



ELSEVIER

Contents lists available at ScienceDirect

# Electoral Studies

journal homepage: [www.elsevier.com/locate/electstud](http://www.elsevier.com/locate/electstud)

## A systematic approach to study electoral fraud<sup>☆</sup>

Lucas Leemann<sup>a,\*,1</sup>, Daniel Bochsler<sup>b,1</sup><sup>a</sup> Department of Political Science, Columbia University, International Affairs Building, 420 W 118th Street, New York City, USA<sup>b</sup> University of Zürich, NCCR Democracy, Affolternstr. 56, CH-8050 Zurich, Switzerland

### ARTICLE INFO

#### Article history:

Received 4 January 2013

Received in revised form 24 December 2013

Accepted 26 March 2014

Available online 13 April 2014

#### Keywords:

Electoral fraud  
 Electoral forensics  
 Benford's law  
 Referendums  
 Switzerland

### ABSTRACT

Integrity of elections relies on fair procedures at different stages of the election process, and fraud can occur in many instances and different forms. This paper provides a general approach for the detection of fraud. While most existing contributions focus on a single instance and form of fraud, we propose a more encompassing approach, testing for several empirical implications of different possible forms of fraud. To illustrate this approach we rely on a case of electoral irregularities in one of the oldest democracies: In a Swiss referendum in 2011, one in twelve municipalities irregularly destroyed the ballots, rendering a recount impossible. We do not know whether this happened due to sloppiness, or to cover possible fraudulent actions. However, one of our statistical tests leads to results, which point to irregularities in some of the municipalities, which lost their ballots: they reported significantly fewer empty ballots than the other municipalities. Relying on several tests leads to the well known multiple comparisons problem. We show two strategies and illustrate strengths and weaknesses of each potential way to deal with multiple tests.

© 2014 Elsevier Ltd. All rights reserved.

## 1. Introduction

Election fraud is not necessarily confined to young and fragile democratic states. While a large part of the election fraud literature has looked at democratizing or non-democratic countries, this article investigates fraud that

might have occurred recently in one of the oldest democracies,<sup>2</sup> and aims at presenting a forensic toolbox for detection of manipulations of ballots and the vote count. This is done based on a new, systematic empirical approach. It is built on two theoretical insights on election or referendum fraud: first, election fraud or misconduct can occur in many different instances of the election process, and in many different ways. Therefore, electoral forensics are strongest when a number of different tests are conducted. Second, each type of fraud is rooted in a specific micro-foundation, which should inform the empirical tests. This has important implications for the analysis of the integrity of elections or referendums. This approach is applied to a specific example: on February 13th 2011 the people in the Swiss canton of Berne voted on a motor tax (*Motorfahrzeugsteuer*). The very close outcome sparked hope that a recount might change the final outcome, which was granted after a legal battle. This is when the public

<sup>☆</sup> We thank Kurt Nuspliger (Staatschreiber, Kanton Bern) for answering a long list of questions regarding the exact procedure and the cantonal rules pertaining ballot storage and vote counts. The interview was about the administrative practice and we did not discuss fraud allegations. We thank Werner Seitz (Bundesamt für Statistik) for supplying us with additional data. We have received helpful comments from Sebastian Fehler, Andrew Gelman, Oliver Strijbis, Marc Bühlmann, Julian Wucherpfennig, Hanspeter Schaub, and Christian Rubba. An earlier version was presented at the annual meeting of the Swiss Political Science Association in February 2012. Lucas Leemann gratefully acknowledges the financial support by SAGW (Reisezuschuss).

\* Corresponding author.

E-mail addresses: [lt2108@columbia.edu](mailto:lt2108@columbia.edu) (L. Leemann), [daniel.bochsler@uzh.ch](mailto:daniel.bochsler@uzh.ch) (D. Bochsler).

<sup>1</sup> Both authors contributed equally to the paper.

<sup>2</sup> See also Cox and Kousser (1981) and Alvarez and Boehmke (2008).

learned that almost one in twelve municipalities had violated the electoral law and destroyed the ballots instead of retaining them for one year (Nuspliger 2011). We ask whether this was pure carelessness, or possibly the attempt to hide electoral misconduct. Our forensic tests show that those municipalities that have destroyed the ballots have reported surprisingly few empty ballots in the electoral results. This paper applies several election forensic approaches to investigate the suspicion that results in the Bernese municipalities that have lost their ballots might have been manipulated. To do so, it makes several suggestions how the electoral forensic methods might be applied in a theory-driven way.

A quickly growing literature has developed two types of tools of vote forensics (e.g., Filippov and Ordeshook, 1996; Breunig and Goerres, 2011). One part of the literature discusses whether the analysis of single digits of the reported electoral results at the ward level can reveal that these numbers are based on the actual count of the votes, or whether they have been altered, relying on Benford's law on the frequency distribution of digits in numbers. A second literature investigates the plausibility of electoral results from wards, and is based on circumstantial information. This paper, first, provides a clear framework in which electoral forensics are carried out and to move away from ad-hoc hypotheses testing towards a more firmly rooted set of micro-foundations. This can help to derive much more precise empirical implications of fraud. Second, it considers that usually election fraud does not occur in a whole country, but is more likely in particular electoral wards (Alvarez and Boehmke, 2008).<sup>3</sup> We rely on models that suggest how election outcomes look in a fair election. These models can be tested on those municipalities where we do not expect fraud to have happened, and we can compare the results to municipalities with possible manipulations. Furthermore, we argue that different forms of manipulation vary in their likelihood, and tests of fraud should start with the formulation of a micro-logic of fraud (see also Beber and Scacco (2012)).

First, we lay out the different potential ways how fraud could occur in these votes. After deriving a micro-logic we connect each of the potential fraudulent acts with a specific tailored test statistic. Finally, we carry out all four derived tests and show how one can combine the different tests into an overall assessment. Substantively, we first investigate the plausibility of the electoral result and the number of invalid and empty ballots, relying also on historical vote data. Second, we rely on Benford's law, focusing on the last digit of the vote figures. We test whether voting results from those thirty municipalities which are unable to produce the ballots show implausible distributions of the last digit.

The next section discuss the literature on electoral fraud, and introduces the referendum of February 13th 2011. Section 3 proceeds with a discussion of statistical methods to detect electoral fraud. We lay out a number of plausible ways in which manipulation could have occurred which

leads to the formulation of four distinct hypotheses. The results of these tests are presented and discussed in section 4. Finally, the concluding remarks are in section 5.

## 2. A Systematic approach for electoral forensics

Electoral fraud occurs in many different ways. The variety of forms of fraud reflects the long list of criteria that need to be established, so that elections can be considered free and fair. Some forms of misconduct occur before or during the election campaign, others on election day or during the vote count; some in the central election administrations, others decentralized (Schedler, 2002). This should be reflected in the approaches to prevent and detect fraud. On election day, the local electoral commissions might invalidate or remove ballots, stuff the ballot box with irregular ballots, change the content of the ballots, miscount the expressed votes, or alter the figures ex-post.

This variety of misconduct is reflected in a variety of actors and forms of behavior related to it, and most of all to very diverse approaches how fraud might be prevented or detected. While the prevention of fraud relies on instruments such as multi-partisan compositions of election commissions, transparency of the election process, exit polls, or election observers,<sup>4</sup> the post-hoc detection of possible fraud (election forensics) is less developed. One method, which has gained increasing attention in the literature, relies on the statistical properties of the distribution of digits in aggregated election results, based on Benford's law (e.g., Mebane, 2008, 2010b, 2011; Deckert et al., 2011; Beber and Scacco, 2012). Benford's law is suited, however, only to detect one very particular, and not always very likely form of fraud.

Systematic forensic approaches should be interested in a variety of traces, which result from the specific forms of electoral misconduct one wishes to detect. This has several implications. First, forensic methods should be based on micro-logics of fraud, which are plausible in the specific setting where the election takes place. Therefore, we first need to gain knowledge of the electoral process, as only this allows us to identify the leeway that involved actors have to commit fraud, and possible logics of fraud.<sup>5</sup> Second, we can only rule out fraud, once we investigated all possible instances and forms of it. This cannot be fully implemented in practice, as some forms of fraud might not be detectable.<sup>6</sup> Still, it is worth to consider the most important instances where fraud might have occurred. Third, the analysis of the context of the election should also discuss the difficulty and effectiveness of different forms of fraud, in order to identify those most likely to occur. A set of hypotheses, addressing the traces of fraud, should therefore be derived from this discussion of micro-logics of fraud, and from the discussion of their relative likelihood. Following these suggestions, we

<sup>4</sup> See, among others, Hyde and Marinov (2008) and Mozaffar and Schedler (2002).

<sup>5</sup> For a nice exception in the literature see the paper by Myagkov et al. (2005) where they employ different tests and approaches.

<sup>6</sup> And with too many parallel tests, we would most likely find some positive results, even at the absence of fraud.

<sup>3</sup> See also Myagkov et al. (2008: 195). In contrast, in our model, the 'fraud suspicion' variable is exogenous to the model.

next move to a discussion of the referendum of February 13th 2011 in the Swiss canton of Berne, and the election and referendum authorities in this canton.

