying number of votes is added to manipulated polling station). The column *Simulated mechanism* gives the corresponding mechanism of the artificially manipulated data. The last two columns indicated whether, according to the literature, the particular

combination of fraud type and coordination level should be detectable by 2BL and LD test.

To simulate the manipulated election data, the following protocol is applied to the Election Day[6] vote counts of the Conservative Party in each district.

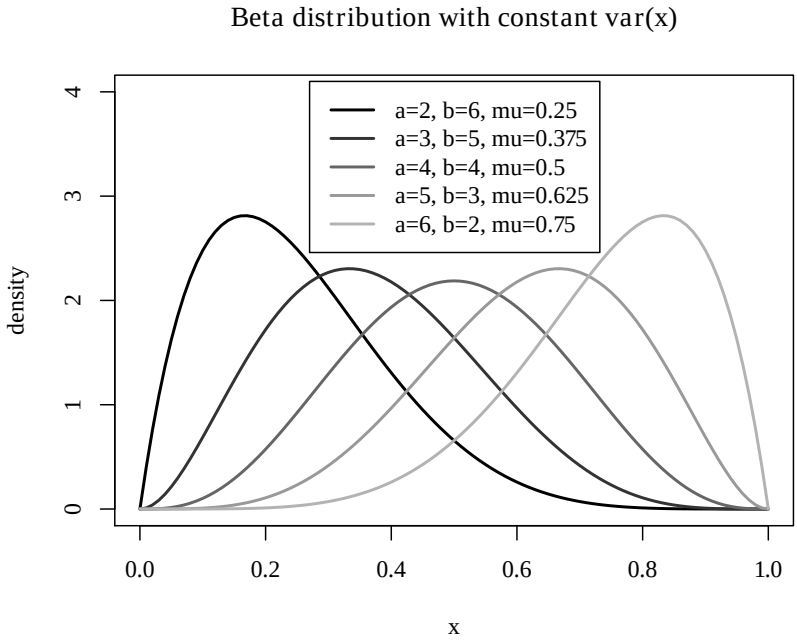


Figure 3.1: Keeping the product of a and b constant when a is increasing and b is decreasing gives constant variance across the distributions. Furthermore, it increases the expected value and therefore the number of manipulated districts.

1. Coordinated fraud at district level

   (a) The fraud intensity is defined. Calculate the number of votes that are necessary to win. This is specified through the parameter win that gives different winning margins: $win \in [1\%, 5\%, 10\%, 20\%, 35\%]$.

   (b) The spread of fraud is defined: Draw from the beta-binomial distribution to get the number of districts M that are manipulated. Keep the variance of $beta(a, b)$ constant. Therefore, keep the product of a and b constant, while parameter a increases and b decreases. This increases the expected value and thus the probability that a district is selected through a Bernoulli draw (compare Figure 3.1).

   (c) the votes that are manipulated according to (a) are distributed among the manipulated districts given in (b) by to the following rule:

[6]The fraudulent activities target the Election Day vote counts. It excludes absentee voting and polling stations with special voting rules.

i. **Coordinated and constant manipulation of nonvoters (CC nonvoters)**: Manipulate nonvoters equally across fraud districts. This means, shift all additional votes from nonvoters to the Conservative Party. The shifted votes ($votes_c$) is the number of votes calculated in (a) divided by M (calculated in (b)).

ii. **Coordinated and constant manipulation for mixed fraud types (CC mixed)**: Manipulate nonvoters and other parties equally across fraud districts. The number of shifted votes is again $votes_c$, but now subtract them from all other parties and nonvoters. For simplicity, $votes_c$ are always taken to 2/3 from all other parties and to 1/3 from nonvoters.

iii. **Coordinated and varying manipulation of nonvoters (CV nonvoters)**: Manipulate the nonvoters in fraud district m. Keep the mean across all manipulated districts M similar to $votes_c$, but slightly vary the added votes to the Conservative Party in district m. Therefore, draw the additional votes as follows: $votes_v \in N(votes_c, \frac{votes_c}{3})$.

iv. **Coordinated and varying manipulation for mixed fraud types (CV mixed)**: Manipulate all other parties and nonvoters in district m. As before, calculate $votes_v$ and add it to the votes of the Conservative Party in district m. For simplicity $votes_v$ are always taken to 2/3 from all other parties and to 1/3 from nonvoters.

(d) Repeat steps (b) and (c) 1000 times.

2. Uncoordinated fraud at polling stations within districts

(a) The spread of fraud is defined: Draw from the beta-binomial distribution to get the number of districts M that are manipulated. Keep the variance of $beta(a, b)$ constant. Therefore, keep the product of a and b constant, while parameter a increases and b decreases. This increases the expected value and thus the probability that a district is selected through a Bernoulli draw (compare Figure 3.1).

(b) The fraud intensity is defined: Draw the percentage p_m that is manipulated in district m with $p_m \in U(min, max)$. The parameter for min and max change according to the following intervals [1%, 10%]; [1%, 50%]; [1%, 100%]; [20%, 80%]; [80%, 100%].

i. **Uncoordinated manipulation of nonvoters (U nonvoters)**: Calculate the percentage p_m of only nonvoters in district m and add the resulting votes to those of the Conservative Party in district m.

ii. **Uncoordinated manipulation for mixed fraud types (U mixed)**: Calculate the percentage p_m of nonvoters and other parties in district

m and add the resulting votes to those of the Conservative Party in district m.

(c) Repeat steps (b) and (c) 1000 times.

Therefore, varying fraud intensity and spread are investigated for coordinated and uncoordinated fraud. Within coordinated fraud, it is differentiated between a constant and varying number of votes that are changed in favor of the Conservative Party. The scenario to add a constant number of votes to each manipulated polling station might seem to be stylized, but it corresponds to a possible manipulation strategy of voting machines, studied by Di Franco et al. (2004). The authors show that adding a small number of votes, for example two votes to each machine (implemented through a change in the program), can change the election outcome in the USA. The differentiation is further helpful as the paper investigates if more systematic manipulation affects the probability of detection by 2BL. To add a constant or slightly varying number(s) across polling stations changes how systematic the election is manipulated. If the manipulation is uncoordinated, then transferred votes vary across polling stations without coordination of fraud intensity and spread (adding a constant number of votes would be unreasonable). As the manipulation lacks any systematic, it should decrease its detectability.

3.5.2 Manipulated election results

In the scenarios of coordinated manipulation the fraud intensity and spread vary, but winning itself is not at stake. In contrast, if there is no coordination, then it is uncertain if the fraudulent activities lead to winning. The percentage of wins under each parameter combination is displayed in the appendix.

The simulated results are exemplary displayed in Figure 3.2. The rows show results for different mechanisms while the columns give the fraud intensity. For reasons of space, the spread of fraud is kept constant to $beta(4, 4)$.[7] The simulated results of CC and CV nonvoters and also CC and CV mixed barely differ. This means, it is visibly not differentiable between more systematic manipulation and (a little) less systematic manipulation. Logically, the manipulation of nonvoters increases the vote share by the same proportion as it increases turnout. The empirical data clusters as expected and the manipulated data (grey shades) blur to the upper right regions of the plots. The strength of the blur depends on the fraud parameters. Strong blur patterns (also visible in Figure 3.2) were identified by Klimek et al. (2012, 2) as sign of extreme election fraud. Since these patterns are present in the manipulated data, it is concluded that the simulations capture relevant manipulation mechanisms.[8] The

[7] Figures are available on request for different parameter combination of *beta*.

[8] Klimek et al. (2012) quantify the blur, or as they call it "smear out", as a measure of fraud

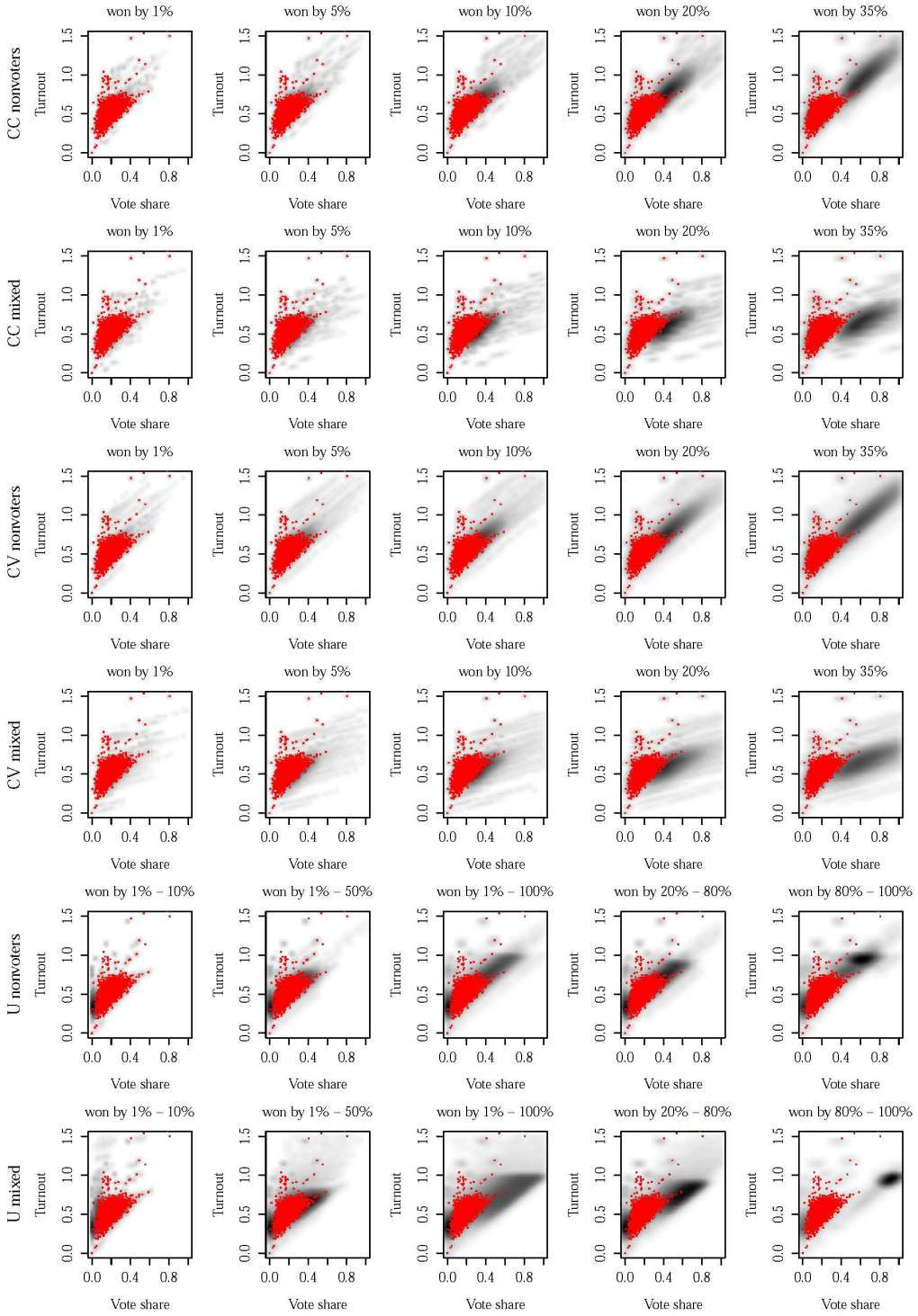


Figure 3.2: Vote share and turnout for different fraud intensities across the 15 districts. The fraud spread is fixed to $beta(4, 4)$ which results in about 50% manipulated polling stations. The red dots mark the empirical observations. The gray shade displays the manipulated vote share and turnout of the calibration. Darker colors indicate higher density of data points.

blur effect increases and gets stronger with increasing fraud intensity across the different versions of coordinated fraud.[9] The spread of fraud effects the shape of the blur pattern. More precisely, less manipulated polling stations at a given fraud intensity caused more bimodality in the data. The patterns for uncoordinated fraud (last two lines in Figure 3.2) are relatively similar to those of coordinated fraud. However, they differ to some extent in the strength of the bimodal pattern. They are stronger if the fraud intensity is relatively high. Interestingly, the uncoordinated fraud in combination with high intensity matches the patterns for Russia and Uganda identified by Klimek et al. (2012). This could indicate that the fraudulent activities in Russia and Uganda were essentially uncoordinated, but carried out with very high intensity.

3.6 Detectability of election fraud

3.6.1 2BLs sensitivity

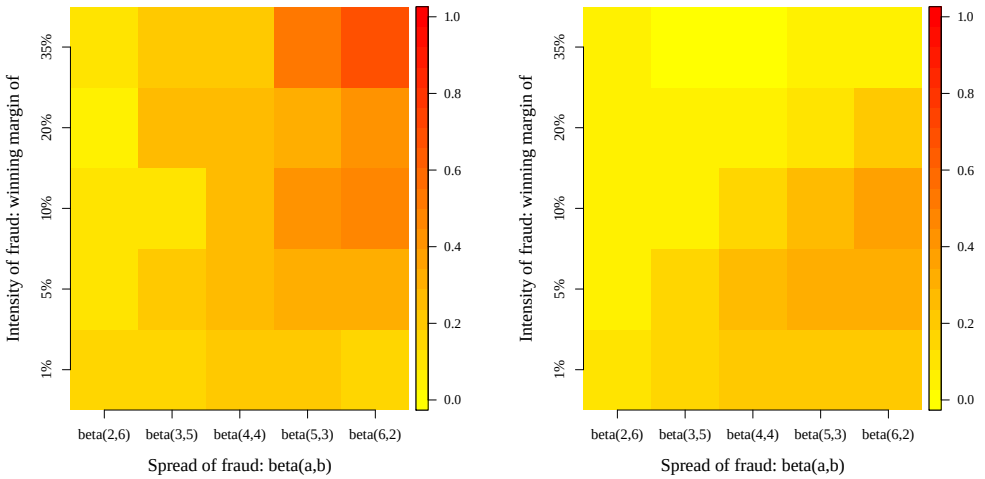
Figure 3.3 displays the proportion of correctly identified manipulated election data. It is displayed across districts and simulations.[10] Results for coordinated manipulation are displayed in Figure 3.3(a) and 3.3(b). While 3.3(a) shows the proportion of significant statistics for the more systematic coordination, 3.3(b) does the same for slightly less systematic coordination. If the exact same number of votes is added to each manipulated polling station, then increasing spread of fraud and increasing fraud intensity increases the overall sensitivity of the 2BL test. If we pay attention to the proportion of correctly identified manipulation, then it becomes obvious that the 2BL test does not perform (much) better than chance, with the exception of the most extreme and systematic manipulations. The 2BL test could detect about 70% correctly when the winning margin is 35% and fraud spread is high ($\mu = 75$).[11] The sensitivity of 2BL is worse if coordinated fraud is less systematic. Figure 3.3(b)

at the national level. To identify how much of the pattern is visible at the national level, one simulation of each manipulated district is integrated in the federal election result of 2008, repeated for all parameter combinations. In Figure B.3 (appendix) it is shown that the fraud intensity has to be strong, otherwise fraud patterns are not visible.

[9] For coordinated fraud, there is no restriction to 100% turnout. This is why turnout could exceed the natural boundaries, which is especially the case if nonvoters are manipulated and a high winning margin is requested. For uncoordinated fraud mechanisms, the manipulation is restricted to 100% of the eligible voters. Therefore, turnout does not exceed 100%. There are a few empirical data points (red dots) that approach or exceed 100% turnout. Those belong to polling stations with special voting rules and were excluded from the simulation (not manipulated) and again included afterwards.

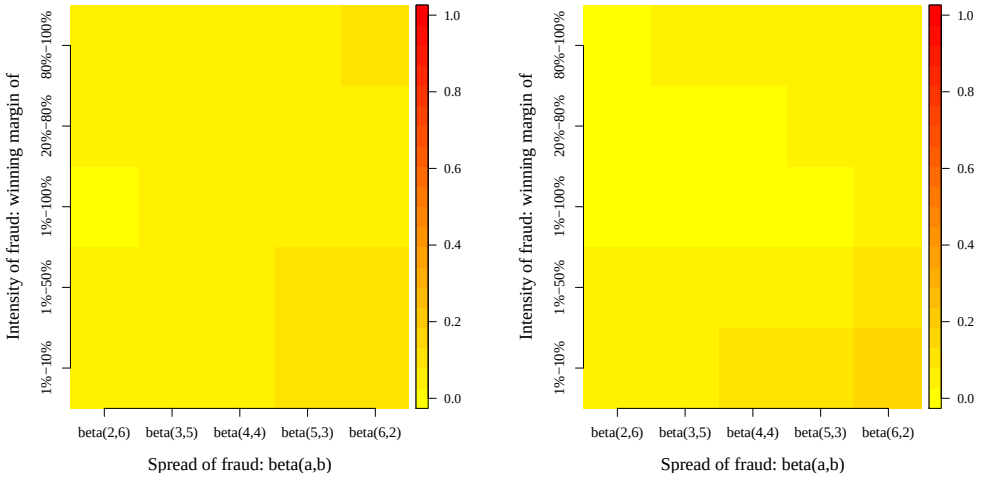
[10] The statistic signals fraud or anomalies if it exceeds the χ^2 statistic of 16.92. The procedure is applied to all parameter combination of fraud, each containing 15000 statistics (1000 simulations for each parameter, repeated for the 15 districts.

[11] This pattern differs in some districts strongly. The two most extreme examples are displayed in the appendix. The first example detects high proportion of manipulation correctly and the second identifies only about 10% of the manipulation.