### 2.1. A vote on taxes – the controversial vote on motor vehicle taxes

On February 13th 2011, the people of the canton of Berne were called to vote on the amendment of the law on motor vehicle taxes. The vote was an optional referendum, where two proposed amendments opposed each other.

The canton of Berne allows the people to vote in such optional referendums. Generally, every new cantonal law and every amendment of a cantonal law that is passed by parliament is voted on in a cantonal referendum, if 10,000 voters (out of some 700,000) demand so. In 1993, the canton introduced new referendums with people's amendments (*Referendum mit Volksvorschlag*). Now, committees cannot only oppose a law or an amendment of the parliament, but also propose an alternative bill, which is voted on. Subsequently, voters can choose between three options on their ballot: the one proposed by the parliament, the alternative proposition by the referendum committee ("people's amendment"), and the status quo. The

**Table 1**  
Reported vote outcome.

	Yes	No	Empty
Parliament bill	172,427 (49.01%)	154,792 (44.00%)	24,597 (6.99%)
People's amendment	166,860 (47.43%)	164,325 (46.71%)	20,631 (5.86%)
	Parliament bill	People's amendment	Empty
Tie-break Question	165,614 (47.07%)	165,977 (47.18%)	20,225 (5.75%)
Turnout: 49.4%			

introduction of this new option has been accompanied by another important change to the voting procedure for three-option referendums. On the ballots, both amendments (or new laws) are separately set in opposition to the status quo. An additional question asks for the voters' preferences between both reform options (see Fig. 1).

If either of both reform options tops the status quo, it wins. If both reforms are favored over the status quo, the reform option that beats the other proposal will be enacted (see also Bochsler, 2010). The referendum-with-people's amendment has substantially increased the number of multi-option referendums.

The amendment of the motor vehicle tax bill, as proposed by the parliament of the canton of Berne, foresaw changes of the motor vehicle taxes, which would have benefitted low-emission vehicles, and taxed high-emission vehicles more heavily. This was opposed by a people's amendment, which was proposed by a committee formed around car dealers and supported by the Swiss People's Party (SVP). Their alternative bill foresaw a general decrease of the motor vehicle taxes. Both amendments obtained a narrow majority of approvals, so that the tie-break question was decisive for which of the two versions would become law. This tie-break question was decided by a narrow margin, with 165,977 to 165,614 votes, in favor of the people's amendment (Table 1).

This sparked hope that a recount might change the final outcome. After a legal battle said recount was ordered, due to the narrow result.<sup>7</sup> This is when the public learned that almost ten percent of the municipalities had violated the election laws by destroying the ballots instead of retaining them for one year. 29 out of 30 municipalities, which have lost the ballots, declared that they destroyed them due to misfortunes, or communication mistakes Nuspliger (2011).<sup>8</sup> The chancellor of the 30th municipality, Oberwil bei Büren, had given a very similar declaration to the media

**Stimmzettel für die kantonale Volksabstimmung vom 13. Februar 2011**  
**Bulletin de vote pour la votation cantonale du 13 février 2011**

**Gesetz über die Besteuerung der Strassenfahrzeuge**  
**Loi sur l'imposition des véhicules routiers**

1 Wollen Sie die **Vorlage des Grossen Rates** annehmen?  
Acceptez-vous le **projet du Grand Conseil**?

Antwort: ja oder nein  
Réponse: oui ou non

**nein**

2 Wollen Sie den **Volksvorschlag** annehmen?  
Acceptez-vous le **projet populaire**?

Antwort: ja oder nein  
Réponse: oui ou non

**ja**

Die Fragen 1 und 2 können je mit «Ja» oder «Nein» beantwortet werden.  
Vous pouvez répondre aux questions 1 et 2 par «oui» ou par «non».

**Stichfrage**  
Für den Fall, dass sowohl die Vorlage des Grossen Rates als auch der Volksvorschlag angenommen werden.

**Zutreffendes ankreuzen**  
**Mettez une croix dans la case qui convient**

so/ainsi:

**Question subsidiaire**  
Si les deux projets sont acceptés.

3 Soll die **Vorlage des Grossen Rates** (Vorlage GR) oder der **Volksvorschlag** in Kraft treten?  
Lequel de ces deux textes doit entrer en vigueur, le **projet du Grand Conseil** (projet GC) ou le **projet populaire**?

Vorlage GR  Volksvorschlag

Bei Frage 3 darf nur ein Feld angekreuzt werden; sonst gilt die Frage als nicht beantwortet.  
Quant à la question 3 veuillez cocher une seule case, car sinon, la réponse est réputée non valable.

Vom Stimmausschuss auf der Rückseite abstempeln lassen  
Veuillez faire timbrer au verso par le bureau électoral

162 278

**Fig. 1.** Reprint of a ballot (provided by referendum committee). Ballot is written in German and French and contains three questions. Question 1 asks if one supports the parliament bill, question 2 asks if one support the people's amendment, and question 3 asks what should become law (parliament bill or people's amendment) in case both receive more than 50% of the votes.

<sup>7</sup> The Court ruling only refers to the narrow result, and does not name any irregularities, which would motivate a recount. *Urteil des Verwaltungsgerichts (Verwaltungsrechtliche Abteilung) vom 22. Juni 2011 i.S. X. und Y. gegen Kanton Bern* (VGE 100.2011.69/100.2011.86).

<sup>8</sup> According to this special report, there were a variety of reasons for the 'loss' of the ballots. The municipality of Habkern claims that they had a new city manager and he was not aware of the proper procedure. The municipality of Ringgenberg claims to have stored the old ballots in the wrong box. Finally, the administration of Alchenstorf was doing some spring cleaning and the ballots were unfortunately thrown out by an apprentice.

in August 2011: Allegedly, he had destroyed the ballots by mistake in early March (Sansoni, 2011a). Only three weeks later, he declared that he found the destroyed ballots again (Sansoni, 2011b).

In this paper we perform a number of tests that would allow us to distinguish expected pattern and unexpected patterns. It is surprising for outside observers that close to ten percent of the municipalities violate electoral code and destroy ballots.

## 2.2. The administration of referendums in the Canton of Berne

After ballots have been lost, not only the result of the before mentioned referendum cannot be verified. Possibly even more important, if ballots are lost or destroyed after referendums, this prevents any transparency in the vote counting process, and the possibility to verify that the vote count is accurate, in general. This evokes questions about the management of referendums in the canton of Berne, and whether the counting procedures might allow electoral fraud. The organization of referendums is heavily decentralized. Referendums are administrated and counted at the level of 372 wards, which almost match the 383 municipalities of the canton (numbers of 2011).<sup>9</sup> Many of these wards are small, 57% count less than 1000 registered voters, and less than 6% are larger than 5000 voters. In large municipalities, precision balances are used to count the ballots, instead of counting them by hand, but they are only allowed if they allow a higher degree of reliability than human counting, and if they do not involve any rounding of the resulting number of votes.

Detailed knowledge about the counting process in practice is not available, not at least because this process is heavily decentralized. Local all-party committees are in charge of the administration. They are composed of non-professional members, and often supported by the professional staff of the municipal administration. Even within the same canton, there are important differences. The local electoral committees are elected locally, and usually multi-partisan, but their composition is not only unknown to the cantonal authorities, but even the rules of their composition vary: for instance, some municipalities oblige their citizens to be part of the electoral committee, others not, and some municipalities stronger rely on professionals in the vote count.<sup>10</sup>

The supervision of the local elections and referendum administration is exercised through the cantonal authorities, especially through the offices of the (elected) district governors. But they do not regularly control the vote count, especially there are no spot checks, and the election administration at the municipal level is widely a matter of trust in local electoral committees. The cantonal authorities stress, however, that the high frequency of elections and referendums (there are usually around 5 elections or referendums days per year) helps

establish a professional routine, even in non-professional committees.<sup>11</sup>

Irregularities in the vote count are detected, if the results appear implausible, e.g. if the number of votes appears conspicuously high or low, and there are routine plausibility checks by several instances. At the absence of a proper control, we argue that other irregularities or fraud would remain undetected. Apart from the high level of general trust in the accuracy of the process, the main protection against fraud is the law, which prescribes that at no instance of the counting process, the ballots are in the hands of only one person. While the cantonal authorities cannot imagine that this rule is ever infringed on, there are no checks of the counting process. The acceptance of elections and referendums is, hence, a matter quasi-professional routine in a (non-professional) militia system and trust. Blind, or maybe even naive trust? There is a series of limited incidents, that show that the formal rules of democracy in Switzerland are occasionally infringed by singular actors. Occasionally, cases where candidates cast ballots on behalf of fellow citizens, using the means of the postal vote, come to court. Violations of the voting procedures can also be observed in the highest authorities, e.g. the national parliament. Its first chamber (the National Council) needed to improve its electronic voting system, after a MP was observed to cast a vote on behalf of his seat neighbor in absence.