(a) Coordinated and constant manipulation

(b) Coordinated and varying manipulation



(c) Uncoordinated manipulation: nonvoters

(d) Uncoordinated manipulation: mixed

Figure 3.3: The proportion of χ^2 statistics exceeding the critical value of 16.92 at a significance level of 5% across the 15 districts and 1000 simulations.

shows a slightly better sensitivity with increased spread, but overall 2BL inaccurately detects less manipulation such as one could when flipping a coin. While visually it was barely differentiable between very systematic coordination and coordination that is less systematic it makes a difference concerning 2BLs sensitivity. If the manipulation is uncoordinated, then 2BL did detect at most 20% of the manipulation as such correctly. Thus, 2BL cannot detect manipulation at all if it is not systematic.[12]

3.6.2 2BLs specificity

For a comprehensive assessment of 2BLs performance, its specificity has to be quantified as well. This is feasible if the actual election outcome is assumed to be one out of many possible outcomes. For this purpose, the paper calibrates the multi-party election model (Katz and King, 1999) with the 2008 Canadian election. First, the multivariate distribution for each district is estimated via the empirical vote count distribution. Then, election results for polling stations of each district are randomly drawn from this distribution. This is done as follows: The proportion of votes for party j in polling station k is given by V_{jk}. $V_{jk} \in [0,1]$ for all j in k, so that the proportion of votes for all parties add up to the sum of one.[13] The votes for the last party J depend on the votes of the other parties:

$$V_{Jk} = 1 - \sum_{j=1}^{J-1} V_{jk}. \tag{3.2}$$

Let Y_k be the vector of $J-1$ log-ratios with $Y_{jk} = \ln(\frac{V_{jk}}{V_{Jk}})$, for party j ($j = 1, ..., J-1$) relative to party J. Y_j is assumed to be multivariate Student t distributed and denoted as $Y_{jk} \sim f_{mt}(\mu, \Sigma, \upsilon)$, with the expected values of μ_{jk}, the variance of $\Sigma\upsilon/(\upsilon - 2)$, and $\upsilon(\upsilon > 0)$ is the degrees of freedom parameter. The paper uses the Bayesian method with non-informative priors to estimate these parameters for each district of the 2008 Canadian election. Afterwards, the estimated parameters are used to draw 1000 Y_{jk} in each district and to calculate the simulated vote percentage with:

$$V_{jk} = \frac{exp(Y_{jk})}{1 + \sum_{j=1}^{J-1} exp(Y_{jk})}. \tag{3.3}$$

If V_{jk} is multiplied by the number of polling stations, it gives the simulated vote

[12] For comparison, the last digit test detects less than 10% of the manipulated data correctly. The different specifications of the manipulations do not make a difference for detectability. This is relevant whenever fraudulent activities are not bounded to human vote count manipulation and different fraud types might be used. The results for all last digit applications are available upon request.

[13] The notation follows that of Shikano and Mack (2011) but does not lose equivalence to Katz and King (1999).

counts for the fraud-free election results and thus, the necessary information to conduct the specificity of 2BL.

Applying 2BL to each simulated district election result gives the proportion of statistics that are incorrectly classified as manipulated. Across all 308 districts and simulations, the 2BL test identifies 33% of the fraud-free vote counts of the Conservative Party as manipulated, hence its specificity is 67%. If, for comparability, we restrict the sample to the 15 districts that were artificially manipulated, then the proportion of false positives reduces to 13%, hence the specificity of 2BL is 87%.[14] Based on the significance level, no more than 5% deviation is expected, however, that is by far exceeded. Therefore, the specificity of 2BL has to be classified as insufficient.

3.7 Conclusion

The focus of this article is on 2BL, as it is a widely used method to detect election fraud, but 2BL is also rarely thought through and investigated in terms of assumptions, applicability and implications. First of all the assumption is investigated that systematic changes in numbers (vote counts) should lead to significant deviations from 2BL. To investigate this aspect, a distinction is made between coordinated and uncoordinated election fraud. This distinction is usually implicitly, but almost never directly discussed by articles on fraud detection, and thus the effects of systematics in fraud are ignored. Both fraud mechanisms are combined with the theoretically conceivable types of fraudulent activities in order to determine the theoretical detectability of the fraudulent activities. Ultimately there are many reasons why election fraud is carried out at different intensities and at varying spread across polling stations, which again can influence the detectability, but again it is rarely mentioned by 2BL applications.

In order to take a closer look at these different aspects, different fraud strategies are developed and calibrated by a supposedly fraud-free election, in this case the 2008 Canadian election. In the first strategy, different fraudulent activities are coordinated across a district. In the second strategy, fraudulent activities are carried out independently from each other. The simulations, based on the calibrated fraud strategies, give artificially manipulated election data to that full knowledge about the quantities of the manipulation exists. Therefore, it is possible to assess the sensitivity of 2BL concerning the intensity of fraud, its spread across polling stations, and if systematic manipulation is easier to detect. To define the specificity of 2BL, a calibrated multi-party election model is used to simulate a set of fraud-free election data.

The paper finds the correct fraud detection using 2BL depends on how system-

[14] For comparison, the proportion that are incorrectly classified as manipulated by the last digit test is 5.3% across all districts and 4.6% for the 15 districts.

atically the manipulation is carried out. Uncoordinated fraudulent activities across a district make the manipulation undetectable using 2BL. In this case, one could improve the probability of correct detection by flipping a coin. If there is coordination across a district and therefore manipulation is to some extent systematic (about the same number of votes are added to the Conservative Party in manipulated polling stations), then increased spread of manipulation across the district increased detectability. This finding should be treated with caution as it increases 2BLs probability to detect fraud correctly, but it is, in the best scenario for 2BL, around 50%. If manipulation is carried out extremely systematic (the same number of votes are added to the Conservative Party in manipulated polling stations), then 2BL improves the probability of correct detection compared to a coin. Again, it is important that many polling stations are manipulated, but also that many votes are manipulated. In this scenario, 2BL can detect about 70% of the manipulation correctly. Therefore, the overall sensitivity of 2BL to detect fraudulent activities is not satisfying. In fact, one is often better off to take a guess or flip a coin. Quantifying the specificity of 2BL supports the conclusion to not trust it as technique for detecting election fraud.

The findings limit the fraudulent activities that are detectable by 2BL as follows: Order the manipulation tactics according to how systematic it can be implemented across a district. If the tactic targets individual people it usually contains random components and therefore is not detectable by 2BL. This excludes any version of vote buying, intimidation, obstacles for registration or voting by those who are ineligible. If ballot box stuffing, tempering with vote counts (or registration lists) and voting multiple times are carried out *very* systematic *and* in many polling stations of a district, then we have about a 70% chance to detect the manipulation correctly. The corresponding situation is only conceivable if e. g. the program of voting machines is systematically manipulated and the same number is added to the votes in (almost) all polling stations of a district. If such obvious manipulations are used, it is very likely that it does not require 2BL for detection, especially if almost 30% of erroneously displaying election fraud is found.

The study calibrated specific fraud mechanisms by Canadian election results. Thus, findings are not necessarily generalizable as they are not exhaustive concerning the investigated countries and the fraud mechanisms. Moreover, the most promising finding that very systematic manipulation is often detectable and varies considerably between districts as shown in the appendix. This supports the conclusion to distrust 2BLs fraud detection in different settings.

The results are worrying considering the recent applications of 2BL. For example, Montgomery et al. (2015) integrate 2BL in the framework of Bayesian additive regression trees, while it is questionable if 2BL can contribute anything to fraud detection other than noise. Others like Mebane (2015) establish a forensic toolbox that combines all kinds of fraud detection techniques to accumulate fraud indications. In

general, this idea is helpful as different indicators should capture different fraudulent activities. However, if the different indicators are not reliable fraud detection techniques it accumulates wrong indications of fraud, which can lead to false accusations. The use of 2Bl is therefore to be evaluated critically both in the form of the Toolbox and in combination with machine learning. Research should focus on examining each of the combined fraud detection techniques in detail and evaluate their contribution to actual fraud detection.

4

Election Fraud, Digit Tests and How Humans Fabricate Vote Counts

Verena Mack

Abstract

The last digit test is a notable method to detect election fraud. It is based on a strong distributional assumption that a manipulator replaces the vote counts of an election result sheet with man-made numbers, but will fail to make the numbers look random. Allegations of election fraud are based on this mechanism, however, the assumed mechanism might be too simple to capture the strategic behavior of humans while manipulating election results. This could result in the following difficulties: Someone could manipulate only a certain number of vote counts. Other manipulation strategies are used instead to replace counted votes with man-made numbers, or man-made numbers are generated differently in this context than in the numerous experiments in which people deliberately try to generate random numbers. This paper uses a laboratory experiment to investigate how humans might manipulate the vote counts of election results, something that has never been done before. It shows how strategically the vote count manipulation is handled. The strategic behavior affects the number of manipulated vote counts and which manipulation strategies subjects use. Furthermore, it shows that even with man-made numbers, the last digit test does not function as sufficiently as fraud indicator.

4.1 Introduction

Statistical detection of election fraud emerged as a challenging task, thereby the last digit (LD) test became a notable detection method (Cantú, 2014). The detected fraud is attributed to a specific mechanism: human vote count manipulation. Last digits electoral returns (vote counts) distribute uniformly and it needs very particular distributional assumptions so that the last digits distribute uneven. Humans that replace vote counts with made up number in combination with their incapability to fabricate random numbers fulfill the criterion (Beber and Scacco, 2012; Weidmann and Callen, 2013). The test is applied to elections were election officials are suspected to falsify election result sheets by hand. This is suspected of occurring when election officials have unsupervised access to the results and/or the aggregation process is opaque. Such fraudulent activities are mostly present in developing countries were the election process is more fragile. One prominent example is the 2007 election in Nigeria were it is claimed that barley any polling stations received result sheets (The Economist, 2007).

The paper uses a laboratory experiment to investigate how people manually manipulate the outcome. This gives valuable insights about how an individual(s) might conduct election fraud. The detailed knowledge about the applied manipulation strategies can also help to develop actions to prevent fraud. Moreover, it investigates the fundamental assumption of the digit test. It is assumed that humans replace vote counts by made up numbers, however, it neglects the strategic behavior of humans. A manipulator could use a dice as random generator, change individual digits of the actual vote count or switch the vote counts of candidates. All of these strategies would undermine the assumed process that causes deviations in the LD.

The experiment implements different manipulation aims and settings to cover a wide range of frameworks of election manipulation. This is important as manual manipulation is not observable and there is little knowledge about manipulation strategies. The analysis of the experimental data investigates the different strategies that are applied and if subjects adapt the manipulation strategy based on the manipulation frameworks. Additionally the analysis captures how and if particular strategies affect the last digits that are produced through manipulation by hand. Therefore, the experiment accounts for the possibility that strategies and the specific context of an election could influence which numbers and hence digits are fabricated by humans.

To this end, subjects have to manipulate fraud-free election results from the 2008 Canadian federal election. Subjects get monetary incentives to manipulate inconspicuously in favor of a specific party. Incentives to manipulate inconspicuously are implemented through evaluation of other subjects. The experiment is comprised of four sessions that correspond to different treatments. Treatment 1, Treatment 2

and Treatment 3 (hereafter T1, T2 and T3) investigate the manipulation process for different intensities of manipulation. This accounts for the possibility that a manipulator adjusts their strategies depending on how much they have to manipulate or similarly by how much they want their preferred party to win. It corresponds to the findings by Simpser (2013) that many elections, especially in authoritarian regimes are won by large winning margins to send signals of power. The fourth session is conducted as a Setup Treatment (hereafter SuT) that asks the question if manipulators use different strategies when they fill out an empty sheet, or if they manipulate within the election return sheet. In the 2007 Nigerian election, for example, it appears the election officials filled out empty sheets. For SuT the level of manipulation is equivalent to T1.

The manipulation by subjects leave "fingerprints" that are often not obvious to the eye, but which are traceable in the data. Subjects have the tendency to manipulate as few polling stations as necessary to reach their manipulation aim. Therefore, the extent of manipulation increases when the intensity of fraud increases. This changes when the circumstances of the manipulation changes. Smaller but more frequent manipulations are conducted when subjects fill out empty election sheets. Subjects do not only manipulate the whole number, but also individual digits. Especially when the intensity of fraud is extremely high, they prefer to manipulate one of the leading digits instead of the whole number. Moreover, subjects use strategies such as vote swapping, but mostly restrain from increasing turnout. The extent as well as the manipulation strategy affect detectability by the LD test.

The paper is structured as follows: the next section discusses the LD test, humans' capability to produce random numbers and strategic aspects of the manipulation process. It is followed by the third section that presents the experimental design. The fourth section presents expected manipulation behavior and the fifth section the results. The final section discusses and concludes the paper.

4.2 The LD test and how humans manipulate electoral returns

Digit tests are one specific category of "election forensics". Such tests focus on the distributions of digits in the electoral vote counts. It is differentiated between the Benford's Law test (or some version of it) (Mebane, 2006*b*) and the LD test (Beber and Scacco, 2012). While the Benford's Law test has been heavily criticized (Deckert, Myagkov and Ordeshook, 2011; Shikano and Mack, 2011)[1], the LD test remains mostly unchallenged and is therefore by some scholars considered to be the better alternative.

[1] Compare also Chapter 2 and 3

The last digit test investigates the deviation of the last digits of vote counts from the uniform distribution. It "require[s] very particular distributional assumptions in order for last digits not to be distributed uniformly" (Beber and Scacco, 2012). Thus, strong assumptions about fraud actions and mechanism which cause digits to distribute differently are necessary. Scholars identified a possible fraudulent activity where manipulators replaced the vote counts with made up numbers that fulfilled the criteria. It is a well-established argument in the literature that humans are bad random number generators and therefore cannot produce uniformly distributed digits (Nickerson, 2002; Boland and Hutchinson, 2000; Rath, 1966). Therefore, a person that manipulates election return sheets by hand through the replacement of vote counts with made up numbers should produce LD that deviate from the uniform distribution.

Beber and Scacco (2012) apply the LD test to Sweden's 2002 parliamentary election, Senegal's 2000 and 2007 presidential elections and Nigeria's 2003 presidential election. Supporting their expectation, they find evidence for human vote count manipulation in Senegal 2007 and Nigeria 2003. Weidmann and Callen (2013) validate the LD test with the application to the challenged election results of 2009 in Afghanistan and its recount. They investigate the performance of the LD test in three provinces of the presidential election where human vote count fabrication is reasonable. Based on fraud suspicion the election results of those provinces had to be recounted and the LD test is able to distinguish between both counts. Two studies further use the LD test as an additional information source for fraud detection. Sjoberg (2014) uses the LD test to specifically identify vote count fraud, distinguishing between monitored and non-monitored polling stations in Azerbaijan. The results suggest that manipulators shift for example from ballot box stuffing to vote count manipulation in monitored polling stations (significant LD statistic), while they use different manipulation mechanisms in those that are non-monitored. Leemann and Bochsler (2014) investigate a Swiss referendum using different techniques to detect different fraud actions. Within this framework the LD test is applied to detect vote count manipulation, however, the results contradict their expectations.[2]

All of these LD test applications emphasize the importance of the manipulation mechanism for this fraud detection. "[V]oting results in manipulated elections are made up by people attempting to make these numbers look random. In generating these seemingly random numbers, however, people follow certain patterns which deviate from true randomness and, therefore, can be detected by statistical tests" (Weidmann and Callen, 2013, 63f). Thus, the human incapability to produce ran-

[2]To be precise, the authors investigate whether the last digit deviates from the probability given by BL for the last digit. For the Swiss referendum this is usually the third or fourth digit. However, if we follow Beber and Scacco (2012) this should not be significantly different from the uniform distribution and therefore is equivalent to the LD test.

dom numbers when manipulating vote counts is crucial when using the LD test for fraud detection. As this is the key assumption for LD test, it is surprising that scholars did not assess the assumption very carefully.

A closer inspection reveals two conditions that have to be fulfilled in order for the LD test to detect human vote count manipulation. The first condition is that a person replacing vote counts with fake numbers has to try to make up numbers that look random (man-made numbers). The second condition is that this person has to replace the vote counts of almost all polling stations in an investigated district with such man-made numbers in order that LD test can capture the manipulation. This means, the test can only capture human inability to produce random numbers if such numbers strongly dominate the investigated data.[3]

For now, lets assume the first condition is fulfilled (vote counts are replaced by man-made numbers) and lets address the second condition: almost all vote counts are manipulated. Why should one manipulate almost all, some or only a few vote counts? Manipulators could perceive very different manipulation approaches as inconspicuous. On the one hand, a manipulator could choose to manipulate a few polling stations that are effective in changing the election outcome. Such a manipulator might perceive fewer interventions to election results as less suspicious and maybe even as more (time) efficient. On the other hand, a manipulator could think it is inconspicuous to manipulate many or most polling stations just a little bit. In the first scenario the original vote counts will dominate the data, while in the second scenario the man-made numbers will dominate the data. As the LD test can only detect the incapability of humans to produce random numbers if such numbers dominate the investigated data, it is likely that LD test can produce adequate results in the second scenario, but not the first. This argumentation shows that the manipulation approach affects the extent of manipulation which is again crucial for detectability by the LD test.

It is not possible to observe the manipulation process of humans in elections. Replacing vote counts by man-made numbers seems to be a simple and straight forward way of manipulation and is therefore considered to be an adequate assumption. However, this assumption, above referred to as the first condition, can be challenged as well. First of all, assuming that vote counts are simply replaced by man-made numbers neglects the strategic behavior that is usually attributed to humans. There are many examples in economic theories and studies of political behavior that show how strategic humans act in elections. The most obvious example might be strategic voting. People's strategic behavior, however, is also discussed in electoral manipulation, such as in articles on vote buying (Gans-Morse, Mazzuca and Nichter, 2014; Hidalgo and Nichter, 2015) or articles discussing incentives of strong electoral manip-

[3]How the extent of manipulation across polling stations affects the detectability by Benford's Law is tackled in Chapter 3.

ulation (Simpser, 2013). How manipulation is conducted can be affected by strategic considerations of the manipulator. For example, the specific aim of manipulation, like inconspicuous manipulation, just winning, reaching a specific winning margin or even sending a "message" to a particular group through manipulation (Simpser, 2013). Thus, a manipulator that is missing many votes might only manipulate front digits while a manipulator that focuses on inconspicuous manipulation might prefer to replace vote counts by man-made numbers or even swap vote counts between parties.

If the behavior of the manipulator deviates from the assumed manipulation strategy, (i. e. vote counts are not replaced by man-made numbers but in another way), then the assumption that LD is not uniformly distributed is violated. Some conceivable manipulations that violate the assumption are exchanges of votes between two parties (vote swapping), exchange of individual digits with the last digit unchanged, systematically adding (or subtracting) numbers or generating numbers by throwing dice.

However, there are also considerations that can challenge the first condition. Given a manipulator choses a random number generation as a manipulation strategy, then the number generating process could still be restricted by certain limitations. For example, the eligible voters are usually a given limitation and can be tricky to manipulate. Moreover, in cases were manipulation is not supposed to be obvious, the number of voters should be smaller than the number of eligible voters and additionally correspond to the sum of vote counts of different parties in each polling station. Those are the most basic considerations, but they emphasize just how complex the number generating process and the manipulation in general can be. It questions if number generation, in the context of election manipulation, is comparable to those studies that showed that humans are bad random number generators when they are specifically asked to produce random numbers (or a sequence of random numbers) (Nickerson, 2002; Boland and Hutchinson, 2000; Rath, 1966). One aspect that could support this argument is based on findings of Beber and Scacco (2012) and Weidmann and Callen (2013). The digit distributions that are identified as critical by both studies do not match the distributions showing humans are bad random number generators. While 0 and 5 are by far the most favored digits in Weidmann and Callen (2013) and Beber and Scacco (2012), 0 is always the least favored digit if humans are asked to generate random numbers. The second least favored digit is 5 which is also very prominent in the applications. According to the experiments, favorable digits are 1, 2 and 3 (Boland and Hutchinson, 2000; Rath, 1966). If the process of random number generation in election manipulation reflects the general process of humans generating random numbers, it should also reflect their preferences and aversion of specific digits. This discrepancy as well as the mentioned strategic considerations are neglected by Beber and Scacco (2012), Weidmann and Callen (2013) and the

follow-up applications.

A similar argument has been raised by Medzihorsky (2015), who also questions the strong distributional assumption, and considers that different priorities and constraints of fraudsters could violate it. Instead of null hypothesis testing as it is done using the χ^2 statistic, his solution is to use a latent class analysis and relax the distributional assumption within this framework. Investigating how humans actually manipulate election results by hand goes a step further than Medzihorsky (2015) and investigates the distributional assumption instead of relaxing it. Such an investigation requires an experimental approach in which the subjects manipulate election return sheets.[4]

4.3 Experimental design

In the last section it was shown that three questions have to be answered for convincing election fraud detection using the LD test. Firstly, how many of the polling station results are actually falsified when manipulators try to manipulate inconspicuously? Secondly, what manipulation strategies are actually used in the case of human vote count fraud? Third, how good (or bad) are people at producing random numbers when they falsify election results? The experimental setup integrates possible influencing factors on the chosen manipulation strategy and amount of manipulated polling stations results under controlled conditions, while the quality of man-made number can be only assessed empirically. A distinction is made between two manipulation settings, (i. e. whether an election result is manipulated within a list or an empty list is filled with new election results). Further, different levels of fraud intensity are investigated. This section only presents the experimental design and the following section presents the expected manipulation behavior for the different treatments.

Participants of the experiment are instructed to manipulate three election sheets and afterwards evaluate the originality of three other election sheets. The participants are randomly assigned to two groups that received different sets of election results to ensure that all displayed data is unfamiliar to participants. Each set is divided into three sheets and each sheet contains vote counts of 30 polling stations that are randomly sampled from the Canadian federal election in 2008, an election which was very likely not manipulated.

The participants are instructed to manipulate the sheets inconspicuously so that party B, the runner up party, wins the overall result of the sheet with a given winning margin.[5] The treatment is the pre-defined winning margin (range) that differs for each session while all other experimental factors are kept constant. This is to in-

[4] Unfortunately, the latent class analysis cannot be applied in the experimental context, which is explained in more detail in the appendix.

[5] Instructions are available upon request.

vestigate different fraud intensities that can also resemble aims of manipulation like "just" winning (small winning margins) or sending a signal of power (large winning margins). In the first treatment (T1), participants had to reach a winning margin between 0.1% and 5%, which is considered to be marginal fraud. In the second treatment (T2), the range of the winning margin of party B increased to the range of 10% and 20%, reflecting considerable fraud. In the third treatment (T3), extreme fraud is imposed by a winning margin of party B between 40% and 50%. In these three treatments and for each sheet, participants have the option to choose if they fill out an empty sheet or if they load the election results and make as many changes as they wish within the sheet. An additional experimental session, the setup treatment (SuT), eliminates this option. SuT is equivalent to T1 in terms of the winning margin, but participants have to fill out the empty table by hand. The detailed procedure and graphical examples are presented in the appendix. Participants did not receive instructions or limitations concerning the amount of polling stations they manipulated. This is because manipulators that have the possibility and capacity to conduct such manipulation are unlikely to be limited in the extent of the manipulation. In the example of Nigeria, manipulators could fill out a complete election return sheet with new numbers if they wished. Participants had to reach the given winning margin for party B before they could continue with the next step of the experiment. This implies that the intensity of fraud is fixed for each treatment.

The primary goal of the evaluation is to incentivize participants to manipulate inconspicuously. In combination with the chosen payment this is very important to ensure the quality of the experimental data and therefore the outcome of this study. Both aspects have been given very careful consideration and are explained in detail in the appendix.

The experiment uses election data from the Canadian federal election in 2008. To the best knowledge, there are no indications of election fraud or anomalies for this election. Prior to the experiment, it was ensured that there are no significant last digit deviations across districts and individual sheets. To guarantee neutral framing, all party names were replaced with the labels party A, party B etc. The two different data sets (used for the manipulation task) are samples of 90 polling stations that are further split into three sheets each containing 30 polling stations, of the districts Guelph (number 35027) and New Westminster-Coquitlam (number 59017). Moreover, the districts were randomly selected out of all districts that party B lost by less than 5%.[6]

The experiment was conducted in the LakeLab of the Thurgau Institute of Economics (TWI) at the University of Konstanz. It was programmed with z-Tree (Fischbacher, 2007) and used ORSEE (Greiner, 2004) for recruiting. Participants are

[6]It was further ensured that within each sheet (30 polling stations), the differences between party A and party B was less than 5%.

undergraduates from various fields between the age of 18 and 27. In May 2015, 4 experimental sessions were conducted, each containing 26 and in total 104 participants.[7] An experimental session lasted approximately 90 minutes and participants earned 17.5 € on average.

4.4 Expected manipulative behavior

As explained, it is not sufficient assuming manipulators simply replace vote counts by man-made numbers which causes LD to deviate from the normal distribution. Instead, human vote count manipulation is likely to comprise several choices as the manipulation strategy (how to manipulate the vote counts) and the extent of manipulation (the vote counts of how many polling stations are manipulated).

Next to the obvious option to replace vote counts by man-made number, there are many different strategies. A manipulator can replace the original vote count with a new number or individual digit(s). The LD test can only capture modifications of the last digit, therefore the share of changed LDs of party B's vote count is considered. A manipulator can also decide to manipulate the first or second digit instead of the LD, which is captured in the share of first or second digit changes.[8] As the underlying assumption of the LD test looks for manipulators to replace vote counts with man-made number, another strategy captures random number generation. This strategy is an approximation of whether an individual digit or the whole number is replaced by a participant. More specifically, participants are assumed to have generated a random number if they replaced the first and the last digit of a vote count.[9] Subjects could also decide to swap votes between the original winner (party A) and the manipulating party (party B). If this strategy is used efficiently, then a few swaps in polling stations with the highest difference between both parties can change the election outcome. In the experimental context, seven of such swaps can be sufficient to reach a winning margin between 0.1% and 5%, required in T1 and SuT. Last but not least, subjects can increase the vote share of party B if they exclusively increase party B's vote

[7] Prior to the experiment one pretest with 8 participants was conducted that is not included in the data.

[8] Digits were assigned fixed positions to be able to keep track if the length of the number changed. Vote counts for party B had a maximal length of 3 digits and never less than 2. If the data originally contained a two digit, and after manipulation a three digit number (or the other way around), it is considered a change in the first digit.

[9] There is no comprehensive identification of a random numbers, but capturing a change in the first and last digit seems the best approximation for the following reasons: If someone generates random numbers it could be constituted out of any possible digits, however, the original digit has to be neglected to differentiate between new numbers and individual digits that are replaced. This should underestimate the numbers that are generated randomly. Moreover, most vote counts consist of two or three digits. The first and last digits provide the frame of the number, and it is therefore plausible that their modification indicates a new number. Furthermore, one can reduce the effect of underestimating random numbers and further manage increases from two to three digits more efficiently if changes of the first and last digit are considered.

counts. This strategy has the same effect as ballot box stuffing and will increase the turnout of the election result. Therefore, turnout is expected to increase across T1, T2 and T3, along with higher intensity of fraud. The SuT should contain equivalent or more turnout than T1 as both have the same winning margin, but more changes are expected in the SuT.

A manipulator also decides whether to manipulate vote counts of one, a few or many polling stations. Therefore, in the experimental setting, participants did not receive an instruction or limitation concerning the amount of polling stations they manipulated. However, as pointed out, they could perceive distinctive manipulation approaches as inconspicuous, which affects the number of polling stations they manipulate. Two hypotheses can be formulated concerning the extent of manipulation.

A manipulator can choose to manipulate fewer polling stations that are effective in changing the election outcome. Such a manipulator likely perceives fewer interventions of election results as less suspicious and maybe as more (time) efficient.

Hypothesis A: A manipulator manipulates as many votes as necessary, but as few vote counts of different polling stations as possible.

In contrast, a manipulator can also choose to make very little changes to most vote counts of a district. Such a manipulator likely perceives manipulation as inconspicuous if it is just a little bit everywhere.

Hypothesis B: A manipulator manipulates as many votes as necessary and distributes them across most vote counts of a district.

The intensity of manipulation is predefined for each session through the winning margin that participants have to reach. They could only continue with the next step of the experiment if they fulfilled the criterion. In the experimental context, replacing a few numbers or digits of a sheet is an efficient strategy. Therefore, it is expected that participants who have the possibility to load election results will consider hypothesis A as inconspicuous manipulation and manipulate the loaded sheets according to it. If they act according to this logic it means that the additional or switched votes have to be distributed across more polling stations if the winning margin increases. The expectations are summarized in Figure 4.1. Therefore, participants are expected to opt for manipulating the sheet that contains the results and conduct the most changes in T3, followed by T2 and the fewest in T1.

Contrary to T1, T2 and T3, participants that have to fill out an empty sheet (SuT) are expected to perceive the manipulation approach of hypothesis B as inconspicuous. This is because they have to write down a number for each party and polling

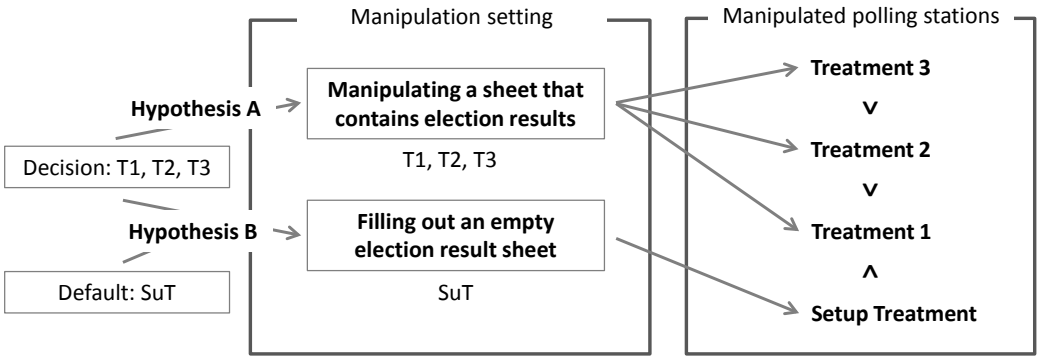


Figure 4.1: Expectations for the extent of manipulation for different treatments.

station, whether this is the original number, a slightly adapted number or a made up number. To conduct small changes in most vote counts is effective and it does not require more effort than filling in the sheet with the original numbers.

Fraud detection via LD test captures deviations in the LD. Two measures of these deviations are presented and are used to analyze the experimental data. The first is the Pearson's χ^2 statistic, which is most often used for the LD test.

$$\chi^2_{LD} = \sum_{i=0}^{9} \frac{(d_i - d * 0.1)^2}{d * 0.1}, \tag{4.1}$$

where d_i is the empirical frequency of the last digit i in a district and d is the sum of polling stations in the district. The expected relative frequency for each digit is 0.1 as LDs are expected to be uniformly distributed. The statistic is assumed to be distributed according to a χ^2 distribution with 9 degrees of freedom, this gives a critical value of 16.92 at a significance level of 5%. The second and closely related measurement is the probability of uniformity. It is defined as the probability of obtaining a result equal to or more extreme than what was actually observed, assuming that LDs are uniformly distributed. Therefore, the smaller the probability of uniformity is, the more likely is it to reject that LDs of party B are uniformly distributed. Its reference distribution is obtained with 1 million simulations.

4.5 Results

To get a first impression of the experimental outcome, subjects were asked to describe how they manipulated the election results in a short questionnaire at the end of the

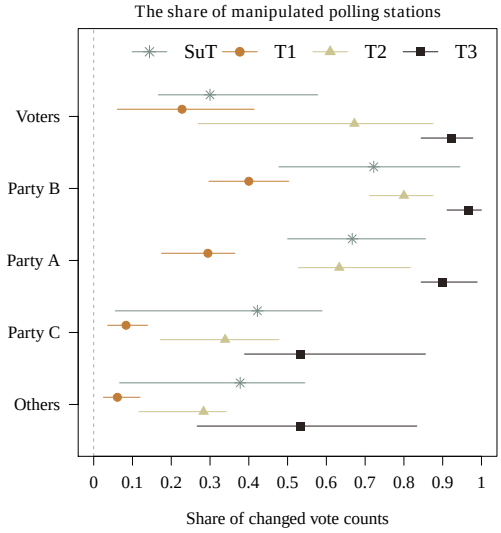


Figure 4.2: The extent of manipulation for different treatments and parties. The line represent subjects between the 25% and 75% percentile. The symbol indicates the median of subjects.

experiment. Only 3 out of 104 declared that they did not apply any strategy at all and most subjects mentioned multiple aspects they payed attention to while manipulating. Strategies that were mentioned by multiple subjects are the attempt to keep the turnout more or less constant, swapping votes of party A and party B, transferring votes between parties and making as little changes as necessary. A few subjects mentioned that they tried to make up unsuspicious numbers. The frequency and importance of the different strategies are investigated empirically. First, the results concerning the extent of manipulation are presented and afterwards the manipulation strategies that have been applied by subjects. Finally, it is presented what impact the extent of manipulation and the manipulation strategy have on detectability by LD test.

Based on the assumption of real effort tasks, participants that have the possibility to load election results are expected to do so.[10] In T1 no subject chose to fill out all three sheets by hand, two subjects chose to do so in T2 and one in T3 (compare Table C.1 in the appendix). This behavior is in line with the expectations. Figure 4.2 presents the median of changed vote counts across subjects, as well as the 25% and 75% percentile, separately for each treatment. It visualizes the manipulation behavior concerning the extent of vote count manipulations across a district, displayed separately for parties and treatments. The extent of manipulation was expected to increase across T1, T2 and T3 due to the increasing winning margin for party B,

[10]Two subjects were excluded from the experiment as they did not take the experiment seriously. This concerns one subject in T3 and one in the SuT that gave vote counts of 0 for party A and party B in about one-third to half of the polling stations.

which is clearly the case. The extent of manipulation and strategies are analyzed through randomized inference (compare Gerber and Green, 2012, 65) of the original data and treatments as well as between treatments. Results are presented in Table 4.1.

The results show that vote counts of party B were manipulated most often. In T1 most subjects manipulated 36 out of 90 vote counts of party B, while in T3 most subjects manipulated 87 out of the 90 vote counts. The changes in vote counts of party B highly correlate with changes in vote counts of party A (correlation between 0.8 and 0.94), except for the changes in T3 (correlation of 0.32). Changes of the vote counts for party C, others and voters were less correlated with changes of party B and differed across treatments. Therefore, participants' manipulation effort is not random and focuses on the two parties that were relevant for the electoral outcome. [11]

The experimental results confirms hypothesis A: subjects that have the choice to manipulate an existing election sheet will manipulate as many votes as necessary, but as few vote counts of different polling stations as possible. As the necessity of manipulation increases under higher intensity of fraud, it was then further expected that subjects of T3 manipulated most polling stations, those of T2 a bit less and subjects of T1 manipulated the least. This is not only represented in the individual average treatment effect (ATE) estimated for no manipulation in each treatment, but also that treatments significantly differ from each other (compare Table 4.1). Furthermore, the results confirm that subjects of the SuT (that had to fill out empty result sheets) manipulated differently under the same intensity of fraud than in T1. The ATE of the SuT compared to no manipulation estimates about 70% manipulated polling stations, and a significant increase of about 31% manipulated polling stations compared to T1 (compare Table 4.1). This is in line with hypothesis B: subjects manipulate as many votes as necessary and distribute them across the most vote counts of a district. Therefore, the manipulation extent depends on the circumstances (filling out empty return sheets or manipulating within an return sheet) as well as the intensity of fraud. As shown by the second condition in section 4.2, the LD test can only detect election fraud if the manipulated data dominates the election results. In many of the presented scenarios this is not the case, which limits the applicability of the LD test. The additional challenge is that in real election fraud cases it will be difficult to judge how many polling stations are potentially manipulated.