Finally, going back half a century there is a court ruling with regard to fraud in the counting process. In the canton of Berne, in the municipality of Wimmis (1734 inhabitants), in an election in 1956, names were crossed out from the ballots, most plausibly by one member of the election committee. While the counting process usually evolves in teams of two persons, one member of the committee might have hindered his colleague from controlling the process, using some of the ballots to screen his actions. Given that the counting process evolved in a chaotic matter, many details could not be accurately establish by the court. Apparently, committee members had also consumed alcohol during the counting process, and this apparently infringed the regularity of the process (Wylter, 2011). Smaller errors occur regularly. Municipal administrations occasionally confuse the number of 'yes' or 'no' votes, and some electoral committees do not know the correct procedure. Investigating the accuracy of the referendum on motor vehicle taxes, the Administrative Court cites even one electoral committee which did not know how to fill in the result sheets.<sup>12</sup> This concern is even more important as there seem to be larger differences in the handling of invalid votes, which seems only to be loosely regulated and harmonized across the 26 Swiss cantons,<sup>13</sup> although there

<sup>9</sup> Only a few very small municipalities are merged to larger counting wards.

<sup>10</sup> Information collected by Miriam Hänni and Marc Bühlmann.

<sup>11</sup> Practical information about the administration of elections and referendums in practice relies on an interview with the State Chancellor of Bern, Kurt Nuspliger, December 20th 2011. This interview was solely about the administrative practice and we did not discuss any fraud allegations.

<sup>12</sup> Court decision; *Urteil des Verwaltungsgerichts (Verwaltungsrechtliche Abteilung)* June 22 2011 i.S. X. und Y. gegen Kanton Bern (VGE 100.2011.69/100.2011.86), page 16.

<sup>13</sup> *NZZ am Sonntag*, 11.20.2011, "Bund will 33,000 ungültige Wahlzettel untersuchen" (No. 47, page 11).



is no information with regards to the local practice. Such occasional evidence does not show any regular fraud, but that the counting process is not very well controlled, and manipulations are possible.

For the referendum of February 13th 2011, no evidence or suspicion of fraud, which would explain the destruction of the ballots in 30 municipalities, was made public. However, the loss of the referendum ballots comes as a surprise, and the statements made in the investigation report about the reasons for the destruction of the ballots, jointly with the apparently wrong (and later corrected) statements in the press, undermine our trust in the counting process. Local committees might have a particular motive in losing the ballots if such a recount was to reveal irregularities in the counting process.<sup>14</sup>

### 3. Detecting fraud

How can we detect electoral fraud? The answer to this question depends on the type of committed fraud. Lacking access to the proof (the ballot papers), researchers have started to develop statistical methods to detect irregularities in the reported election results, which might be due to illegitimate manipulations.

Fraud can occur in almost all steps of the election process, and in many different ways, and each way requires its own methodology to detect it. Note that the distinction of acceptable and illegitimate practices has changed over time, and varies across countries and regions. Today, democracies usually consider vote buying illegal and illegitimate, while practices of patronage, which involve violations of the vote secrecy are more widespread.<sup>15</sup>

In this paper, we focus at the level of the ward, and on the role of the local electoral committee, i.e. the vote count and the reporting of the electoral result (in Switzerland this is mostly the municipality level). Fraud at the ward level can occur by altering individual ballots after they have been cast, invalidating valid ballots (or validating invalid ones), or forging the ballot return sheet and changing the numbers.

In general, there are two ways to go about detecting electoral fraud. We focus on the returns at the lowest levels possible and we try to compare outcomes with expectations. The origins of these expectations distinguish the two instruments we have. First, we may rely on ecological information. Knowing the political structure of a village may allow us to predict the voting pattern we should observe (Alvarez and Boehmke, 2008). This approach relies on regression style models based on a subsample where we can (with large confidence) outrule fraud.

Second, we can focus solely on the return sheets (the reported numbers). We compare these figures not with

other returns but with a theoretical distribution of digits. As it turns out, our interest will lie mostly in the last digits which are inconsequential for the outcome but can be an invaluable source to detect fraud. The basic idea is that when someone makes up numbers they fail to produce numbers that are truly random in the way they would be in a truly fair election or vote. But before diving into the empirical tests we derive a number of micro-logics which provide the micro-foundations.

#### 3.1. The micro-logic of fraud

We do not know what happened on February 13th 2011. However, a rich literature of election research provides us with different models that help to predict the outcome of the referendum of the 13 February 2011. We can test these models for optional referendums in the canton of Berne, and we can test whether there were any irregularities in the results of the 30 municipalities that lost their ballots. Therefore, we construct several fictitious scenarios of how an election committee might have manipulated the ballots – each of which require a different degree of criminal energy.

It is of central importance that an investigation is based on a micro-logic of how such potential fraud occurs. We derive four different possibilities and show how we can test for each of them. The derivation is guided by findings of the criminology literature on white-collar crimes (Bannenberg and Jehle, 2010). This body of theories is often referred to as the *fraud triangle* (Cressey, 1980) and regards the likelihood of fraud to depend on opportunity, motivation, and rationalization. Hence, we focus on the effectiveness and severeness of fraud (directly and inversely related to rationalizability) and the amount of criminal energy necessary (motivation) to commit such fraud.

The first potential fraud form we highlight is specific to referendums with tie-break questions. The easiest way to falsify the Bernese ballot is to add a cross on the empty tie break question, while for the other referendum questions, a full word needs to be added. The tie-break question is at least as important as the other two questions on the ballot, and it is often decisive for the outcome of the referendum. There is considerable potential for manipulation, as voters frequently miss to correctly fill in such a ballot (see Fig. 1) and leave the last question out, as a YES and a NO (or vice versa) seems to imply a clear preference.<sup>16</sup> But despite two clear answers to the two proposals at stake, the voter is obliged to say which one she prefers if both obtain a popular majority. The first manipulation occurs when officials fill in left-out tie break responses. Despite this being a fraudulent act it is not hard to see how an official could actually believe to be doing something good as she is not tempering with the intention of the voter. Manipulating empty fields in the tie-break question is also the easiest way to manipulate the Bernese ballot, as only one cross needs to be added to the empty field, while for the other referendum questions, a full word has to be added -

<sup>14</sup> It is heavily implausible that all 30 municipalities have coordinated the destruction of the ballots, and/or done so for the same motive. Carelessness might be an important reason in some of the municipalities, but we cannot exclude that others destroyed the ballot to hinder a recount.

<sup>15</sup> See Kitschelt and Wilkinson (2007: 15–9). A relevant part in the democratic development of Imperial Germany is the absence of the secret ballot and the opportunities to 'bread lords' (employers) to threaten voters into voting differently (e.g., Ziblatt, 2009; Anderson, 2000).

<sup>16</sup> This has also led to legislative action in the past where a part of the Social Democrats of the cantonal legislature demanded a change of the ballot structure (Wasserfallen, 2011).

yielding a higher risk because the different handwritings might be detected. This subjectively least severe form of fraud requires little to no criminal energy.

A second form of fraudulent behavior can be observed when officials fill in empty response fields for the first two proposals. This is more severe because unlike the first category this cannot be justified by trying to help the voter express her preferences. The third form of manipulation requires more criminal energy and is found when an official changes the response on the ballot. This is what happened in the described fraud case in Wimmis (see page 8). This not only requires more criminal energy but is also more severe as it is an act that directly contradicts the voter's choice. A fourth and final possibility is to simply misreport the results of their ward and report different totals to the next bureaucratic level. This blunt contempt of the voters' preferences is the strongest form of fraud in terms of severeness and criminal energy.

These four potential forms lead to a number of hypotheses which can be tested. The first three forms can be tested with a correlational approach in which we specify a parametric model which predicts an outcome variable (e.g. the number of empty ballots) and we include an indicator variable which takes on the value '1' for those municipalities that are at the center of this investigation. Hence, we compare the municipalities which kept all ballots to those that did not and see if they systematically deviate. The fourth form of fraud can be tested by relying on Benford's law which allows under certain circumstances to discriminate between naturally evolving numbers and made-up reported numbers.

**Hypothesis 1.** The number of empty ballots for the tie break question is lower for the municipalities which "lost" the ballots than for those which were able to produce the ballots for a recount when controlling for other covariates.

**Hypothesis 2.** The number of empty ballots for the two proposal questions is lower for the municipalities which 'lost' the ballots than for those which were able to produce the ballots for a recount when controlling for other covariates.

Some municipalities which lost their ballots might have altered the ballots, or written in new content (in any of the three referendum questions). Therefore, in lost-ballot-municipalities, the aggregated results should deviate from the expected results. As manipulations might have occurred in different directions, we expect that the results of the lost-ballot-municipalities are more difficult to explain, compared to other municipalities.

**Hypothesis 3.** The variance of the regression error of the municipalities which lost ballots is higher than the variance from the other municipalities.

Finally, to test for the most severe form of electoral fraud, we perform a test which is able to detect made up numbers and should indicate fraud if the 30 municipalities reported phony digits.

**Hypothesis 4.** The distribution of the last digit of the reported yes and no votes does not follow the theoretical distribution (Benford) for those municipalities which 'lost' the ballots.

In terms of assessing the likelihood we rely on *rationalizability* and *criminal energy*. We operate under the prior that the behavior described in [Hypothesis 1](#) and [2](#) is more likely than what is underlying [Hypothesis 3](#). The least likely micro-logic is captured in [Hypothesis 4](#). Given that empty tie-break questions can be perceived as being left out by mistake, but are still important (rationalizability), even though there is very little criminal energy necessary for altering them, one can argue that this is the most likely form of fraud. On the other hand, blatantly misreporting the vote totals is logistically difficult to do (as there are more than one person observes the vote totals prior to submitting them) as well as it is hard to rationalize. In the following two paragraphs we lay out how we can test these four hypotheses. Note, that the hypotheses assume fraud and if we eventually reject the null hypotheses that would constitute circumstantial evidence for irregularities.