Concerning the manipulation strategy, the current literature assumes that if a person changes the vote count of a particular party she replaces it by randomly made up numbers (Beber and Scacco, 2012; Weidmann and Callen, 2013). This is,

[11] The variation between the subjects, the manipulation of their individual sheets, the differences between treatments and the connection to changes in party B are visualized in Figure C.7 in the appendix.

Table 4.1: Randomization inference

	Original data and treatments				Between treatments			
	Comparison	Variable	ATE	p-value	Comparison	Variable	ATE	p-value
Extent of manipulation	T0 and SuT	Voters	0.406	0.000	T1 and SuT	Voters	0.149	0.058
	T0 and T1	Voters	0.256	0.000	T1 and T2	Voters	0.310	0.000
	T0 and T2	Voters	0.566	0.000	T1 and T3	Voters	0.578	0.000
	T0 and T3	Voters	0.835	0.000	T2 and T3	Voters	0.268	0.002
	T0 and SuT	Party B	0.698	0.000	T1 and SuT	Party B	0.312	0.000
	T0 and T1	Party B	0.385	0.000	T1 and T2	Party B	0.388	0.000
	T0 and T2	Party B	0.774	0.000	T1 and T3	Party B	0.557	0.000
	T0 and T3	Party B	0.943	0.000	T2 and T3	Party B	0.169	0.000
	T0 and SuT	Party A	0.644	0.000	T1 and SuT	Party A	0.335	0.000
	T0 and T1	Party A	0.309	0.000	T1 and T2	Party A	0.327	0.000
	T0 and T2	Party A	0.635	0.000	T1 and T3	Party A	0.562	0.000
	T0 and T3	Party A	0.871	0.000	T2 and T3	Party A	0.236	0.000
	T0 and SuT	Party C	0.395	0.000	T1 and SuT	Party C	0.293	0.000
	T0 and T1	Party C	0.102	0.000	T1 and T2	Party C	0.237	0.000
	T0 and T2	Party C	0.339	0.000	T1 and T3	Party C	0.477	0.000
	T0 and T3	Party C	0.579	0.000	T2 and T3	Party C	0.240	0.002
	T0 and SuT	Others	0.385	0.000	T1 and SuT	Others	0.302	0.000
	T0 and T1	Others	0.083	0.000	T1 and T2	Others	0.196	0.000
	T0 and T2	Others	0.279	0.000	T1 and T3	Others	0.446	0.000
	T0 and T3	Others	0.529	0.000	T2 and T3	Others	0.249	0.002
Manipulation strategies	T0 and SuT	LD	0.588	0.000	T1 and SuT	LD	0.306	0.000
	T0 and T1	LD	0.282	0.000	T1 and T2	LD	0.225	0.001
	T0 and T2	LD	0.507	0.000	T1 and T3	LD	0.244	0.000
	T0 and T3	LD	0.526	0.000	T2 and T3	LD	0.019	0.797
	T0 and SuT	Digit 1 or 2	0.570	0.000	T1 and SuT	Digit 1 or 2	0.248	0.000
	T0 and T1	Digit 1 or 2	0.322	0.000	T1 and T2	Digit 1 or 2	0.396	0.000
	T0 and T2	Digit 1 or 2	0.718	0.000	T1 and T3	Digit 1 or 2	0.610	0.000
	T0 and T3	Digit 1 or 2	0.932	0.000	T2 and T3	Digit 1 or 2	0.214	0.000
	T0 and SuT	Random number	0.403	0.000	T1 and SuT	Random number	0.217	0.000
	T0 and T1	Random number	0.186	0.000	T1 and T2	Random number	0.189	0.000
	T0 and T2	Random number	0.375	0.000	T1 and T3	Random number	0.246	0.000
	T0 and T3	Random number	0.432	0.000	T2 and T3	Random number	0.057	0.334
	T0 and SuT	Swapping A & B	0.170	0.000	T1 and SuT	Swapping A & B	0.087	0.125
	T0 and T1	Swapping A & B	0.082	0.000	T1 and T2	Swapping A & B	-0.022	0.656
	T0 and T2	Swapping A & B	0.061	0.000	T1 and T3	Swapping A & B	-0.058	0.108
	T0 and T3	Swapping A & B	0.024	0.010	T2 and T3	Swapping A & B	-0.036	0.098
	T0 and SuT	Turnout	0.024	0.008	T1 and SuT	Turnout	0.022	0.027
	T0 and T1	Turnout	0.002	0.404	T1 and T2	Turnout	0.014	0.005
	T0 and T2	Turnout	0.016	0.001	T1 and T3	Turnout	0.051	0.009
	T0 and T3	Turnout	0.053	0.007	T2 and T3	Turnout	0.037	0.067
LD deviation	T0 and SuT	χ^2 statistic	5.653	0.000	T1 and SuT	χ^2 statistic	5.235	0.002
	T0 and T1	χ^2 statistic	0.436	0.582	T1 and T2	χ^2 statistic	8.530	0.000
	T0 and T2	χ^2 statistic	8.966	0.000	T1 and T3	χ^2 statistic	4.835	0.009
	T0 and T3	χ^2 statistic	5.289	0.001	T2 and T3	χ^2 statistic	-3.695	0.278
	T0 and SuT	Prob. uniformity	-0.256	0.000	T1 and SuT	Prob. uniformity	-0.245	0.003
	T0 and T1	Prob. uniformity	-0.012	0.838	T1 and T2	Prob. uniformity	-0.295	0.000
	T0 and T2	Prob. uniformity	-0.307	0.000	T1 and T3	Prob. uniformity	-0.192	0.039
	T0 and T3	Prob. uniformity	-0.206	0.004	T2 and T3	Prob. uniformity	0.103	0.221

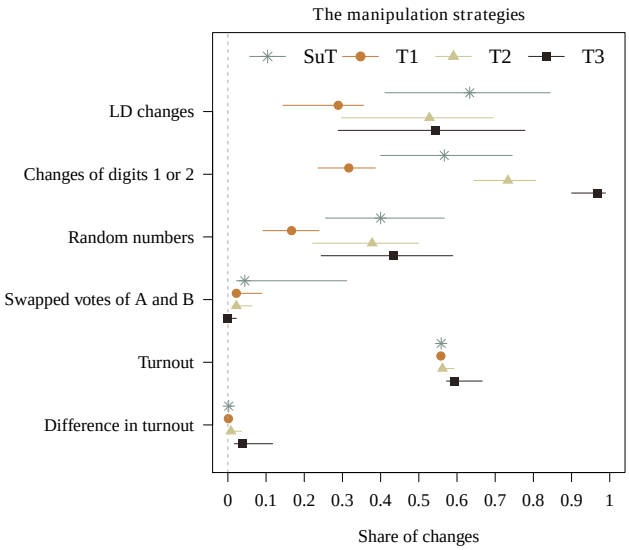


Figure 4.3: Strategies that are used for manipulation. The line represent subjects between the 25% and 75% percentile. The symbol indicates the median of subjects.

however, only one out of many possibilities. Figure 4.3 shows the median, 25% and 75% percentile for different strategies that are traceable in the experiment. The first displayed strategy is changes in the LD of party B. While changes of party B's vote count are often associated with changes in the LD, the correlation decreases across the different treatments and is very low for T3. Therefore, subjects seem to change how they manipulate if they have to manipulate a lot. This is further represented in the share of manipulated first and second digits that significantly increase between treatments. The fewest changes are in T1, followed by T2 and almost all first or second digits are manipulated in T3. The literature suggests that vote counts are replaced by made up numbers, which is, according to the presented strategies, not always the case. The random number generation captures this aspect more directly, but shows similar results. While the intensity of manipulation increases across T1, T2 and T3, only T1 differs significantly according to randomized inference from T2 and T3. Subjects of T3 do not significantly produce more random numbers than those in T2 even though they manipulate significantly more polling stations. A clear difference between the manipulation approach can be established for SuT and T1 as subjects of the SuT generate about 22% more random numbers than those of T1 (compare Table 4.1). To disentangle the extent of manipulation and the use of a particular strategy for a more profound understanding, they are displayed as relative shares to the changed vote counts of party B in Figure C.6 in the appendix.

Another strategy that can represent very efficient manipulation is vote swapping between party B and party A. While the share of swapped votes between party A and party B is small, the randomized inference of Table 4.1 confirms that such swaps

are significantly different from 0 for SuT, T1, and T2. The ATE of T1 compared to no manipulation is about 8% which transfers to 7 swapped vote counts of party A and party B. If the vote counts in polling stations with the highest difference between party A and party B are swapped, than 7 swaps are enough to make party B the new winner of the district. The number of swaps is even higher for SuT and lower for T2. Therefore, swapping should be considered a relevant strategy that humans might apply when manipulating. Swapping becomes less useful the higher the (requested) winning margin for the new winner is since swapping is no longer sufficient in fulfilling the required manipulation. Subjects adjust to this insufficiency as they do not significantly use swapping as a manipulation strategy in T3.

A simple way to increase the vote share of party B would be to add votes to this party and therefore increase turnout. Subjects mentioned that they considered increasing turnout as suspicious and therefore tried to avoid it. According to randomized inference, subjects do not significantly increase turnout in T1, slightly in the SuT and T2 and only by about 5% in T3 were extremely high winning margins are required. This means that subjects rather subtracted votes from other parties, mostly from party A due to the high correlation between both. Up to now, all the different investigated aspects show that election manipulation is conducted very strategically. Participants chose efficient strategy that fit the particular context.

The LD test is designed to detect deviations of the LD of vote counts. What happens to the LD test, or its probability to detect manipulation, if humans manipulate only a few or some vote counts? Moreover, what happens if they use strategies besides replacing vote counts with made up numbers? The results suggest that there are significant differences between the LD test of the original data and the manipulated data in T2, SuT and T3. The presented order corresponds to the strength of the deviation, captured by the χ^2 statistic and the probability of uniformity (compare Table 4.1). While the ATE is considerable, it is not necessarily high enough to cause significant deviation of the LD test.[12] Randomized inference gives no significant increase of T1s LD tests compared to no manipulation. Subjects' manipulation in T1 is not sufficient to increase the χ^2 statistic and decrease the probability of uniformity compared to no manipulation. On average the χ^2 statistic is higher for SuT than T1 and respectively lower for the probability of uniformity. The same is found for T2 and T1 as well as T3 and T1, but there is no significant difference between T2 and T3. Therefore, manipulating considerable or extensively does not seem to differ in its effect on the LD test.

Table 4.2 presents the results of beta regressions for different model specifications, estimating the effect of manipulation on the probability of uniformly distributed LDs. The different treatments capture the intensity of manipulation with T1 as

[12] As a reminder, for a significance level of 5% the critical χ^2 value by 9 degree of freedom is 16.92 and the probability of uniformity should be smaller than 0.05.

Table 4.2: The Effect of Manipulation on the Probability of LDs Uniformity: Beta Regression Estimates

	(1)	(2)	(3)	(4)
(Intercept)	0.73	0.95	0.97	0.74
	(0.70)	(0.65)	(0.64)	(0.69)
Setup treatment	−0.23	−0.87	−0.99	−0.30
	(0.61)	(0.60)	(0.60)	(0.61)
Treatment 2	−0.54	−0.91**	−0.85*	−0.35
	(0.42)	(0.33)	(0.34)	(0.45)
Treatment 3	0.47	−0.23	−0.09	0.83
	(0.51)	(0.36)	(0.37)	(0.58)
Changes	−2.21**			
	(0.73)			
LD changes		−2.29***		
		(0.53)		
Random numbers			−3.00***	
			(0.65)	
Changes of digits 1 or 2				−2.67**
				(0.82)
Vote swapping	1.85*	2.45**	2.56**	1.98*
	(0.81)	(0.81)	(0.81)	(0.81)
Δ Turnout	−0.71	−0.94	−0.44	0.12
	(2.01)	(1.94)	(1.95)	(2.06)
Dataset 2	−0.27	−0.05	−0.21	−0.31
	(0.24)	(0.22)	(0.23)	(0.25)
Loaded results >= 2	0.34	−0.23	−0.26	0.35
	(0.51)	(0.53)	(0.52)	(0.51)
Precision: (phi)	1.74***	1.91***	1.92***	1.76***
	(0.22)	(0.24)	(0.24)	(0.22)
Pseudo R^2	0.22	0.30	0.30	0.23
Log Likelihood	50.66	55.77	56.02	51.20
AIC	-81.32	-91.54	-92.05	-82.40
Num. obs.	102	102	102	102

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Note: Additional controls that were included are the time a subject needed for manipulation as linear and non-linear predictor, subjects ability to correctly evaluate a displayed sheet and their ability to correctly detect manipulated results. None of these controls are relevant or substantially impact the effect of other variables. The results are available upon request.

reference category. As before, T1 and SuT do not differ in intensity, but in how the manipulation is carried out (changing election result sheets compared to filling out empty sheets). As discussed, participants used different strategies for manipulation. It varies in how many polling stations participants manipulate and also how they manipulate them. Different indicators cover this aspect to different extent. The changed vote counts of party B captures the overall manipulation that an individual conducted (compare Model (1)). However, it does not capture if the manipulation targeted for example a specific digit. Therefore, the share of changed LDs is included as a separate indicator in Model (2). Furthermore, it is assumed that humans replace the vote count with made up numbers but fail to produce random ones. The share of random number generation is included in Model (3). For reasons of completeness, the share of changes in the first or second digit is included in Model (4). Furthermore, the share of swapped votes and the difference in turnout are included as manipulation strategies in all models. The following variables are included as controls: Dataset 2 is included to ensure that results do not depend on the original data set. Loaded results 2 is a dummy variable, 1 indicating that a subject manipulated two or three out of the three sheets based on the loaded election results and 0 indicates that a subject filled out all three sheets or at least two by hand.[13]

The indicators for the different manipulation strategies are not independent of each other and at least two are highly correlated. These are the changed vote counts and changes of the first or second digit as well as LD changes and random number generation. The correlation for both across treatments is above 0.9. All indicators have a significant negative effect on the probability of uniformly distributed LDs, but based on the AIC criterion, the LD changes and the random number indicator are the best predictors for deviations. Moreover, the changes of digits 1 or 2 lose their explanatory power, if either one is included in the same model as LD changes or random numbers. The finding supports the argument that humans are bad random number generators also in the context of election manipulation. However, the predictions based on Model 2 show that not even 100% changed LDs would trigger the LD test. This is displayed in Figure 4.4. The solid line displays the median and the dashed line the 5% and 95% credibility interval of the predicted probability of uniformly distributed LDs for increasing shares of changed LDs.[14] This finding is powerful as it challenges the applicability of the LD test even given man-made numbers (100% changed LDs do not cause the LD test to indicate fraud).

Moreover, results show that other strategies can intervene. Vote swapping has the significantly opposite effect of similar size than LD changes. It increases the probabil-

[13]The definition is chosen as it represents the "dominant" manipulation setting, but robustness checks for different definitions as presented in the Appendix in Table C.2. The step-by-step results are available upon request.

[14]The predictions of Model 2 are based on Bayesian Inference for Beta Regression.

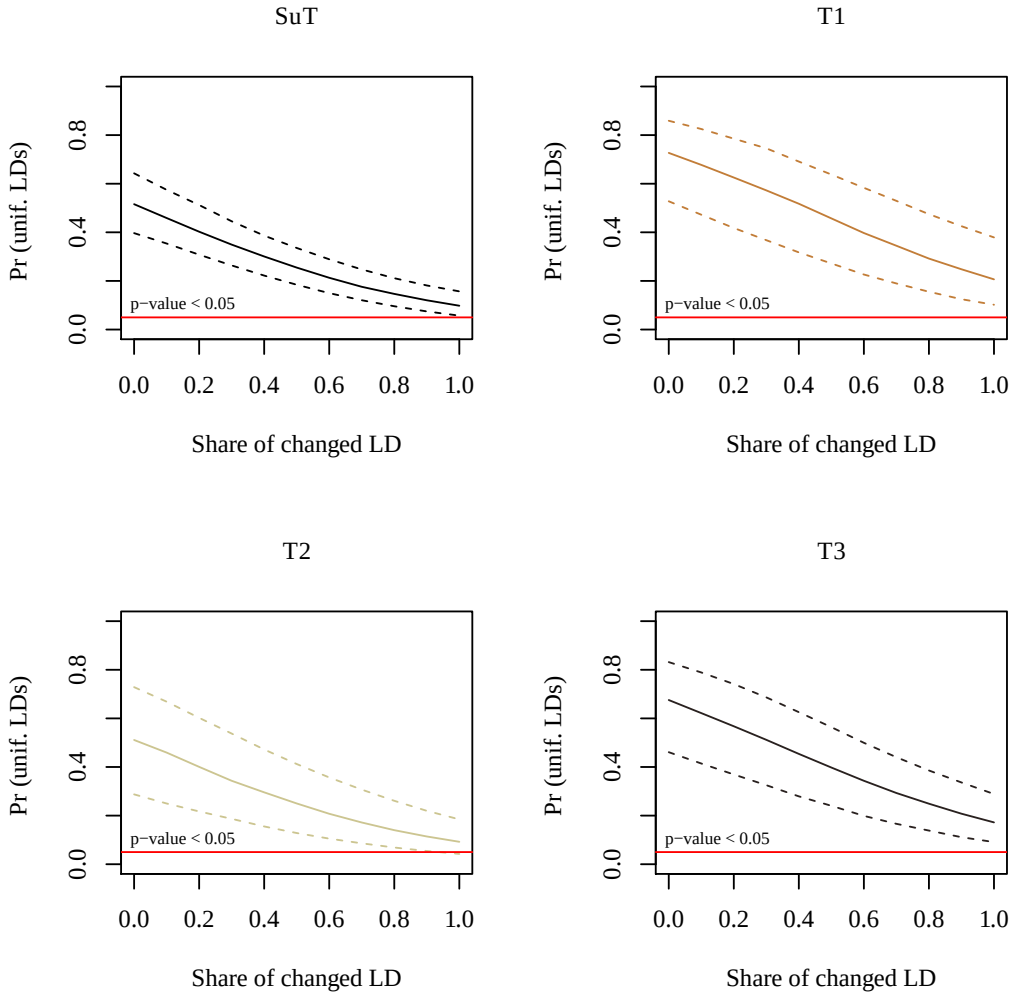


Figure 4.4: Detectability of manipulations by the LD test. The graph shows the predicted probability of uniformly distributed LDs given an increasing share of changed LDs based on the estimations of Model 2. The red line indicated the significant level of 5%, crossing this line would indicate a significant LD test. Based on the results, the LD test will not signal significant deviations even given 100% of LD changes.

ity that LDs are uniformly distributed.[15] The manipulation extent, the manipulation strategies and the inability of the LD test to signal even high levels of LD changes adds up to strong evidence against the LD test as detection method for human vote count manipulation.

4.6 Conclusion

Previous research uses the LD test to investigate if humans conducted vote count manipulation for a particular election. This relays on a strong assumption that humans replace almost all vote counts by made-up non-random numbers because humans are incapable of producing such numbers. The assumption disregards the fact that humans act strategic and therefore might manipulate differently or produce different numbers. In fact, how humans manipulate has never before been investigated. Concretely, the strategic behavior of humans could lead to three difficulties in applying the LD test: Firstly, someone could manipulate only a certain number of votes and, accordingly, not enough vote counts within a constituency for LD detectability. Secondly, other manipulation strategies are used then to replace counted votes with man-made numbers, (e. g. vote swaps). Thirdly, due to the complexity of falsifying election results, man-made numbers are generated differently than in the numerous experiments in which people deliberately try to generate random numbers. The paper utilizes an experimental study to investigate these aspects.

The experiment encouraged inconspicuous manipulation in favor of a particular party through monetary incentives and an evaluation mechanism between subjects. The different treatments tackle the intensity of manipulation and the manipulation setup, which affects the manipulation behavior.

The experimental results show that subjects manipulate as few polling stations as possible if they have the option to manipulate within an existing election result sheet. Therefore, the extent of manipulation increases with increasing intensity of fraud. This confirms the hypothesis that subjects perceive little interventions as inconspicuous as long as they have the option to conduct such efficient manipulation. If subjects do not have such an option and have to fill out empty sheets, they seem to perceive it as less suspicious to manipulate many polling stations just a little bit. Thus, under the same intensity of fraud the extent is significantly and substantially higher for the different setup. Consequently, it is not plausible to assume that almost all vote counts of an election result sheet will be replaced by man-made numbers.

It is assumed that manipulation is conducted through replacing the vote count of the manipulating party with made-up numbers, which is not necessarily the case. There are big differences between treatments whether subjects produced a random

[15] The two main effects, LD changes and vote swapping, are robust for all specifications for loaded results (compare Table C.2 in the appendix).

number or if they changed a specific digit. With increasing intensity of fraud subjects replace more often one of the front digits without changing the LD. LD changes are often conducted when the whole number is replaced. Replacing one of the front digits can be an efficient strategy to increase the manipulated votes favored by subjects if the intensity of fraud is high. Vote swapping between the two main parties is uncovered as an influential strategy. On average subjects did not swap a big percentage of votes, but it can still be enough to change the election outcome. Subjects used the strategy as long as it was sufficient for winning, but they neglected it in insufficient cases as in extreme manipulation. Last but not least, increasing turnout seems to be considered as very suspicious as subjects only slightly increased turnout if the intensity of fraud was very high. These increases are smaller than one would expect. The experiment showed that subjects are creative in choosing manipulation strategies by applying them very strategically when manipulating. Therefore, assuming human vote count manipulation is simply a process were humans try to generate random numbers and replace them with vote counts does not meet the strategic behavior and variety of strategies that have been identified in this experiment.

LD deviations from uniformity depended strongly on the extent of fraud. The extent increased with higher intensity of fraud, but also for low intensity of fraud in combination with a different setup of the manipulation process. Regression analysis contributes to important findings: it shows that changes of LDs and replacing vote counts by random numbers are good predictors for LD deviations. Therefore, humans are bad random number generators when manipulating the LD of vote counts. However, the prediction of the regression analysis shows, not even if 100% of LD are manipulated it would trigger the LD test and signal election fraud.

Concluding, the link between human vote count manipulation and LD deviations is not as simple as suggested. How the election results are manipulated depends on many different factors like the intensity of fraud, the circumstances under which manipulation is conducted and the manipulator itself. Moreover, based on the experimental findings, it seems unlikely that the LD test can properly detect this kind of election fraud at all.

The results of the study therefore question the applicability of LD test as an indicator of election fraud. This raises the question of what was actually identified in cases of significant test results for previous applications. In the most prominent study of Beber and Scacco (2012), it is argued in detail why deviations are not due to rounded vote counts and thus it can be distinguished between negligence of the electoral commission or counting process and deliberate election fraud. The presented results question this interpretation as rounding the numbers is the single remaining mechanism which should lead to deviations captured by the LD test. As Beber and Scacco (2012) points out, under these circumstances it is no longer possible to distinguish between negligence in the counting process and deliberate election fraud.

Therefore, it should be considered carefully if applying LD test can produce additional knowledge and how significant test results are interpreted.

However, the study is only the beginning when it comes to investigating human behavior in the context of election manipulation and has certain limitations. For example, the experiment was carried out with mostly university students of a Western European country (Germany), which might use very different manipulation strategies than a manipulator in another country. One strategy that has been identified emphasizes this aspect that is the subjects' aversion to increase turnout. Subjects obviously perceive an increase in turnout as very suspicious, which is likely the result of past election experience in Germany with moderate changes in turnout and the negative stigma of high turnout and manipulated elections in the DDR. The aversion to increase turnout can be very distinctive in different countries depending on what is perceived as convincing. What seems general is the subjects' tendency to use different manipulation strategies, which is not necessarily replacing vote counts by man-made numbers. Moreover, subjects adapt their behavior given the circumstances and the intensity of fraud, which makes it difficult to formulate a distinct assumption about manual manipulation. Nevertheless, these core findings should also be examined in more detail in other countries and under different conditions. This would provide valuable insights to how human vote count manipulation is conducted and contribute to fraud detection by election observers and statistical analysis.

A

Supplementary Information for Chapter 2

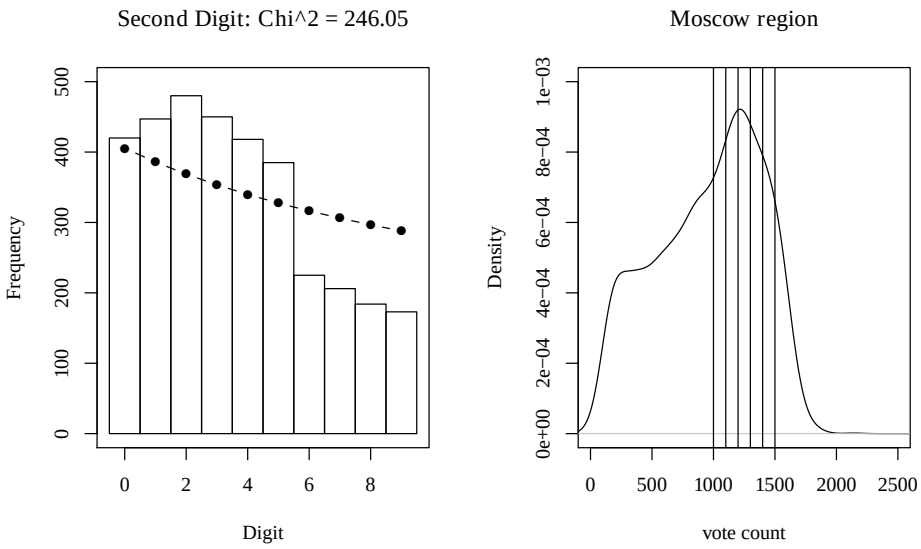


Figure A.1: The Figure to the left shows a histogram which gives second digits of vote counts and the black line the expected frequency of digits according to 2BL. The Figure to the right shows the density distribution of votes for Putin in the Moscow region (unit 58). The vertical line marks the numbers 1000 - 1500.

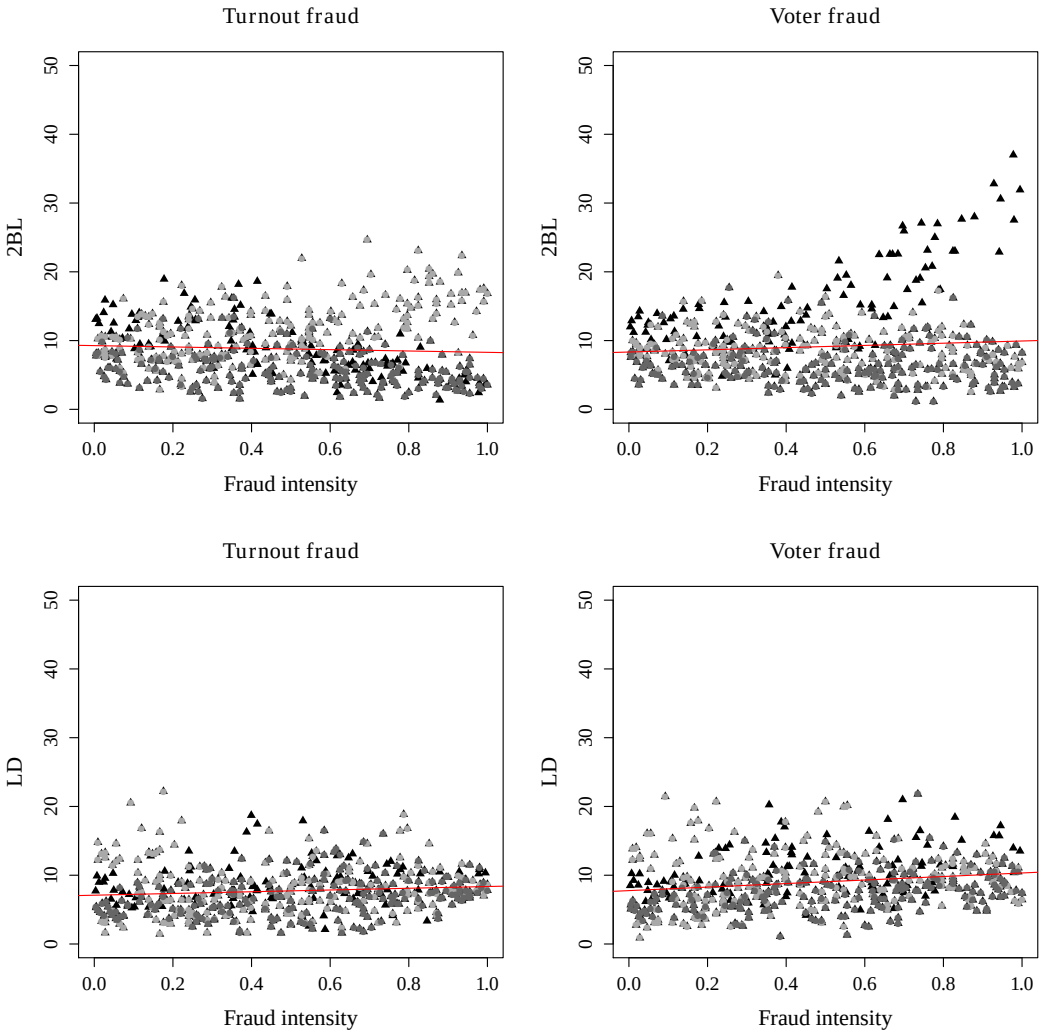


Figure A.2: The upper plots scatters the fraud intensity and 2BL statistic in a district, the lower once scatter fraud intensity and LD statistics. The dark grey points display vote counts from France, light grey points those from Finland and the black triangle are Russian vote counts. The red line shows the regression of digit tests (2BL or LD) on the fraud intensity. There is no linear relationship between the relative share of manipulated polling stations and the digit statistics. Hence, if more polling stations in a district get massively manipulated it does not increase one of the digit test statistics. Again, we find one exception which concerns Russia: if in nearly all polling stations vote counts of Putin are replaced with the total number of voters, we get an inflated 2BL statistic. This hinds to the fact that vote counts of voters are not distributed according to 2BL which was also identified through fraud mechanism 1.

B

Supplementary Information for Chapter 3

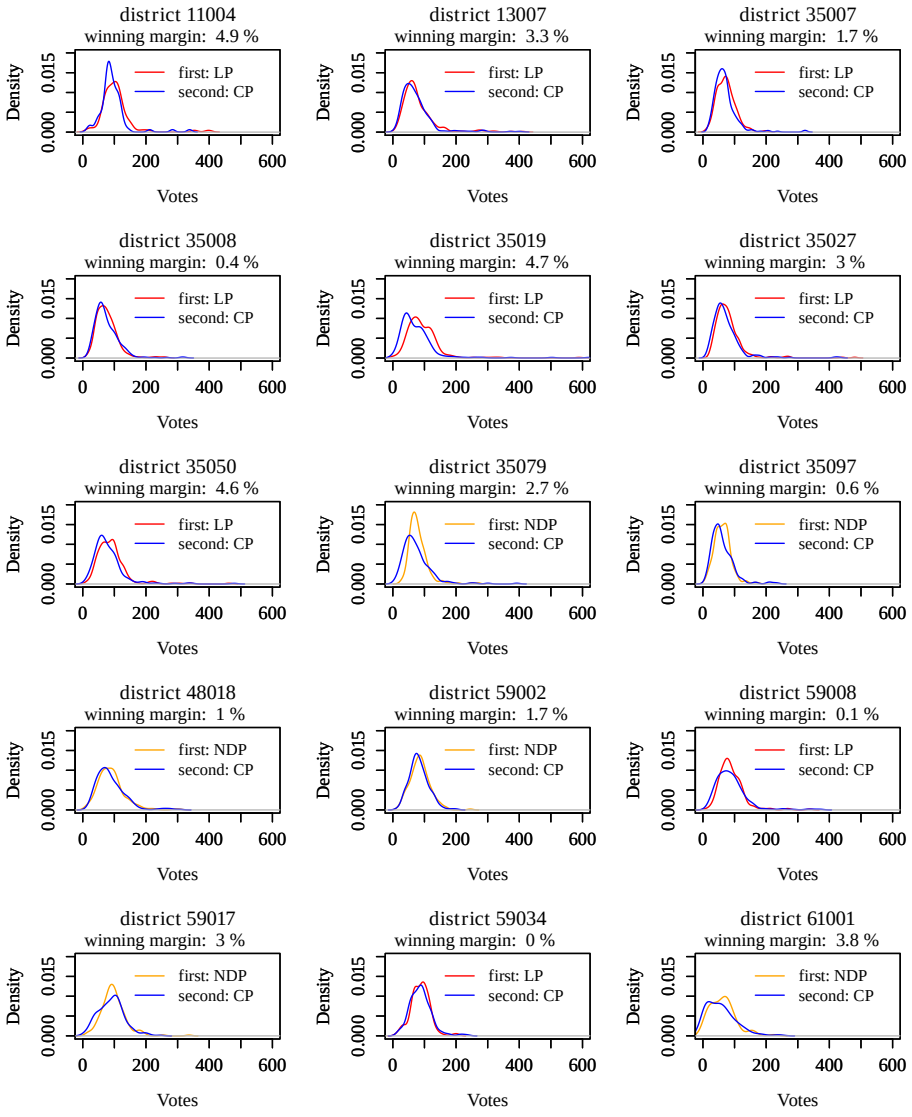


Figure B.1: Distribution of votes for all districts with a winning margin of less than 5%.

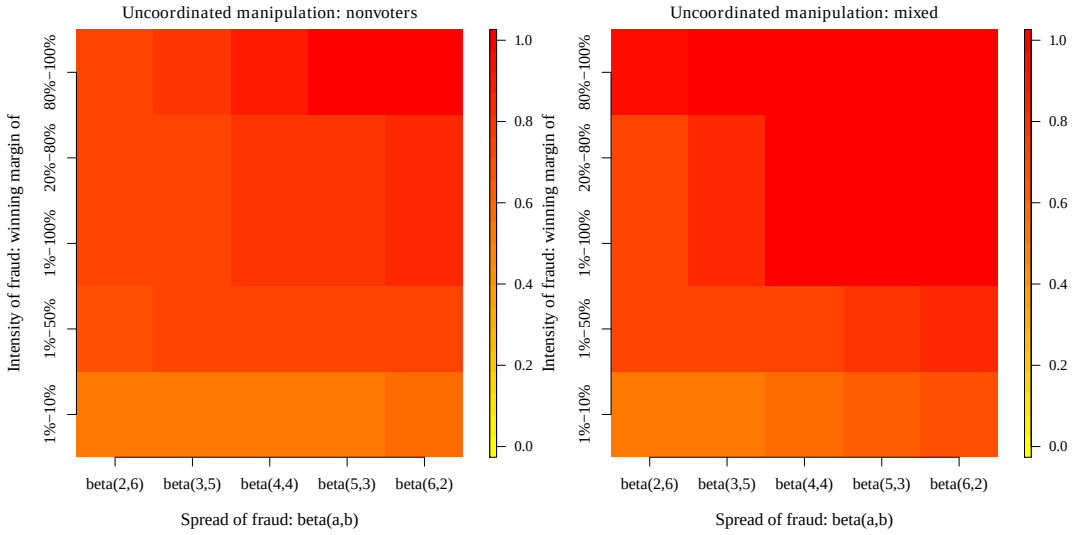


Figure B.2: It shows the proportion of wins for the Conservative Party across the 15 manipulated districts and 1000 calibrations. Logically, applying a mixture of fraud types makes winning easier, but they win only about 50% of the time if fraud is marginal.