### 3.2. Detecting different forms of fraud

#### 3.2.1. Ecological approach to test the first three hypotheses

First, we predict the referendum result for each ward (i.e. mostly identical with municipalities), and we analyze the deviations from this prediction. We expect that the deviations from the expectations should be most pronounced for the 30 municipalities, which lost their ballots, as explained in hypotheses 1 to 3. The three hypotheses relate to different aspects of the election results (dependent variables): [Hypothesis 3](#) relates to the accuracy of the model prediction (unexplained variance of the yes/no votes), while hypotheses 1 and 2 relate to the number of empty ballots.

The literature of election research provides us with different models that help to predict the outcome of the referendum in February 2011. We can test these models for this particular and several other optional referendums in the canton of Bern, and we can test whether there were any irregularities in the results of the 30 municipalities that lost their ballots. As we have constructed several fictitious scenarios how an election committee might have manipulated the ballots, each of them requires a different effort to test whether a manipulation might have occurred.

Three types of control models can be helpful to predict the referendum results in an optional referendum. The first model (*interdependence of referendum questions*) states that the answers to the three referendum questions on the same ballot are not independent from each other. On the one hand, certain voters who reject both proposed amendments of the law might renounce from answering the tie-break question on the ballot. On the other hand, voters who reject one of the two bills might not answer the tie-break question, because they misunderstand the meaning of the tie break question, and do not realize that everybody is entitled to answer it.<sup>17</sup> Also, certain voters might be more

<sup>17</sup> The ballot states that the tie-break question applies if the two amendments should both be accepted. Voters might misunderstand this statement, and assume that this applies to individual ballots. Hence, they might not answer the tie-break question, if they rejected either of the two amendments (see also [Wasserfallen, 2011](#)).

inclined to leave several of the questions on the ballot unanswered, so that the number of empty fields should correlate on the ballots, and within the municipalities. Therefore, we have expectations about correlations between the results of the three questions on the same ballot.

The second model (*historical model*) states that there are local, idiosyncratic characteristics that might explain parts of the results, and these aspects should be constant in all optional referendums in the last few years. Especially, if certain voters repeatedly cast empty ballots, then the number of empty ballots should correlate across referendums. The third model (*party model*) argues that referendum results can be explained with the partisan composition of the electorate. All optional referendums considered in this study were politicized along party divides. The referendum committees, which are proposing the people's amendment, are often formed or at least heavily supported by political parties. Therefore, the party preferences of the electorate are an important predictor of referendum results. Each of these models works for different types of dependent variables, and therefore, each can only be applied to two hypotheses.

Our set of further control variables includes the language spoken in the municipality (binary indicator variable for French, as opposed to German), and the size of the electorate in the municipality (we use the logarithm of the number of registered voters),<sup>18</sup> which should also control for possible population effects.

The first set of hypotheses (*Hypothesis 1 and 2*) refers to the mean number of empty votes registered per municipalities. We expect that possible manipulation might decrease the number of empty ballots. Therefore, we rely on models that explain the mean share of empty ballots as a percentage of all ballots cast in the referendum in the particular electoral ward. We estimate the models with Goodman regression models for ecological data (Goodman, 1959). These are based on OLS models with robust standard errors, and assume linear effects. As using OLS on fractional data comes with a certain cost, we also rerun the models in Tables 4 and 5 while relying on a fractional logit model as described in Papke and Wooldridge (1996). As our control models are solely aimed at giving accurate predictions of the outcomes, we are indifferent to whether the observed effects are contextual, or occur at the individual level. Goodman regressions and fractional logic models allow us to test models with several explanatory variables, including variables that are not based on aggregate statistics of individuals, in our case dummy variables for French-speaking municipalities and for the municipalities that lost their ballots.

*Hypotheses 1 and 2* are thus tested in the following model, where  $\mathbf{X}$  are the variables included in the control model, and  $\Delta_{\text{lostballots}}$  indicates the municipalities that lost their ballots.

<sup>18</sup> The size of the municipality also serves as a proxy for different types of communities. If we assume that the size of municipalities affects the electoral returns, we find it plausible that the effect on the vote share in a referendum rather follows the relative increase in size of a municipality, rather than an absolute increase. The effect is not altered if the number of registered voters is not transformed.

$$y = \beta_0 + \beta' \mathbf{X} + \beta_{\text{LB}} \cdot \Delta_{\text{lostballots}} + \varepsilon$$

As *Hypothesis 3* relates to the variance part of the estimates, and not to the mean, we need to test it using variance models. They are based on a maximum likelihood estimator that establishes the parameters of the outcome term and the variance simultaneously (Davidian and Carroll, 1987; Braumoeller, 2006).  $\mathbf{X}$  is a matrix of explanatory variables for the mean function;  $\mathbf{Z}$  is the matrix of control variables for the variance function. Both,  $\beta$  and  $\gamma$ , are vectors of parameters for both functions,  $\alpha_\mu$  is the constant in the mean term, and  $\alpha_\sigma$  the constant in the variance term.

Again, we include terms for the size of municipalities (number of registered voters), and a dummy variable for French-speaking municipalities in the variance part of the model, because we expect that predictions of voting results might be more accurate in larger municipalities.

$$y \sim N(\mu, \sigma^2)$$

$$\mu = \alpha_\mu + \beta' \mathbf{X} + \beta_{\text{LB}} \cdot \Delta_{\text{lostballots}}$$

$$\sigma^2 = \exp(\alpha_\sigma + \gamma' \mathbf{Z} + \gamma_{\text{LB}} \cdot \Delta_{\text{lostballots}})$$

We first run the three models for earlier cases of optional referendums in the canton of Berne. This allows us to select the models that have the best explanatory power, and to maximize the accuracy of the predictions of the municipal referendum results. Thereafter, we run the models in order to examine the results of the referendum on February 13th 2011.

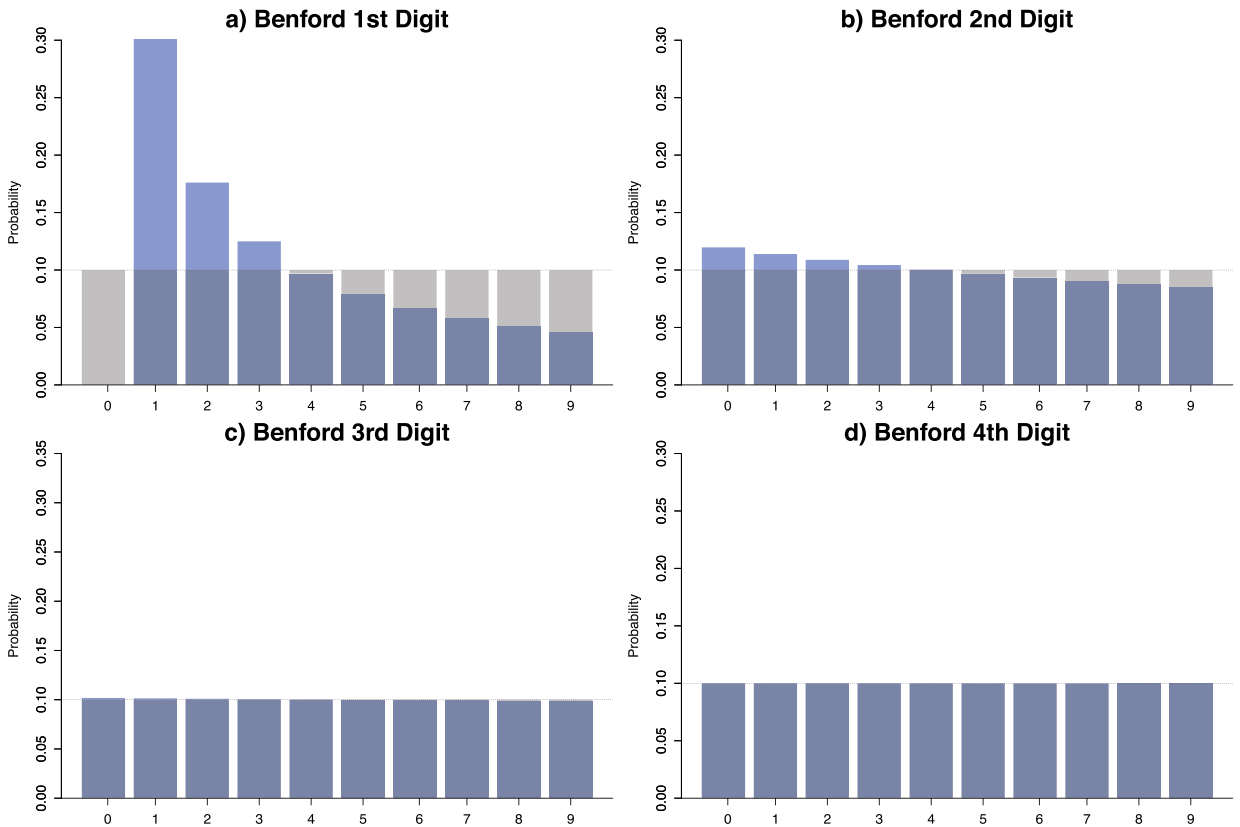
To rule out a possible endogeneity of the 30 selected municipalities, we have tested several hypotheses (partisan approach, size of the municipalities, language group, and interactions of these variables), in order to explain why certain municipalities might have lost their ballots. None of these hypotheses is able to contribute to the explanation of the losses of the ballots.

### 3.2.2. Digit based test for *Hypothesis 4* – can Benford help?

Recently, Benford's law has been applied by several social scientists to distinguish between genuine numbers and 'made-up' or 'manufactured data' (Diekmann, 2007; Mebane, 2010b). It has been shown over and over again, that when individuals make up numbers they tend to pick too often some digits and other digits are chosen too rarely. This psychological bias – the inability to truly pick random numbers – can be exploited for a forensic test. Benford reports in a paper from 1938 that the first couple of pages of a table of common logarithms are used far more often than others (Benford, 1938).<sup>19</sup> This sparked his interest in the frequency of specific digits. Benford derived a distribution that describes amazingly well the frequency of digits for many different processes (Diekmann, 2007; Raimi, 1969).

According to Benford's law the first digit of a number follows a simple distribution where the digit '1' is more

<sup>19</sup> The observation that the first couple of pages seem to be used more is ascribed to Newcomb (1881).



**Fig. 2.** Predicted Probabilities. Notes: The blue bars display the frequencies according to Benford's law. The gray bars indicate a uniform distribution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

probable than the digit '2', the digit '2' is then more frequent than the digit '3' and so on. The distribution is defined as  $P(z_i) = \log_{10}(1 + 1/z_i)$ , hence the probability to find the digit '2' should be  $p(z_i = 2) = \log_{10}(1 + 1/2) = 0.176$ . That means that if digits actually would follow a Benford distribution almost one out of five digits should be a two.

It should not be overlooked that Benford provides more than just a distribution for the first digit. Benford provides a probability mass function for any digit at any position ( $p$ : position,  $d$ : digit). Equation (1) describes the probabilities for the leading digit and Equation (2) describes the probabilities for any digit at any position if  $p > 1$  (not leading):

$$P(Z_1 = d) = \log_{10}\left(1 + \frac{1}{d}\right) \tag{1}$$

$$P(Z_p = d) = \sum_{i=10^{p-2}}^{10^p-1} \log_{10}\left(1 + \frac{1}{10i + d}\right) \tag{2}$$

In a review essay Hill (1995) describes many different processes which seem to follow a Benford distribution (e.g. physical constants, population of counties, income tax data). On top of that Hill also offers an explanation based on a variant of the central limit theorem assuming that all

numbers stem from a random selection of random variables.<sup>20</sup>

It is important to highlight that the frequency of any digit  $d$  depends on its position. The digit "1" has a probability of about 0.3 to appear as the first digit, while it has only a probability of about 0.1 to appear if we are looking at the fourth digit. The following plot (Fig. 2) illustrates the probabilities for first, second, third, and fourth digits.

Table 7 (Appendix) shows the distributions for the first four digits according to Benford's law. From Fig. 2 we see that the first digit follows a distinct *non*-uniform distribution but as we move back in position (increasing  $p$ ) we find that the distribution approximates a uniform. It can be shown that as  $p \rightarrow \infty$  the distribution is uniform (Hill, 1995).

The discussion of Benford's law so far may give the impression that we can use it to detect fraud on return sheets. If people in charge of reporting the results from the ward level manipulate the numbers, we might be able to detect that. Unfortunately, Benford's law does not say that every series of numbers follows automatically a Benford distribution. Phone numbers for example do not follow

<sup>20</sup> For an extensive review of the mathematical literature see Raimi (1976).



Benford's law.<sup>21</sup> The first digit of vote return data might not always stem from Benford's law (Mebane, 2008; Deckert et al., 2011). In a recent article Deckert et al. (2011) argue against the use of Benford's law based on using the mean of the second digits and extensive simulations (see also Shikano and Mack, 2011). Whereas we do not doubt their results, we want to highlight that we are not performing any tests on the means of digits nor on the second digit (Mebane, 2011). Instead, we follow Beber and Scacco (2012) in focusing on the last digit and an emphasis on a micro-logic of fraud.

### 3.2.3. Testing digits

Hence, we rely on Benford's law only for the last few digits and dismiss the first and second digit. We will use Benford's law while focusing on the last digits; the inconsequential ones. One may argue that focusing on third and fourth digits would be meaningless because elections are not won by altering those numbers. But if numbers are made up entirely, we expect that little care is given to the last digits and we should find significant deviations in the frequencies of digits.<sup>22</sup>

Regardless whether care is given or not to fraudulent return sheets, humans are almost incapable of generating good fake data. A large number of experimental research shows that individuals are incapable of producing pseudo-random numbers (see Beber and Scacco, 2012; for an extensive overview of the experimental literature). This inability is of great value to election analysts which would like to test vote reports for accuracy.

To test whether digits follow a specific distribution or not we need a test statistic. We will use a common  $\chi^2$  test (see e.g. Snedecor and Cochran, 1989). This test is only asymptotically valid as many other tests. This Pearson  $\chi^2$ -test computes the weighted squared deviations from the theoretical expectation for each digit and sums it up. Readers familiar with the  $\chi^2$ -test for  $n \times m$  tables will see the similarity between the two tests since the basic idea is the same. The basic idea is that if the data we are dissecting is similar to the theoretical prediction we expect the differences to be about 0 and the test statistic should be small. Let  $t_i$  be the expected frequency of observed digit  $i$  and let  $z_i$  be the frequency of observations of  $i$ :

$$\mathcal{B} = \sum_{i=0}^9 \frac{(z_i - t_i)^2}{t_i} \quad (3)$$

$$\mathcal{B} \sim \chi_9^2 \quad (4)$$

This test statistic  $\mathcal{B}$  follows under the null hypothesis – that the data measured in  $z_i$  stem from the theoretical

expectation – a  $\chi^2$ -distribution with 9 degrees of freedom. A potential weakness of testing Benford's law with  $\chi^2$  test is that the power of such a test hinges on the sample size.<sup>23</sup> This is not a concern if one inspects a large number of wards or municipalities, but becomes a problem when the sample size is small. In the application which follows we use this test for a subset of municipalities and the smallest sample has only 30 observations. The draw-back is that failure to reject the null hypothesis can be because the data follows the theoretical distribution well but can also be due to a lack of statistical power. This has to be taken into account when interpreting the test results.

If we were looking at the  $n^{\text{th}}$  digit and  $n$  would be large, the theoretical distribution is uniform, hence  $t_i = t = 1/10 \forall i$ . In our application we will encounter votes from small municipalities with a few hundred votes but also larger ones with more than a thousand votes. Hence, the last digit falls on the third, fourth, or rarely even the fifth position. Instead of brushing away the inconvenience we derive for every case the appropriate mixture distribution (usually based on 3rd and 4th digits). Details on deriving the mixture distribution are provided in the Appendix (A.3).

## 4. Results

This paper proposes two different approaches to deal with vote return data. Both approaches are based on the basic idea that we have two sets of municipalities; the ones that followed the law and kept the ballots and the other municipalities which did not do so. Both approaches (ecological and digit based) are used to investigate whether the two groups are distinct. If loosing or destroying ballots was a mistake we would expect that the subgroup of municipalities, who lost ballots, would perform the same way on both tests. But if these thirty municipalities have very atypical vote returns, this would raise suspicion whether actually fraudulent acts were committed and the ballots not lost but rather destroyed to render a recount unfeasible.

### 4.1. Berne 2011 – ecological tests

The first three hypotheses address the questions, whether the 30 municipalities have reported lower figures of empty ballots, referendum results which deviate more strongly from the expectations, compared to the other municipalities. Therefore, we first estimate control models to predict the number of empty ballots and to predict the vote returns. We test these models on four referendums with optional questions (see Appendix A.5 for a list), before

<sup>21</sup> Diekmann (2007) rejects the use of the first digit based Benford test for coefficients in published articles. His objective is to detect academic fraud and he argues to only use the second digit.

<sup>22</sup> Beber and Scacco (2012) argue forcefully for the use of last digits and rely on a uniform distribution. Since Benford essentially is uniform for later digits (third or more) Beber and Scacco are supporting the use of Benford's law on last digits.

<sup>23</sup> There are however small sample correction factors for some alternative tests, which should increase test power. One such alternative is the Kolmogorov–Smirnov (K–S) test (see Stephens, 1970; for an adjusted critical value). Even though this test is for continuous distributions there exists the claim that one can also adjust for discrete distributions (see a working paper by Morrow, 2010). The problem here is that all K–S related tests have low power if the distribution is not trend shaped but rather multimodal (Pettitt and Stephens, 1977). However, we show the performance of the two tests for a specific distribution and show why we rely for this application the Pearson  $\chi^2$  test (see Appendix A.2).