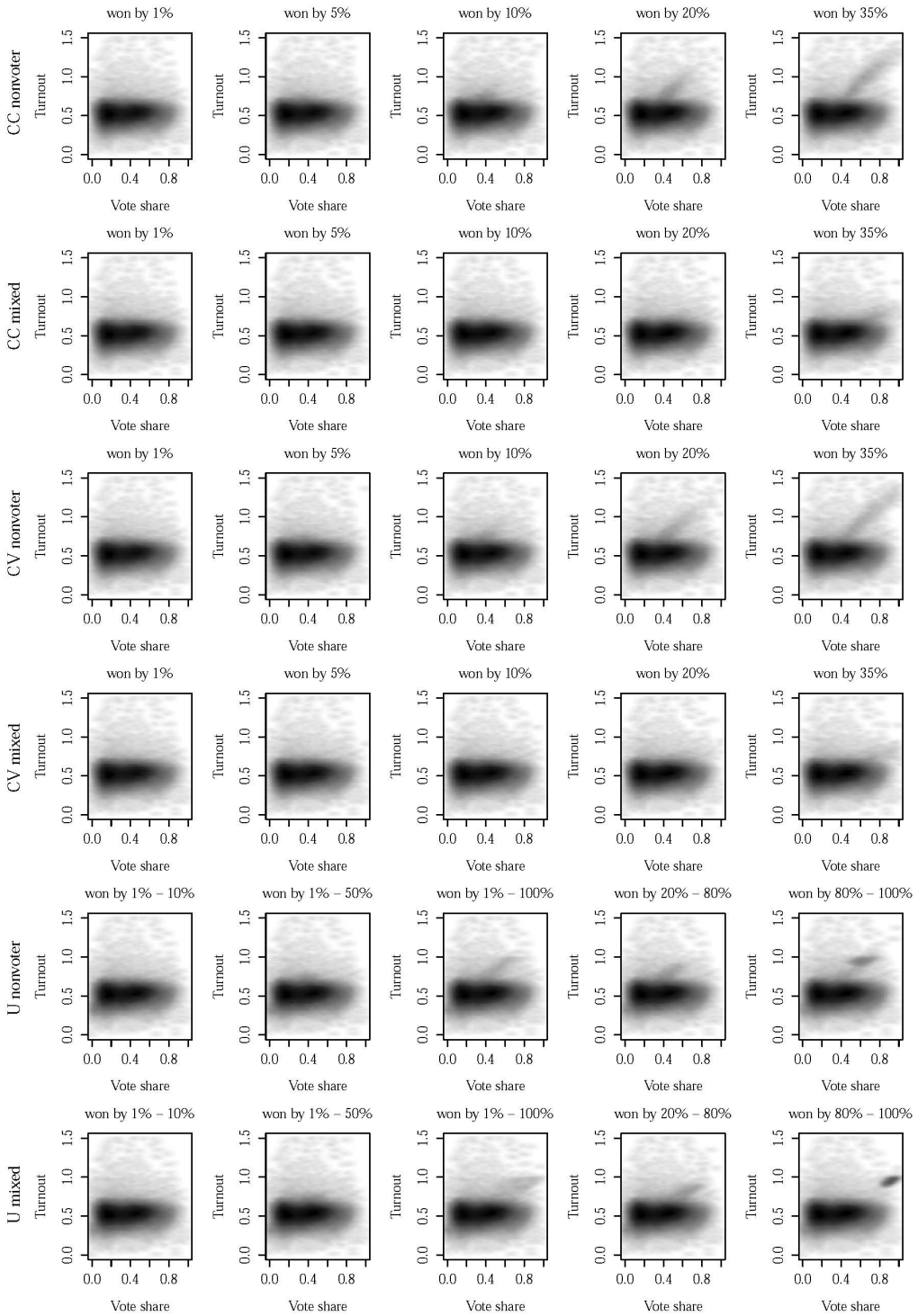


Figure B.3: Vote share and turnout across the Canadian election 2008 is combined with one calibration of each of the 15 districts that were artificially manipulated. It is displayed for the different mechanisms and fraud intensity, the spread is kept constant to $beta(4, 4)$. The gray shades display the density of data points.

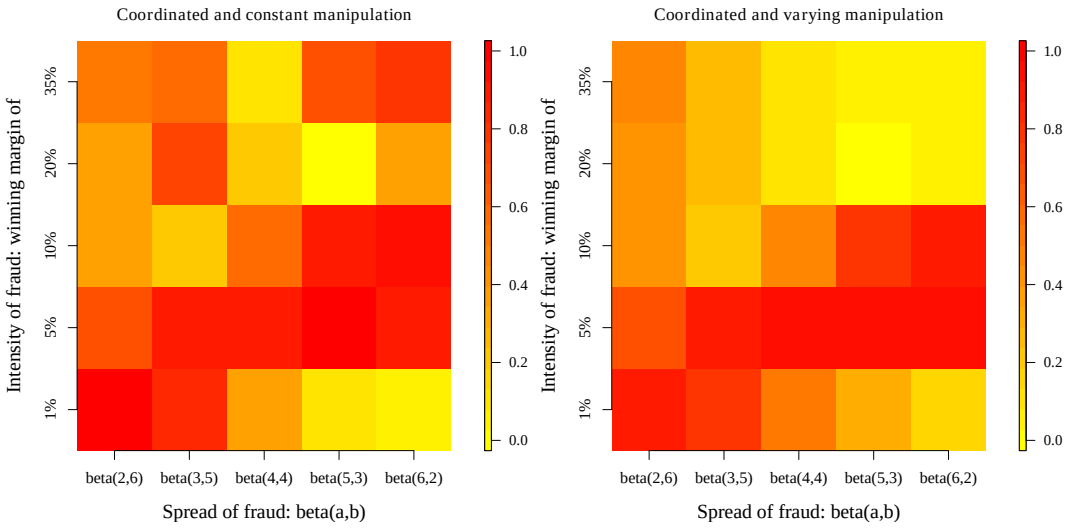


Figure B.4: Coordinated fraud in district 59034. It shows the proportion of χ^2 statistics exceeding the critical value of 16.92 at a significance level of 5% across the 15 districts and 1000 simulations.

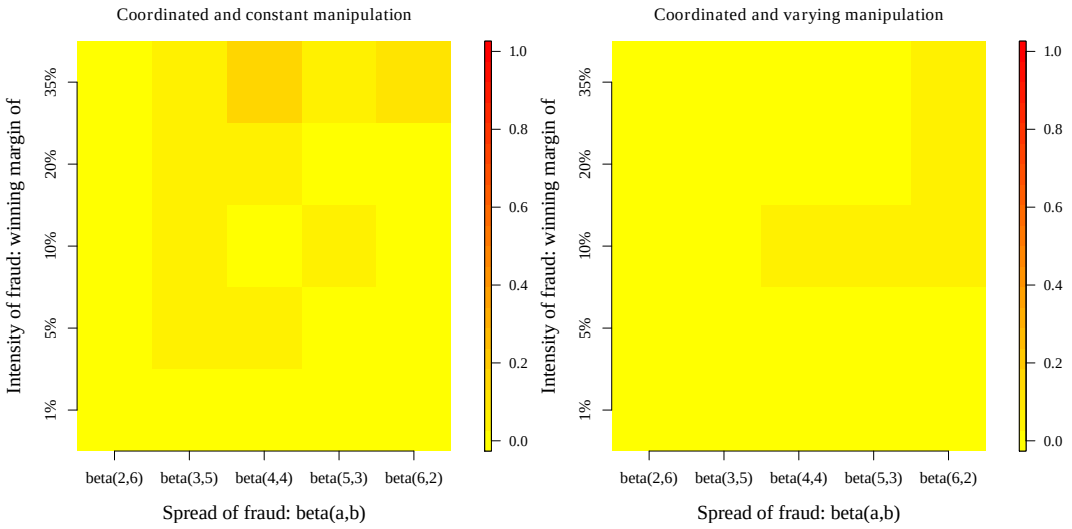


Figure B.5: Coordinated fraud in district 61001. It shows the proportion of χ^2 statistics exceeding the critical value of 16.92 at a significance level of 5% across the 15 districts and 1000 simulations.

C

Supplementary Information for Chapter 4

C.1 Experimental design and procedures

This section gives additional information to section 4.3. It goes into details of the experimental design and procedures.

Participants have full information about the experimental procedure. This means that they are familiar with the procedures of the manipulation and evaluation task, including that they have to decide in the evaluation phase if a displayed sheet corresponds to real vote counts or if it was manipulated by a different participant. The detailed procedure of the manipulation and evaluation task is outlined in the following paragraphs.

Manipulation: on the left side of the screen, the table with the real outcome of the 30 polling stations is displayed. It shows eligible voters, voters and the votes for party A, party B, party C and other parties. The overall counts are displayed on the bottom of the table, as well as the vote share of each party. On the right side of the screen, the same table is displayed, but it contains the eligible voters and empty fields to enter (new) vote counts for each party and polling station (compare Figure C.1). The participants can decide whether they complete the table by hand or load the actual results and change as many vote counts of each party and polling station as they wish (compare Figure C.2).

Within this setting, it is not possible to enter higher numbers of voters than for eligible voters.[1] As the experiment aims to investigate manipulation strategies and the randomness of humanly produced vote counts, but not (individual) mathematical skills, the program enables participants to calculate (and immediately display) the (new) voters in each polling station. The feature also calculates the (new) overall voters, the total vote count for each party and the vote share of each party (compare Figure C.3). The manipulation and the calculation can be repeated infinitely. Further, participants have plenty of time (i.e. no time limit is imposed). Participants have to confirm when they finished manipulating a sheet and cannot continue with the experiment if they do not reach the requested winning margin.

The setup treatment (SuT) explores manipulation if subjects cannot choose to load the actual election results and instead have to fill out the table on the right side by hand. The treatment is equivalent to T1 in terms of the smallest winning margin, therefore compare Figure C.1, and only misses the button to load results (lade Wahlergebnis).

Evaluation: This phase is conducted after all participants finished the manipulation task to ensure that they are not influenced by manipulation strategies of others.

[1] More specifically, the smallest number that participants could enter is 0 and the highest they could enter is the number of eligible voters in each polling station. It is not possible that the overall sum of voters exceeds the overall sum of eligible voters, but voters could exceed the eligible voters in individual polling stations. However, participants are instructed that more voters than eligible voters are considered an obvious indication of election fraud.

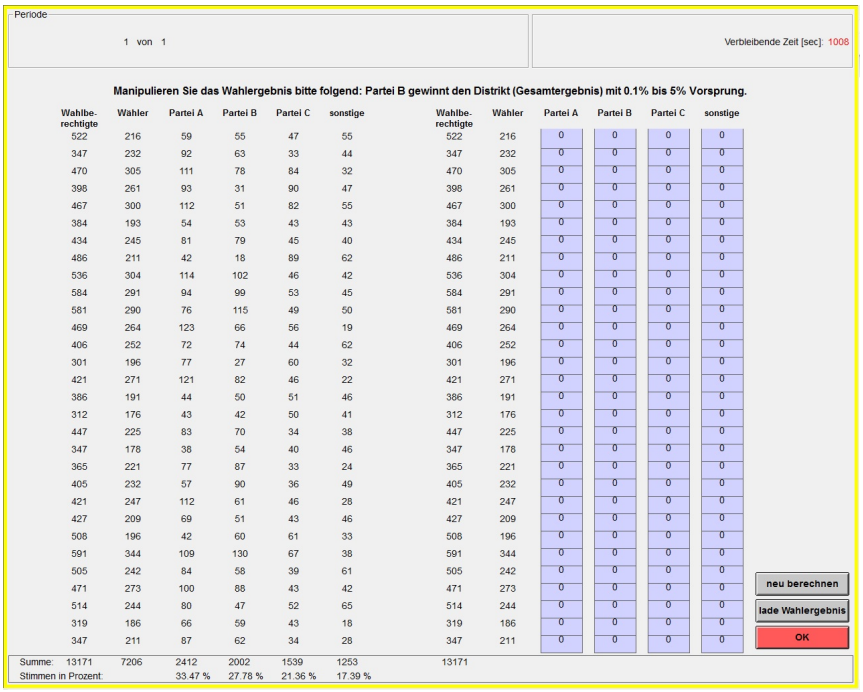


Figure C.1: Manipulation task: empty table to the right.

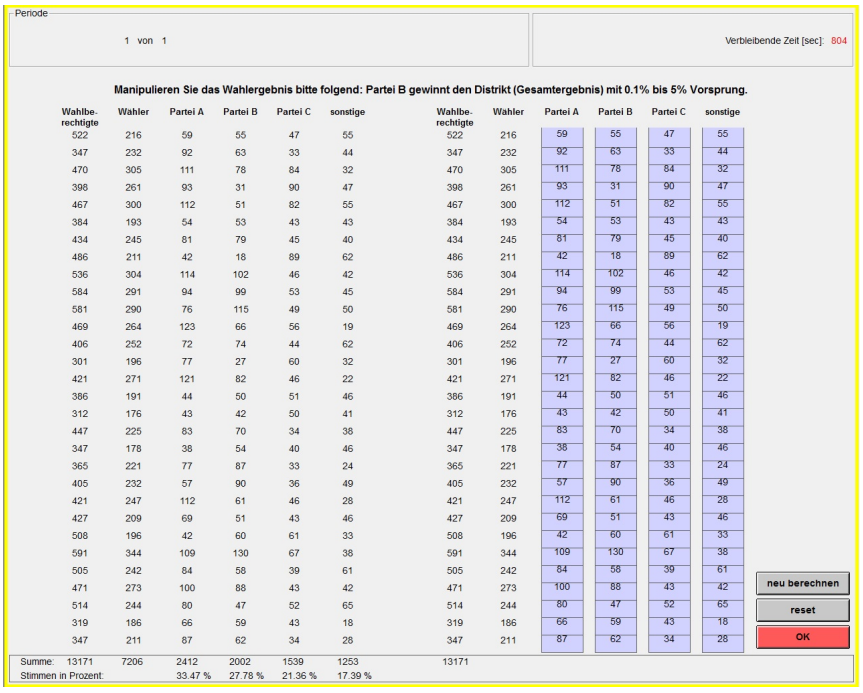
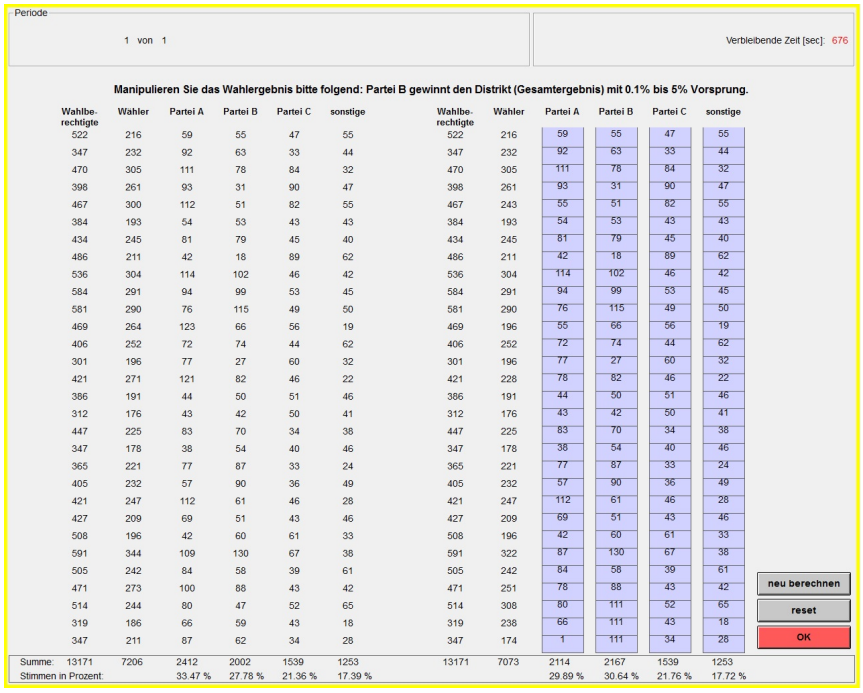


Figure C.2: Manipulation task: loaded results to the right.