**Table 2**  
Explanation of the empty ballots in the tie-break question (H<sub>1</sub>).

	Public employees (control case)	Hospital (control case)	Taxes (control case)	Energy (control case)	Motor vehicle taxes (test case)
Empty – PB	0.090 (0.076)	0.168* (0.080)	0.065 (0.080)	0.017 (0.040)	0.165* (0.069)
Empty – P'sA	0.377** (0.075)	0.381** (0.099)	0.314** (0.067)	0.027 (0.104)	0.385** (0.100)
'Yes' – PB	–0.239** (0.036)	–0.280** (0.029)	–0.104* (0.042)	–0.048* (0.019)	–0.298** (0.041)
'Yes' – P'sA	–0.189** (0.045)	–0.227** (0.039)	–0.133** (0.051)	–0.117** (0.032)	–0.290** (0.042)
French (d)	0.008 (0.008)	0.006 (0.006)	0.019** (0.007)	0.005 (0.004)	0.014** (0.004)
Reg. voters (log)	0.000 (0.001)	0.002(*) (0.001)	–0.001 (0.002)	0.002* (0.001)	0.001 (0.001)
Lost ballots	0.010 (0.007)	–0.005 (0.006)	0.001 (0.006)	0.001 (0.004)	–0.008* (0.003)
Constant	0.251** (0.038)	0.291** (0.035)	0.168** (0.057)	0.115** (0.032)	0.297** (0.041)
N	372	372	372	372	372
R <sup>2</sup>	0.475	0.646	0.305	0.151	0.553

Note: OLS and robust standard errors. PB = Parliament bill, P'sA = People's amendment. (log) = logarithm, (d) = dummy.  
\*\* $p < 0.01$ , \* $p < 0.05$ , and (\*) $p < 0.1$ .

we will use them to estimate whether there is statistical evidence for any of the three hypotheses on the referendum results on February 13th 2011 (for hypotheses 2 and 3, the tests of the control models for the four other referendums are reported in [Tables 8, 9 and 10 in Appendix A.6](#)). We first discuss the results for [Hypothesis 1](#), according to which we expect fewer empty votes cast for the tie-break questions. The historical model (number of empty votes in previous optional referendums) performs badly. However, we can explain the number of empty votes for the tie-break question based on the interdependence of referendum questions. We argue that the voters' decisions on questions that were asked on the same ballot, and for the same matter, can be related to each other (see above, [subsection 3.2.1](#)). First, we have the control model for four reference cases, as reported in [Table 2](#).

In all four cases, the model contributes considerably to the explanation of the empty ballots in the tie-break questions. All included variables are statistically significant for at least some of the four optional referendums, and explain up to 64% of the variance. The model also performs well for the referendum of February 13th 2011, on motor vehicle taxes. After controlling for the correlations within the electoral ballots, the 30 municipalities that lost their ballots still show some deviating results. On average, we count 0.2–1.4%<sup>24</sup> fewer empty fields for the tie-break question, compared to similar ballots cast in other municipalities. Hence, we find that there was an effect diminishing the number of empty fields for the tie-break questions in those municipalities that have lost their ballots. The reasons for this difference cannot be answered in this paper. While one option is (as hypothesized) that crosses might have been added to the empty fields of the

tie-break questions, the effect might also have emerged from a different practice of distinguishing valid from invalid votes. Based on the 18,162 ballots that were cast in the 30 concerned municipalities, the overall effect might be anywhere in between 30 and 250 votes. We have also re-run the models relying on fractional logit model, and results substantially remain the same.

Second, we build models that explain the number of empty votes for all three referendum questions. These models allow us to evaluate whether possibly in the 30 municipalities that lost their ballots, empty fields on the ballots were filled in general. This time, we need a different control model: we are investigating a possible manipulation that might have affected the number of empty ballots for any of the three referendum questions, or all three simultaneously. As a consequence, we cannot rely anymore on the interdependency of the referendum questions within the same ballot for the control model. Instead, we rely on the historical model, and investigate whether the level of empty ballots in municipalities can be explained with the records from the last earlier optional referendums. As for [Hypothesis 1](#), we first tested the explanatory power of the model, relying solely on earlier optional referendum. In some cases, the number of empty ballots correlates between the two referendums, and the control model covers some 10–20% of the observed variance (see [Table 8 in A.6](#)).

We use the same model to predict the number of empty ballots in the referendum of February 13th 2011, with the tax referendum of February 24th 2008 as our lagged case. As there are three referendum questions, this results in three separate models for the empty votes for each of the three questions (see [Table 3](#)). The explanatory power of the model is weak, for all three dependent variables, and there is no statistically significant effect of the lost ballot municipalities. The substantial magnitude of this effect remains very similar to the previous analysis (see [Table 2](#)), but given the high

<sup>24</sup> Given an effect of 0.8 percent points, a RMSE of 0.3%, and a 95% coverage.

**Table 3**  
Explanation of the empty ballots in all question (H<sub>2</sub>).

TAXES	Parliament bill	People's amendment	Tie-break question
Empty – PB	0.108(*) (0.060)	0.004 (0.052)	0.067 (0.064)
Empty – P'sA	–0.043 (0.077)	0.093 (0.059)	0.031 (0.096)
Empty – TBq	0.174** (0.064)	0.036 (0.052)	0.133* (0.057)
French (d)	0.001 (0.006)	–0.007 (0.004)	0.019** (0.005)
Reg. voters (log)	–0.003(*) (0.001)	0.000 (0.001)	–0.001 (0.001)
Lost ballots	–0.001 (0.005)	0.000 (0.004)	–0.007 (0.005)
Constant	0.080** (0.012)	0.047** (0.012)	0.049** (0.012)
N	372	372	372
R <sup>2</sup>	0.091	0.040	0.146

Note: OLS and robust standard errors. PB = Parliament bill, P'sA = People's amendment, TBq = Tie-break question. (log) = logarithm, (d) = dummy. \*\**p* < 0.01, \**p* < 0.05, and (\*)*p* < 0.1.

standard error of the model, it is not statistically significant. We have performed several robustness checks, which did not alter the results.<sup>25</sup> Results do not change if we take other optional referendums as reference (see Table 9 in A.6).

Finally, we wonder whether we can find any effect on the valid (i.e. non-empty) votes. We did not hypothesize how the vote might have been manipulated, but if some of the lost ballot municipalities have manipulated the referendum, their results should be less well predictable than those of other municipalities (Hypothesis 3).

Again, not knowing which of the three referendum questions might have been manipulated, the control model that is based on the interdependence of the three referendum questions is not applicable. Therefore, we have tested two control models, relying on the national parliamentary elections of 2007,<sup>26</sup> and on voting patterns in previous optional referendums.

We include terms for the size of municipalities (number of registered voters), and a dummy variable for French-speaking municipalities in the variance part of the model, because we expect that predictions of voting results might be more accurate in larger municipalities. The explanatory power<sup>27</sup> of the party composition of the municipalities varies for different referendum topics, between 30% (referendum on hospitals of 2005) and 70% (energy law of

2011) (see Table 10 in A.6). The results remain mostly the same, if we logit-transform the parties' vote shares and the vote shares in the referendum.<sup>28</sup>

We have further searched for municipalities for which this control model has already lead to non-accurately explained predictions in earlier referendums. We have run the control model on earlier referendums, including the residual from these models as a lagged variable in the variance part of the model for later referendums. Neither are effects statistically significant, nor does this historical model contribute to the explanatory power. Therefore, we have refrained from applying it to the referendum of February 13th 2011.

The main model (Table 4), including a dummy variable to identify the lost ballot municipalities in the variance part, does not give any statistical evidence that the results might have been manipulated. The dummy variable is not significant in any of the three models. This means that the results from the lost ballot municipalities do not deviate more from the model's prediction than the results from other municipalities. This, while the control model is very powerful as a predictor of the referendum results in February 2011, with 74–76% (variance-weighted) explanatory power. We find no evidence for hypotheses 2 and 3, but we do find patterns consistent with the behavior described in Hypothesis 1.

#### 4.2. Berne 2011 – digit-based tests

In this section we test Hypothesis 4. It states that the last digits of the reported numbers is in line with the theoretical expectation and follows a Benford distribution as laid out in subsection 3.2.2. The fourth hypothesis is based on fraudulent behavior of the official reporting the results to the next administrative level. If fraud occurred in such a manner in those municipalities which illegally disposed the ballots too early, we expect that the last digits of the reported votes does not follow a Benford distribution. If fraud actually occurred by misreporting the final numbers we should find no significant test statistic for the 342 municipalities which kept the ballots but would expect to find significant test statistics for the 30 municipalities which disposed the ballots. Due to the small sample size of the second group it is possible that even if fraud occurred that one would not detect it in this case due to the low power of the test (see subsection 3.2.3).

Our results do not coincide with this expectation. In Table 5 we perform six  $\chi^2$  tests on the reported votes from the 342 municipalities which kept the ballots. To provide a

<sup>25</sup> First, we have taken other optional references as lagged reference cases (see Table 9 in A.6). Second, we also rerun the models in Tables 4 and 5 while relying on a fractional logit model. The results are substantively identical although the exact numbers slightly change, but direction and significance level for all coefficients remain unchanged.