Wahlberechtigte	Wähler	Partei A	Partei B	Partei C	sonstige	Wahlberechtigte	Wähler	Partei A	Partei B	Partei C	sonstige
522	216	59	55	47	55	522	216	59	55	47	55
347	232	92	63	33	44	347	232	92	63	33	44
470	305	111	78	84	32	470	305	111	78	84	32
398	261	93	31	90	47	398	261	93	31	90	47
467	300	112	51	82	55	467	243	55	51	82	55
384	193	54	53	43	43	384	193	54	53	43	43
434	245	81	79	45	40	434	245	81	79	45	40
486	211	42	18	89	62	486	211	42	18	89	62
536	304	114	102	46	42	536	304	114	102	46	42
584	291	94	99	53	45	584	291	94	99	53	45
581	290	76	115	49	50	581	290	76	115	49	50
469	264	123	66	56	19	469	196	55	66	56	19
406	252	72	74	44	62	406	252	72	74	44	62
301	196	77	27	60	32	301	196	77	27	60	32
421	271	121	82	46	22	421	228	78	82	46	22
386	191	44	50	51	46	386	191	44	50	51	46
312	176	43	42	50	41	312	176	43	42	50	41
447	225	83	70	34	38	447	225	83	70	34	38
347	178	38	54	40	46	347	178	38	54	40	46
365	221	77	87	33	24	365	221	77	87	33	24
405	232	57	90	36	49	405	232	57	90	36	49
421	247	112	61	46	28	421	247	112	61	46	28
427	209	69	51	43	46	427	209	69	51	43	46
508	196	42	60	61	33	508	196	42	60	61	33
591	344	109	130	67	38	591	322	87	130	67	38
505	242	84	58	39	61	505	242	84	58	39	61
471	273	100	88	43	42	471	251	78	88	43	42
514	244	80	47	52	65	514	308	80	111	52	65
319	186	66	59	43	18	319	238	66	111	43	18
347	211	87	62	34	28	347	174	1	111	34	28
Summe: 13171	7206	2412	2002	1539	1253	13171	7073	2114	2167	1539	1253
Stimmen in Prozent:		33.47 %	27.78 %	21.36 %	17.39 %			29.89 %	30.64 %	21.76 %	17.72 %

Figure C.3: Manipulation task: after results are calculated.

Participants are asked to evaluate one sheet at a time and decide if a displayed sheet contains real election results or election results that were manipulated by a different participant in the manipulation phase (compare Figure C.4 and C.5). It is randomly assigned with a probability of 50% whether a displayed sheet contains real or manipulated results. In general, the sheets contain very similar information as given in the manipulation task. It comprises the eligible voters, voters, votes for party A, party B, party C and other parties and the overall result of the 30 polling stations, but it does not include the vote share of parties. This is because participants should focus on the plausibility of individual vote counts, instead of the overall vote shares.

Every participant evaluates sheets that contain unfamiliar data, thus no one evaluates the same election results that they previously manipulated. This is secured through the two groups that received different data sets and the ring matching procedure.[2] The real vote counts for the 30 polling stations are random samples from districts of the Canadian federal election in 2008 that fulfill the winning margin requested in the treatment. It is ensured that the subsets of 30 polling stations also fulfill this criterion. Participants are fully informed about the procedures of the experiment.

[2]Participants with odd numbers are assigned to group 1 and participants with even numbers to group 2. The ring matching means that participant 1 evaluates participant 2, participant 2 evaluates participant 3, participant 3 evaluates participant 4 and 4 evaluates participant 1. Each session contains an equal number of participants.

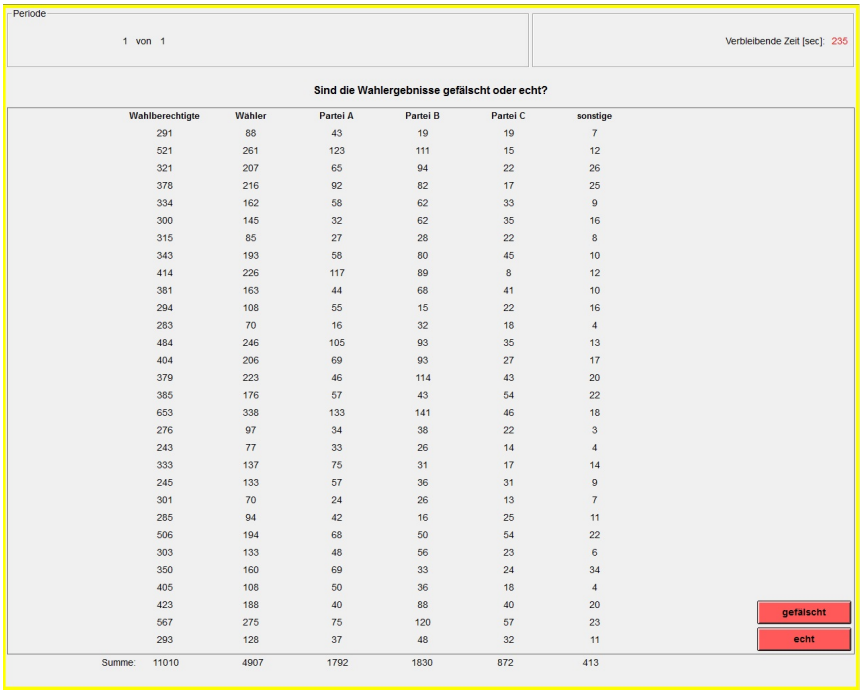


Figure C.4: Evaluation task: real results are displayed for evaluation.

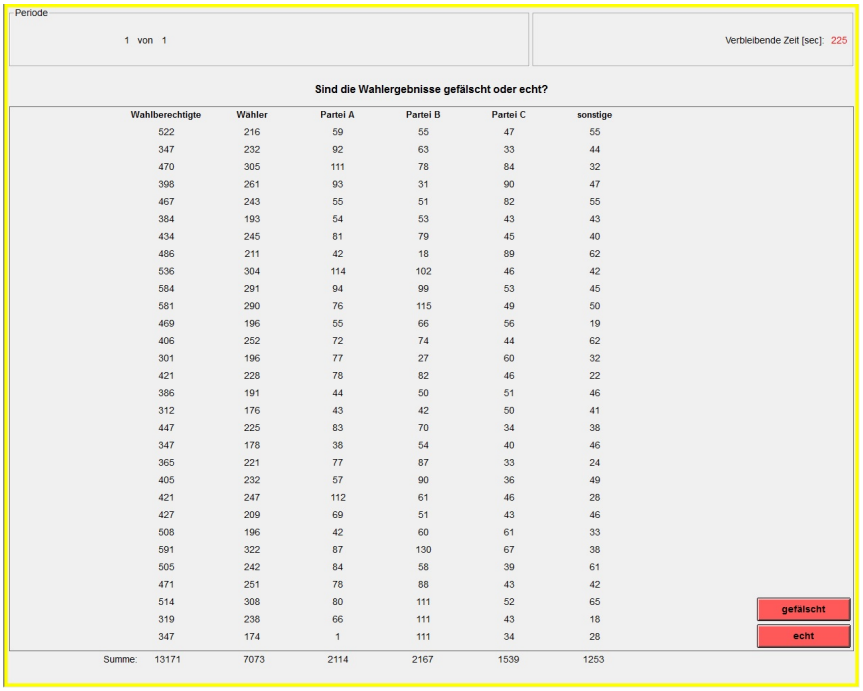


Figure C.5: Evaluation task: (badly) manipulated results are displayed for evaluation.

The process of evaluation and payoffs are very important to ensure the quality of the data and give the right incentives to participants to manipulate inconspicuously. The payoffs for participants are defined as follows: All participants receive 15 points at the beginning of the experiment. They are informed at the introduction that a random mechanism decides which of the three manipulated and evaluated sheets is relevant to determine the final payoff. The random selection of the payoff round is a common strategy in lab experiments to incentivize participants to not change strategies during the experiment. There are strong monetary incentives for participants to perform well as they earn an additional 10 points if they evaluate a displayed sheet correctly and they lose 10 points if their manipulated sheet is identified as such by a different participant. If a manipulated sheet of a participant is not displayed for evaluation, no points are subtracted. For example, if the random mechanism selects the first sheet, then a participant gains 10 additional points if she evaluated the first displayed sheet correctly. Moreover, the participant loses 10 points if her first manipulated sheet is displayed to a different participant and correctly identified as manipulated by this participant. The procedure results in payoffs of 5, 15 and 25. Participants understand that they are evaluating a participant other than the one evaluating them. Thus, there is no incentive to play nice. Before the actual experiment starts, participants have three minutes to familiarize themselves with the features of the experiment using a small example of three polling stations. Each experimental point was converted to 1 €.

C.2 Methods

Most of the digits test use the popular Pearson's χ^2 statistic to identify significant deviations from the investigated digit distribution. Medzihorsky (2015) suggest a latent class analysis that is not based on null hypothesis testing. In this framework the distributional assumption is relaxed. The idea seems promising and further solves problems that arise from large N applications using the χ^2 statistic. However, a replication of the approach using German federal election 2009 shows that it produces very reasonable results when analyzing all polling stations within the country ($N = 88720$), but performs very poorly (it indicates extremely high values for fraud parameters) if the analysis is conducted for the first constituency that contains 291 polling stations.[3] This study uses an experimental approach in that subjects manipulate election return sheets. Within the experimental framework it seems unreasonable to let subjects manipulate more than 100 polling stations without losing motivation and concentration. Unfortunately, this means that the suggested estimation technique for the LD test (Medzihorsky, 2015) is not suitable for the experimental data. Instead, the LD test is based on the popular Pearson's χ^2 statistic, where the fre-

[3]The replication results are available upon request.

quency of the last digits is compared with the expected frequency of the uniform distribution.

C.3 Results

Table C.1: How often subjects loaded sheets for manipulation

Loaded	SuT	T1	T2	T3
0	25	0	2	1
1	0	0	0	3
2	0	3	2	8
3	0	23	22	13

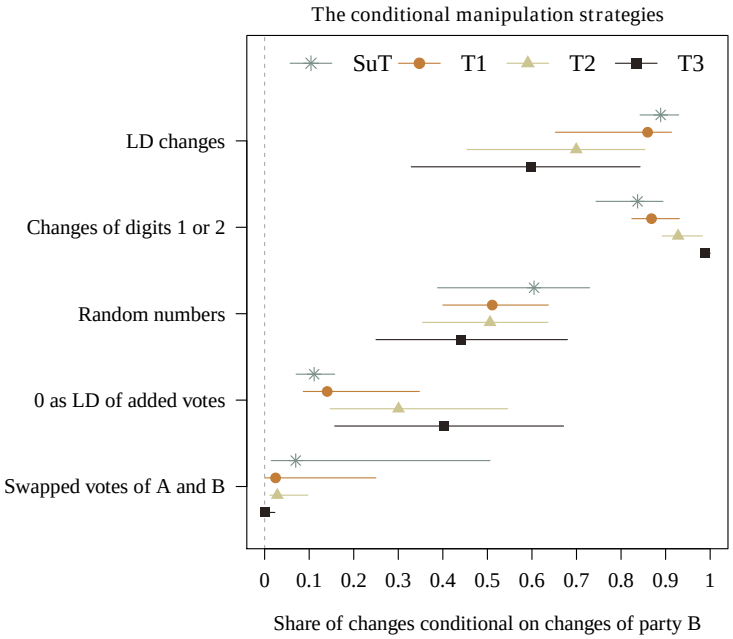


Figure C.6: Conditional strategies, calculated as share relative to changed vote counts of party B for each subject. The line represent subjects between the 25% and 75% percentile. The symbol indicates the median of subjects.

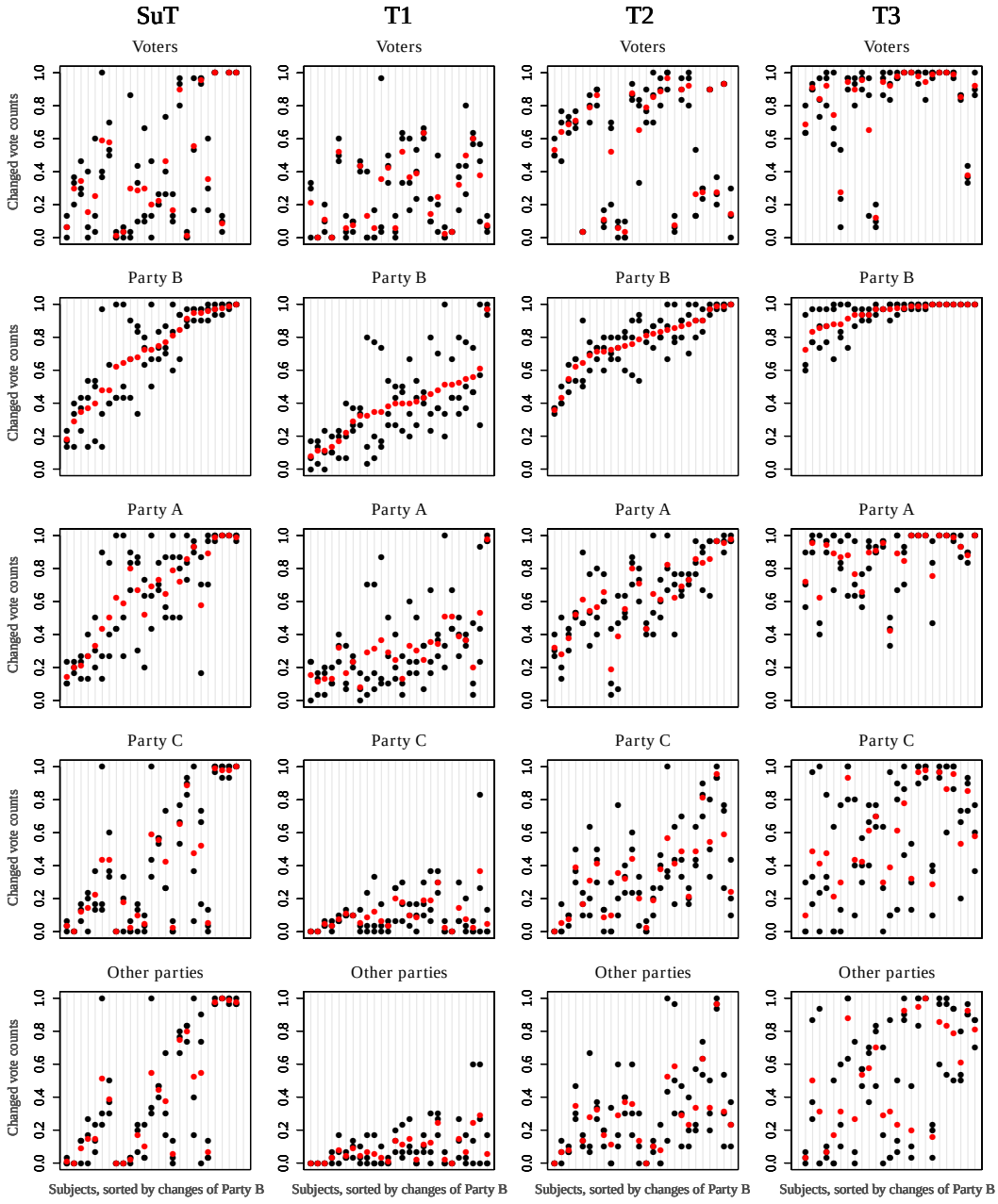


Figure C.7: Manipulation extent by individual subjects. The black dots mark the share of manipulated vote counts for each sheet and the red dot the share across the three sheets that each subject manipulated.

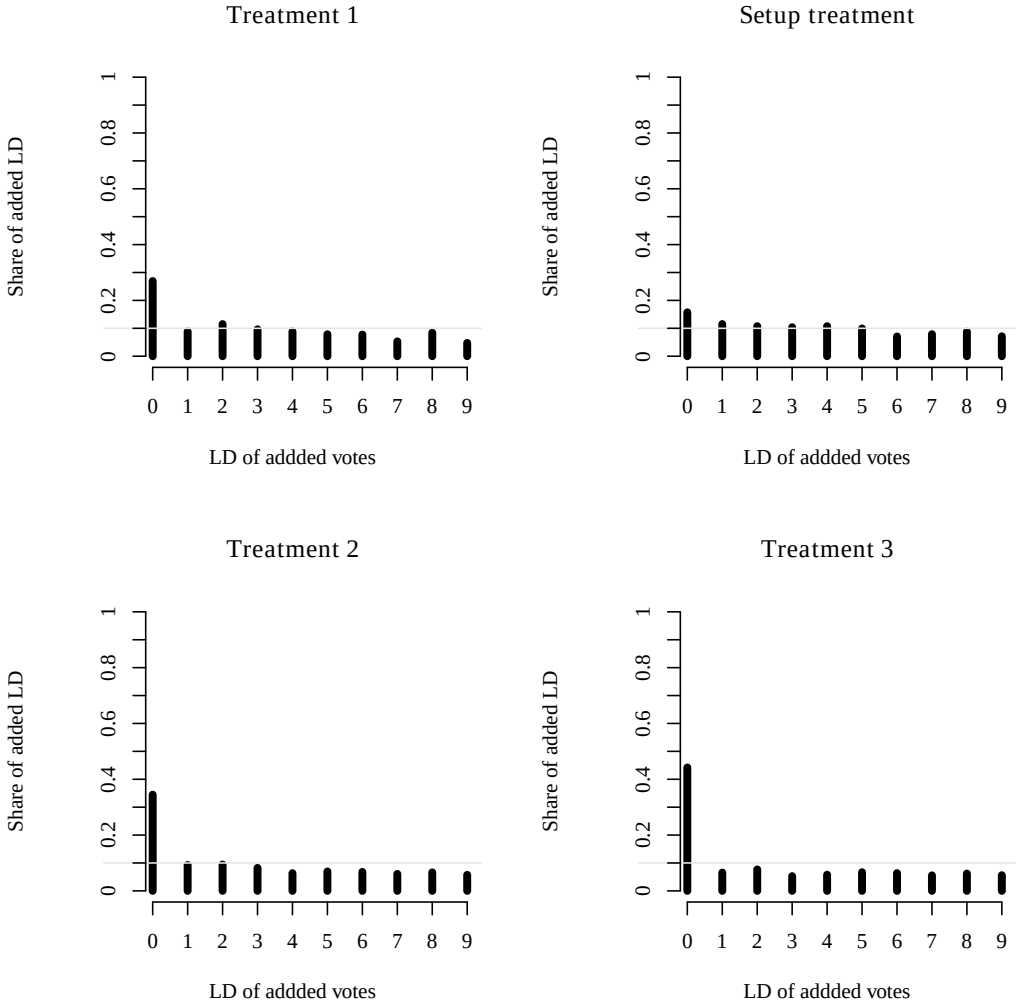


Figure C.8: The LD of added votes to party B. If humans try to make up random numbers, then the added votes to the original vote counts should be a random number. Consequently, the LD of the added votes should be distributed uniformly. Alternatively, humans might have preferences to add particular numbers to the existing vote counts. The gray horizontal line displays the probability of 10% that we would expect for uniformly distributed digits. Instead of making up random numbers, humans might have preferences to add particular numbers to the existing vote counts. The 0 as LD of added votes is overrepresented in all treatments, this is most common in T3 with 44%, and the least with 16% in SuT compared to the expected 10%. Therefore, across subjects the LDs of added votes are not random numbers, especially not if the manipulation is extensive.

Table C.2: The Effect of Manipulation on the Probability of LDs Uniformity: Robustness Checks

	(5)	(6)	(2a)	(2b)
(Intercept)	0.99	1.17	2.68***	1.21*
	(0.70)	(0.69)	(0.80)	(0.48)
Setup treatment	−0.84	−0.72	−2.51***	−1.04*
	(0.63)	(0.61)	(0.76)	(0.44)
Treatment 2	−0.87*	−0.61	−0.99**	−0.89**
	(0.42)	(0.44)	(0.33)	(0.33)
Treatment 3	−0.17	0.26	−0.26	−0.32
	(0.54)	(0.60)	(0.35)	(0.36)
Changes	−0.14			
	(0.95)			
LD changes	−2.23**	−1.99**	−2.39***	−2.40***
	(0.70)	(0.62)	(0.50)	(0.51)
Changes digits 1 or 2		−0.95		
		(0.95)		
Vote swapping	2.45**	2.55**	2.31**	2.42**
	(0.81)	(0.82)	(0.80)	(0.81)
Δ Turnout	−0.90	−0.38	−0.71	−0.93
	(1.97)	(2.03)	(1.88)	(1.89)
Dataset 2	−0.07	−0.17	−0.12	−0.13
	(0.25)	(0.25)	(0.23)	(0.23)
Loaded results >= 2	−0.22	−0.18		
	(0.53)	(0.53)		
Loaded results >= 1			−1.88**	
			(0.69)	
Loaded results = 3				−0.47
				(0.35)
Precision: (phi)	1.91***	1.93***	2.00***	1.94***
	(0.24)	(0.25)	(0.26)	(0.25)
Pseudo R^2	0.30	0.30	0.34	0.31
Log Likelihood	55.78	56.23	58.32	56.60
AIC	-89.56	-90.46	-96.63	-93.19
Num. obs.	102	102	102	102

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

Note: (5) and (6) show that the variables changes and changes of digits 1 or 2 lose their explanatory power if either one is included in a model with LD changes. (2a) and (2b) are comparable to (2), but differ in the specification of loaded results. Loaded results >= 1 means results were at least once loaded (compared to never loaded results) and loaded results = 3 means that they are always loaded. Effects of LD changes and vote swapping do not change, but effects of the treatments and the loaded results differ between specifications.

Bibliography

Alvarez, R. M. and J. N. Katz. 2008. The Case of the 2002 General Election. In *Election Fraud: Detecting and Deterring Electoral Manipulation*, ed. R. M. Alvarez, T. E. Hall and S. D. Hyde. Brookings Institution Press pp. 149–161.

Alvarez, R. M., T. E. Hall and S. D. Hyde. 2008. *Election Fraud: Detecting and Deterring Electoral Manipulation*. Brookings Institution Press.

Bader, M. and H. Schmeets. 2013. "Does International Election Observation Deter and Detect Fraud? Evidence from Russia." *Representation* 49(4):501–514.

Bailey, Delia. 2008. Caught in the Act: Recent Federal Election Fraud Cases. In *Election Fraud: Detecting and Deterring Electoral Manipulation*. Brookings Institution Press.

Baum, Dale and James L Hailey. 1994. "Lyndon Johnson's victory in the 1948 Texas Senate Race: A Reappraisal." *Political Science Quarterly* 109(4):595–613.

Beber, B. and A. Scacco. 2012. "What the Numbers Say: A Digit-based Test for Election Fraud." *Political Analysis* 20(2):211–234.

Benford, F. 1938. "The Law of Anomalous Numbers." *Proceedings of the American Philosophical Society* 78(4):551–572.

Benjamini, Y. and D. Yekutieli. 2005. "Quantitative Trait Loci Analysis Using the False Discovery Rate." *Genetics* 171(2):783–790.

Benjamini, Y. and Y. Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1):289–300.

Boland, P. J. and K. Hutchinson. 2000. "Student Selection of Random Digits." *The Statistician* 49(4):519–529.

Brady, H. E. 2005. "Comments on Benford's Law and the Venezuelan Election.". Unpublished manuscript, Stanford University, January 19, 2005.

Breunig, C. and A. Goerres. 2011. "Searching for Electoral Irregularities in an Established Democracy: Applying Benford's Law Tests to Bundestag Elections in Unified Germany." *Electoral Studies* 30(3):534–545.

Callen, M. and J. D. Long. 2015. "Institutional Corruption and Election Fraud: Evidence from a Field Experiment in Afghanistan." *American Economic Review* 105(1):354–381.

Cantú, F. 2014. "Identifying Irregularities in Mexican Local Elections." *American Journal of Political Science* 58(4):936–951.

Cantú, F. and S. M. Saiegh. 2011. "Fraudulent Democracy? An Analysis of Argentina's Infamous Decade Using Supervised Machine Learning." *Political Analysis* 19(4):409–433.

Carriquiry, A. L. 2011. "Election Forensics and the 2004 Venezuelan Presidential Recall Referendum as a Case Study." *Statistical Science* 26(4):471–478.

Cox, G. W. and J. M. Kousser. 1981. "Turnout and Rural Corruption: New York as a Test Case." *American Journal of Political Science* 25(4):646–663.

Deckert, J., M. Myagkov and P. C. Ordeshook. 2011. "Benford's Law and the detection of election fraud." *Political Analysis* 19:245–268.

Delfino, G. and G. Salas. 2011. "Analysis of the 2004 Venezuela Referendum: The Official Results Versus the Petition Signatures." *Statistical Science* 26(4):479–501.

Di Franco, A., A. Petro, E. Shear and V. Vladimirov. 2004. "Small Vote Manipulations can Swing Elections." *Communications of the ACM* 47(10):43–45.

Dunning, T. 2008. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Research Quarterly* 61(2):282–293.

Dunning, T. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge University Press.

Encyclopædia Britannica Online. 2016. "s.v. election.". Accessed February 27, 2016. **URL:** *http://academic.eb.com/EBchecked/topic/182308/election*

Enikolopov, R., V. Korovkin, M. Petrova, K. Sonin and A. Zakharov. 2013. "Field Experiment Estimate of Electoral Fraud in Russian Parliamentary Elections." *Proceedings of the National Academy of Sciences* 110(2):448–452.

Fischbacher, U. 2007. "z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics* 10(2):171–178.

Fukumoto, K. and Y. Horiuchi. 2011. "Making Outsiders' Votes Count: Detecting Electoral Fraud through a Natural Experiment." *American Political Science Review* 105(3):586–603.

Gans-Morse, J., S. Mazzuca and S. Nichter. 2014. "Varieties of Clientelism: Machine Politics During Elections." *American Journal of Political Science* 58(2):415–432.

Gehlbach, S. and A. Simpser. 2015. "Electoral Manipulation as Bureaucratic Control." *American Journal of Political Science* 59(1):212–224.

Gerber, A. S. and D. P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. WW Norton.

Goldberg, R. 1987. Election Fraud: An American Vice. In *Elections American Style*, ed. J. A. Reichley. The Brookings Institution, Washington, DC pp. 185–194.

Greiner, B. 2004. An Online Recruitment System for Economic Experiments. In *Forschung und wissenschaftliches Rechnen GWDG Bericht 63*, ed. K. Kremer and V. Macho. Gesellschaft für Wissenschaftliche Datenverarbeitung, Göttingen pp. 79–93.

Hall, T. E. and T. A. Wang. 2008. International Principles for Election Integrity. In *Election Fraud: Detecting and Deterring Electoral Manipulation*, ed. R. M. Alvarez, T. E. Hall and S. D. Hyde. Brookings Institution Press pp. 37–49.

Hausmann, R. and R. Rigobón. 2011. "In Search of the Black Swan: Analysis of the Statistical Evidence of Electoral Fraud in Venezuela." *Statistical Science* 26(4):543–563.

Hidalgo, F. D. and S. Nichter. 2015. "Voter Buying: Shaping the Electorate through Clientelism." *American Journal of Political Science* pp. n/a–n/a.
URL: *http://dx.doi.org/10.1111/ajps.12214*

Hill, T. P. 1995. "A Statistical Derivation of the Significant-Digit Law." *Statistical Science* 10(4):354–363.

Hyde, S. D. 2007. "The Observer Effect in International Politics: Evidence from a Natural Experiment." *World Politics* 60(01):37–63.

Ichino, N. and M. Schündeln. 2012. "Deterring or Displacing Electoral Irregularities? Spillover Effects of Observers in a Randomized Field Experiment in Ghana." *The Journal of Politics* 74(01):292–307.

Kalinin, K. and W. R. Mebane. 2011. Understanding Electoral Frauds through Evolution of Russian Federalism: from Bargaining Loyalty to Signaling Loyalty. Prepared for presentation at the Annual Meeting of the Midwest Political Science Association, Chicago, IL.

Katz, J.N. and G. King. 1999. "A Statistical Model for Multiparty Electoral Data." *American Political Science Review* 93(1):15–32.

Klimek, P., Y. Yegorov, R. Hanel and S. Thurner. 2012. "Statistical Detection of Systematic Election Irregularities." *Proceedings of the National Academy of Sciences* 109(41):16469–16473.

Kobak, D., S. Shpilkin and M. S. Pshenichnikov. 2012. "Statistical Anomalies in 2011-2012 Russian Elections Revealed by2D Correlation Analysis." *Physics and Society* . Available at http://arxiv.org/pdf/1205.0741.pdf.

Kossovsky, A. E. 2015. *Benford's Law : Theory, the General Law of Relative Quantities, and Forensic Fraud Detection Applications*. World Scientific.

Leemann, L. and D. Bochsler. 2014. "A Systematic Approach to Study Electoral Fraud." *Electoral Studies* 35:33–47.

Lehoucq, F. 2003. "Electoral Fraud: Causes, Types, and Consequences." *Annual Review of Political Science* 6(1):233–256.

Lehoucq, F.E. and I.M. Jiménez. 2002. *Stuffing the Ballot Box: Fraud, Electoral Reform, and Democratization in Costa Rica*. Cambridge Univ Pr.

Martín, I. 2011. "2004 Venezuelan Presidential Recall Referendum (2004 PRR): A Statistical Analysis from the Point of View of Electronic Voting Data Transmissions." *Statistical Science* 26(4):528–542.

Mebane, W. R. 2006*a*. Detecting Attempted Election Theft: Vote Counts, Voting Machinesand Benford's Law. Annual Meeting of the Midwest Political Science Association, Chicago, IL, April.

Mebane, W. R. 2006*b*. Election Forensics: The Second-Digit Benford's Law Test and Recent American Presidential Elections. Election Fraud Conference, Salt Lake City, Utah.

Mebane, W. R. 2007. Election Forensics: Statistical Interventions in Election Controversies. Prepared for presentation at the Annual Meeting of the American Political Science Association.

Mebane, W. R. 2008*a*. Election Forensics: Outlier and Digit Tests in America and Russia. American Electoral Process Conference, Center for the Study of Democratic Politics, Princeton University.

Mebane, W. R. 2008*b*. Election Forensics: The Second-Digit Benford's Law Test and Recent American Presidential Elections. In *Election Fraud. Detecting and Deterring Electoral Manipulation*, ed. R. M. Alvarez, T. E. Hall and S. D. Hyde. Brookings Institution Press pp. 162–181.

Mebane, W. R. 2010*a*. Election Fraud or Strategic Voting? Can Second-digit Tests Tell the Difference? Summer Meeting of the Political Methodology Society, University of Iowa.

Mebane, W. R. 2010*b*. "Fraud in the 2009 presidential election in Iran?" *Chance* 23(1):6–15.

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

Mebane, W. R. 2012. Second-digit Test for Voters' Election Strategies and Election Fraud. Prepared for presentation at the Annual Meeting of the Midwest Political Science Association, Chicago, IL.

Mebane, W. R. 2015. Election Forensics Toolkit DRG Center Working Paper. Technical report. Working paper for IIE/USAID subaward# DFG-10-APS-UM,"Development of an Election Forensics Toolkit: Using Subnational Data to Detect Anomalies".

Mebane, W. R. and J. S. Sekhon. 2004. "Robust Estimation and Outlier Detection for Overdispersed Multinomial Models of Count Data." *American Journal of Political Science* 48(2):392–411.

Mebane, W. R. and J. Wall. 2015. Election Frauds, Postelection Legal Challenges and Geography in Mexico. Annual Meeting of the American Political Science Association, San Francisco, CA.

Mebane, W. R. and K. Kalinin. 2009. Comparative Election Fraud Detection. Prepared for presentation at the 2009 Annual Meeting of the American Political Science Association, Toronto, Canada, Sept 3–6.

Medzihorsky, J. 2015. "Election Fraud: A Latent Class Framework for Digit-Based Tests." *Political Analysis* pp. 1–12.
URL: *http://pan.oxfordjournals.org/content/early/2015/09/04/pan.mpv021.abstract*

Montgomery, J. M., S. Olivella, J. D. Potter and B. F. Crisp. 2015. "An Informed Forensics Approach to Detecting Vote Irregularities." *Political Analysis* 23(4):488–505.

Myagkov, M. G., P. C. Ordeshook and D. Shakin. 2009. *The Forensics of Election Fraud: Russia and Ukraine*. Cambridge University Press Cambridge.

Myagkov, Mikhail, Peter C Ordeshook and Dimitry Shaikin. 2008. On the Trail of Fraud: Estimating the Flow of Votes between Russia's Elections. In *Election Fraud: Detecting and Deterring Electoral Manipulation*, ed. R. M. Alvarez, T. E. Hall and S. D. Hyde. Brookings Institution Press.

Newcomb, S. 1881. "Note on the Frequency of Use of the Different Digits in Natural Numbers." *American Journal of Mathematics* 4(1):39–40.

Nickerson, R.S. 2002. "The production and perception of randomness." *Psychological Review* 109(2):330–357.

Nigrini, M. 2012. *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*. John Wiley & Sons Inc.

Nyblade, Benjamin and Steven R Reed. 2008. "Who cheats? Who loots? Political competition and corruption in Japan, 1947–1993." *American Journal of Political Science* 52(4):926–941.

OSCE/ODIHR. 2012*a*. France, Presidential Election – 22 April and 6 May 2012: OSCE/ODIHR Needs Assessment Mission Report. Technical report.

OSCE/ODIHR. 2012*b*. Republic of Finland, Presidential Election – 22 January 2012: OSCE/ODIHR Needs Assessment Mission Report. Technical report.

OSCE/ODIHR. 2012*c*. Russian Federation, Presidential Election – 4 March 2012: Statement of Preliminary Findings and Conclusion. Technical report.

Pericchi, L. and D. Torres. 2011. "Quick Anomaly Detection by the Newcomb–Benford Law, with Applications to Electoral Processes Data from the USA, Puerto Rico and Venezuela." *Statistical Science* 26(4):502–516.

Pericchi, L. R. and D. Torres. 2004. "La Ley de Newcomb-Benford y sus aplicaciones al Referendum Revocatorio en Venezuela." *Reporte Técnico no-definitivo 2a. versión: Octubre 01,2004* .

Powell, L.N. 1989. "Correcting for Fraud: A Quantitative Reassessment of the MississippiRatification Election of 1868." *The Journal of Southern History* 55(4):633–658.

Prado, R. and B. Sansó. 2011. "The 2004 Venezuelan Presidential Recall Referendum: Discrepancies Between Two Exit Polls and Official Results." *Statistical Science* 26(4):517–527.

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

Roukema, B. F. 2009. "Benford's Law Anomalies in the 2009 Iranian Presidential Election.". Available at http://en.scientificcommons.org/46955838.

Schedler, A. 2002. "The Menu of Manipulation." *Journal of Democracy* 13(2):36–50.

Shikano, Susumu and Verena Mack. 2011. "When Does the Second-Digit Benford's Law-Test Signal an Election Fraud? Facts or Misleading Test Results." *Journal of Economics and Statistics (Jahrbuecher fuer Nationaloekonomie und Statistik)* 231(5-6):719–732.

Simpser, A. 2013. *Why Governments and Parties Manipulate Elections. Theory, Practice, and Implications*. Cambridge University Press.

Sjoberg, F. M. 2014. "Autocratic adaptation: The strategic use of transparency and the persistence of election fraud." *Electoral Studies* 33:233–245.

The Economist. 2007. "Nigeria: How to steal yet another election.". Published April 18, 2007, accessed January 13, 2016.
URL: *http://www.economist.com/node/9050948*

Wand, J. N., K. W. Shotts, J. S. Sekhon, W. R. Mebane, M. C. Herron and H. E. Brady. 2001. "The Butterfly did it: The Aberrant Vote for Buchanan in Palm Beach County, Florida." *American Political Science Review* 95(4):793–810.

Weidmann, Nils B. and Michael Callen. 2013. "Violence and Election Fraud: Evidence form Afghanistan." *British Journal of Political Science* 43(1):53–75.

Ziblatt, Daniel. 2009. "Shaping Democratic Practice and the Causes of Electoral Fraud: The Case of Nineteenth-Century Germany." *American Political Science Review* 103(01):1–21.