<sup>26</sup> In national parliamentary elections, the whole canton of Berne is a single electoral district. Differently, in the cantonal elections of 2006/2010, there were 8/9 electoral districts, so that the offer of electoral lists varies widely, and affects the electoral returns for the major parties. We rely on the national parliamentary election results of 2007 for all referendums from 2004 to 2011, as we are mainly interested in a good fit of the control variables, whereas causality is a minor concern for the control models.

<sup>27</sup> As measured by a quasi-R<sup>2</sup> measure (VWLS R<sup>2</sup>).

<sup>28</sup> The logit-transformation was performed for all percentage variables, i.e. vote shares in referendum and elections and turnout, following Greene (1993, p. 837). While Greene suggests to correct unanimous votes, by increasing/decreasing them at a small constant rate, so that these cases are not dropped in the logit transformation, we have added one yes and one no vote to all voting results, so that none of the results is unanimous anymore. Both the transformed and the non-transformed model have similarly good fits, and we detect no sign for clear non-linearities in the function in the graphical display of residuals. We have not performed similar logit-transformations for the tests of hypotheses 3 and 4, as we believe that the effects underlying the models are clearly linear in their nature.

**Table 4**Explanation of the referendum results (share of yes vote), party-political model (H<sub>3</sub>).

Referendum	Motor vehicle taxes		
	Parliament bill	People's amendment	Tie-break question
Main part			
Votes SVP%	-0.085 (0.059)	0.062 (0.062)	-0.085 (0.065)
Votes SP%	0.145* (0.067)	-0.075 (0.07)	0.143(*) (0.075)
Votes FDP%	0.305** (0.071)	-0.203** (0.067)	0.263** (0.074)
Votes GPS%	0.887** (0.079)	-0.942** (0.077)	0.975** (0.088)
Turnout elections	0.393** (0.055)	-0.355** (0.061)	0.457** (0.059)
Turnout refer.	-0.156* (0.064)	0.200** (0.071)	-0.230** (0.067)
Lost ballots (d)	-0.008 (0.011)	0.004 (0.010)	-0.006 (0.010)
Reg. voters (log)	0.010** (0.003)	-0.011** (0.003)	0.010** (0.003)
French (d)	-0.086** (0.011)	0.040** (0.011)	-0.080** (0.011)
Constant	0.142* (0.062)	0.778** (0.060)	0.129* (0.064)
Variance part			
Lost ballots (d)	0.559(*) (0.297)	0.143 (0.291)	0.271 (0.331)
Reg. voters (log)	-0.531** (0.076)	-0.660** (0.089)	-0.499** (0.120)
French (d)	-0.229 (0.24)	-0.248 (0.254)	-0.228 (0.264)
Constant	-2.44** (0.532)	-1.391* (0.590)	-2.566** (0.780)
N	372	372	372
Model $\chi^2_{10}$	414.092	330.111	399.456
Prob > $\chi^2$	0.000	0.000	0.000
Pseudo R <sup>2</sup>	-0.5454	-0.4184	-0.5322
VWLS R <sup>2</sup>	0.7641	0.7371	0.7507

Note: Variance model with robust standard errors. (log) = logarithm, (d) = dummy. \*\* $p < 0.01$ , \* $p < 0.05$ , and (\*) $p < 0.1$ .

reading example we look at the test for the reported 'yes' votes on the parliament bill: The test statistic is 6.12 which corresponds to a  $p$ -value of 0.73 (df = 9). Given a confidence level of 0.05 we fail to reject the null hypothesis (digits stem from a Benford distribution).

For five out of the six tests we fail to reject the hypothesis (H<sub>4</sub>), i.e. the observed distribution is not different from Benford's Law. But, we fail to reject H<sub>4</sub> for the vote on the People's Amendment in the tie-break question. This is not in line with our expectations, and it raises the question

**Table 5**Digit based test – 342 municipalities (H<sub>4</sub>).

	Parliament bill		People's amendment		Tie-break question	
	Yes	No	Yes	No	PB	P'sA
Test value ( $\chi^2_9$ )	6.12	12.84	4.74	9.22	8.18	24.88
$p$ -value	0.728	0.170	0.856	0.417	0.516	0.003*
Benford?	✓	✓	✓	✓	✓	✗
df	9	9	9	9	9	9
N	342	342	342	342	342	342

**Table 6**Digit based test – 30 municipalities (H<sub>4</sub>).

	Parliament bill		People's amendment		Tie-break question	
	Yes	No	Yes	No	PB	P'sA
Test value ( $\chi^2_9$ )	6.48	4.34	5.02	8.02	15.39	8.59
$p$ -value	0.691	0.888	0.833	0.533	0.081	0.476
Benford?	✓	✓	✓	✓	✓	✓
df	9	9	9	9	9	9
N	30	30	30	30	30	30

whether this test is valid. In principle, there are two possibilities why one would find a significant test statistic even if there was no fraud; first, the large test statistic is a type-I error, and second, the test is not valid. Unusual cases, where the test leads to statistically significant results, although no fraud is expected or suspected, are reported in the literature.<sup>29</sup> The second possibility is that the test is not appropriate for digits stemming from a tie-break question. It is known that certain circumstances can lead fraud-free election results to not conform to the theoretical distribution.<sup>30</sup> But as laid out above, we explicitly focus on the last digit as it should be unaffected by the known issues. Finally, if there was a specific mechanism at work (yet unknown) which would lead the last digits of a tie-break question to not conform with the theoretical expectation we would expect to also find a large test statistic for other tie-break questions (such as the ones in Table 2) but this is not the case. Altogether, there is no reason to expect the test being inappropriate for tie-break questions.

In Table 6 we show the same eight hypothesis tests for the subsample of municipalities which did not keep the ballots and prevented thereby a recount. If we think that these municipalities forged the vote results by changing the numbers, we would expect to find that the  $\chi^2$  test statistic significantly deviates from 0. In all six cases we find a small  $\chi^2$  value which is lower than the critical threshold (for an  $\alpha$  level of 0.05 and 9 degrees of freedom it is 16.92). As mentioned in subsection 3.2.3, the power of this test is small due to the very small sample of only 30 observations.

Based on the results of the digit based tests, we do not find any meaningful deviations. Keeping in mind the low power of this test, the absence of evidence is not evidence to the contrary. It simply means that one possibility how one could have detected fraudulent elections did not provide evidence for fraud.

#### 4.3. Combining multiple hypothesis tests

The last two paragraphs present the results of the carried out empirical tests. Each of these tests provide a test

<sup>29</sup> An example of this is found in Mebane (2008, 171–172) where he finds significant deviations for Los Angeles county and three other counties in the US.

<sup>30</sup> Note that all known anomalies causing digits to not follow the theoretical distribution are concerned with the first or second digit. This may be due to strategic voting (Mebane, 2010a) or that some vote machines are only used during peak hours (roughly equal division with leftovers, REDLW, see Mebane, 2006).



statistic for a specific hypothesis. The specific hypotheses were e.g. whether empty ballots were manipulated ([Hypothesis 1](#)) or whether officials misreported vote totals ([Hypothesis 4](#)). Altogether we find evidence for one of the four postulated hypotheses, namely that empty ballots were filled in after they had been casted.

The general interest in a paper of electoral forensics does not uniquely lie in distinguishing which form of fraud occurred but also whether fraud occurred at all or not. How should one treat the multiple tests? There are two extreme alternatives; first, one could treat each hypothesis as fully independent and derive a *combined* test statistic under a null hypothesis. But fraud may only occur in one specific way – hence, a highly fraudulent election in which only precincts misreport vote totals (and no other form of fraud) can go by undetected. The second alternative is to forego the issue of multiple testing and to claim that whenever one finds a positive test statistic it has to be proof of fraud. The problem with this approach increases as well in the number of carried out hypothesis test. Since we employ the null-hypothesis-significance-testing we are bound to have positive results if the number increases sufficiently.

The first logic can be implemented assuming full independence or can be amended by relying on Fisher's method.<sup>31</sup> Using Fisher's method in this application leads to a test statistic with a *p*-value of 0.22 which does not allow to reject the null hypothesis of no fraud at any conventional level (one test from [Table 2](#), three from [Table 3](#), three from [Table 4](#), and six from [Table 6](#)). This approach, however, entails the risk of overseeing single forms of fraud. With a growing number of hypotheses, for which we test, we increase the level of the *p*-value, which would be required to indicate fraud in any of the individual micro-scenarios. While we capture a wider array of forms of fraud, it becomes thus more difficult to detect any single form of it.

Instead, we could follow the second approach, treat every hypothesis separately, and if any of them is not rejected, take this as a sign that there was fraud. Again, this may lead to exaggerated claims of fraud when the number of tests is high. In this application one would, based on the results in [Table 2](#), claim that fraud has occurred. With more tests of fraud conducted, the researcher will be more likely

to find evidence of fraud. In the beginning we have emphasized that it is of central importance that one first derives micro-logics which explain how potential fraud could occur. This step and the necessity to justify the micro-logics constrain the inclusion or exclusion of test results. This second approach also retains the possibility to pin point where or how the fraud occurred.<sup>32</sup> After implying that fraud occurred, the immediate following question will be “*How was fraud committed?*” or “*Who committed fraud?*” and this approach allows to have a clear answer to that question.

We support this second approach, for two reasons. First, fraud may only occur according to one micro-logic. It is irrelevant how many other forms of fraud are tested, one should be able to conclude that such an election was manipulated. As a safeguard against an uncontrolled inflation of tests – which under the second approach eventually would produce one or the other wrong indication of fraud – we restrain ourselves to derive precise micro-logics and justify our choices.

In the application, analyzing the tax referendum of 2011, we have tested four distinct fraudulent actions. The test results for three of the four tests are in line with a fair election. Where the data and models raise a red flag is for the first hypothesis which states that fields left empty in the tie-break question were manipulated. The result of that test supports the argument of fraud – there are significantly less empty ballots in those thirty municipalities than in the 342 other municipalities reported. Finding fraud for the most low-scale forms of manipulation (here: [Hypothesis 1](#)), but not for more demanding forms or more easily detectable forms of fraud, is therefore in line with the expectations. With this reasoning, it should not be worrisome, if one does not detect other, more costly forms of fraud. In hindsight it would have been more troubling to find empirical support e.g. for the fourth hypothesis.

#### 4.4. Where to go from here

The tests presented so far only use clearly quantifiable data and treat all potential micro-logics as equally relevant and probable. Before any tests are carried out one has already an expectation of how likely each form of fraud is. In this application it can be argued that filing in empty fields in the tie-break question ([Hypothesis 1](#)) is more likely than outright mis-reporting of vote totals ([Hypothesis 4](#)). The psychological models of fraudulent behavior as well subjective expectations do not enter these tests or their final evaluation. In the future, exploiting this by adopting an explicitly Bayesian approach where one combines this information in form of a prior promises an advance. Using Bayes' rule faces a distinct challenges here as one will have to have an unconditional probability function for fraud. Identifying ways to derive such a function would enable the implementation of a Bayesian contribution to electoral forensics.

<sup>31</sup> Following the logic one has a *overall* statistic which follows a binomial distribution with  $p = 0.95$  and  $n = 4$ . The probability of finding non-significant test statistics in three instances and a significant test in one case (three hypotheses are not supported by the data in this application) is  $p = \binom{4}{1} \cdot 0.95^3 \cdot (1 - 0.95) = 0.17$  which does not warrant to reject the null hypothesis. Such a method is sensitive to the exact significance level one chooses as a rejected null hypothesis is counted as 1 and barely not rejected is counted as 0. That means that relying on this approach we treat a *p*-value of 0.049 very different than a value of 0.051 although the difference between these two values might not be significant ([Gelman and Stern, 2006](#)). There is a more precise way to combine several *p*-values which dates back to [Fisher \(1948\)](#) and is known as *Fisher's method*. It requires to assume that the *p*-values are uniformly distributed and independent. If so, one can compute  $-2 \sum_{i=1}^k \log(p_i)$  (where  $k$  indicates the number of *p*-values) and this sum follows under the null hypothesis a  $\chi^2$  distribution with  $2k$  degrees of freedom ([Westfall, 2005](#); [Westfall and Wolfinger, 1997](#)).

<sup>32</sup> Note, that using Fisher's method one can have a series of *p*-values of which none is below 0.05 but the overall *p*-value is lower than 0.05 (example: five *p*-values of 0.15 lead to an overall *p*-value of 0.04).

## 5. Conclusion

This paper is interested in electoral fraud and electoral forensics. We set out from the observation that most contributions in the field of electoral forensics rely on a specific tool. Our main claim is that there is no one optimal test and that every tests needs to rest on a sound basis. First, we start with the fraudulent act and show for a specific case how one can define the different possible actions that can distort the true results. Second, we identify the wards, where fraud is most likely to happen, and consider this information in the statistical tests. This is guided by what we have labeled micro-logics and provides some prior expectation to its likelihood. The two dimensions of these micro-logics are the probability of detection (and possible prosecution) and the degree to which it is contradicting the voters' true intent. This framework also allows us to derive prior expectations for the illustrative case of each acts likelihood. In a second step we parse out the empirical implications for each of these fraudulent acts which may or may not have occurred. Despite the fact that we are only looking at one vote, we find at least four distinct tests which in turn rely on four different potential acts.

Our illustrative case is the optional referendum on February 13th 2011 in the Swiss canton of Berne. While there are no accusations of fraud, a re-count of the ballots, mandated by courts, was impossible, because 30 out of 383 municipalities declared that they have lost the ballots of the referendum in the meanwhile. One of these municipalities, after declaring that they have shredded the ballots, found them again. A re-count could never take place. We do not know what has happened in the 30 municipalities that declared their ballots as lost. The least harmful reason is sloppiness of the administration, the most harmful might be electoral fraud. If there were irregularities, they regard the 30 municipalities that have lost their ballots. We compare the election returns of the 342 other wards (usually: municipalities) in the same canton. We do find empirical support for irregularities which could be due to the most likely form of fraud. The Bernese ballot counts three referendum questions, the two main questions, which are answered by 'Yes' or 'No', and the (not less important) tie-break question, answered with a simple cross. We find that in the 30 municipalities which lost their ballots, fewer empty ballots were counted for the tie-break question. This empirical pattern is consistent with the hypothesis that some members of the electoral committee in the 30 municipalities might have filled in empty tie-break answer boxes.

The counting of the votes is little professionalized, especially in the smaller municipalities of Berne, and while the rules are clear, the cantonal administration has little knowledge about the counting process in general. This lack of oversight does not only apply to Berne but to Switzerland more generally. It was one of the few objections of the OSCE Election Assessment Mission which observed the national elections in October 2011 (OSCE, 2012). We remain agnostic whether those 30 municipalities actually tampered ballots by filling left empty ballot fields and believe that there is need for reform. We would like to remind the critical

reader, convinced of the properness of the process, of the case in Wimmis in 1956.

This illustration shows how one can fruitfully combine different approaches and tools rooted in theoretical micro-logics to the detection of electoral fraud. This framework allows to go beyond mere digit checking à la Benford's law or comparisons of current electoral results with historical results. The ultimate question, whether fraud actually occurred or not, remains a question which cannot be answered with certainty – but the confidence in our probabilistic assessments is greater once we put them on an appropriate theoretical foundation. Finding irregularities then points to specific acts and procedure which can subsequently be further analyzed.

## Appendix A. Supplementary data

Supplementary data related to this article can be found at <http://dx.doi.org/10.1016/j.electstud.2014.03.005>.

## References

- Alvarez, R.M., Boehmke, F., 2008. Correlates of fraud: studying state election fraud allegations. In: Alvarez, Michael, Hall, Thad E., Hyde, Susan D. (Eds.), *Election Fraud: Detecting and Detering Electoral Manipulation R*.
- Anderson, M.L., 2000. *Practicing Democracy: Elections and Political Culture in Imperial Germany*. Princeton University Press, Princeton.
- Bannenberg, B., Jehle, J.-M., 2010. *Wirtschaftskriminalität*. In: BoD–Books on Demand, vol. 112.
- Beber, B., Scacco, A., 2012. What the numbers say: a digit-based test for election fraud. *Polit. Anal.* 20 (3).
- Benford, F., 1938. The law of anomalous numbers. *Proc. Am. Phil. Soc.* 78, 551–572.
- Bochsler, D., 2010. The Marquis de Condorcet goes to Bern. *Public Choice* 144 (1–2), 119–131.
- Braumoeller, B.F., 2006. Explaining variance; or, stuck in a moment we can't get out of. *Polit. Anal.* 14, 268–290.
- Breunig, C., Goerres, A., 2011. "Searching for electoral irregularities in an established democracy. Applying Benford's law tests to Bundestag elections in Unified Germany. *Elect. Stud.* 30 (3), 534–545.
- Cox, G.W., Kousser, J.M., 1981. Turnout and rural corruption: New York as a test case. *Am. J. Polit. Sci.*, 646–663.
- Cressey, D.R., 1980. *Management Fraud: Detection and Deterrence*. In: *Management Fraud, Accounting Controls and Criminology Theory*, pp. 117–148.
- Davidian, M., Carroll, R.J., 1987. Variance function estimation. *J. Am. Stat. Assoc.* 82 (400), 1079–1091.
- Deckert, J., Myagkov, M., Ordeshook, P.C., 2011. Benford's law and the detection of election fraud. *Polit. Anal.* 19, 245–268.
- Diekmann, A., 2007. Not the first digit! Using Benford's law to detect fraudulent scientific data. *J. Appl. Stat.* 34 (3), 321–329.
- Filippov, M., Ordeshook, P.C., 1996. Fraud or fiction: who stole what in Russia's December 1993 elections. *Demokratizatsiya* 5 (1), 36–52.
- Fisher, R.A., 1948. Answer to question 14 on combining independent tests of significance. *Am. Stat.* 2 (30).
- Gelman, A., Stern, H., 2006. The difference between "significant" and "not significant" is not itself statistically significant. *Am. Stat.* 60 (4).
- Goodman, L.A., 1959. Some alternatives to ecological correlation. *Am. J. Sociol.* 64 (6), 610–625.
- Greene, W.H., 1993. *Econometric Analysis*. Macmillan, New York.
- Hill, T.P., 1995. A statistical derivation of the significant-digit law. *Stat. Sci.* 10, 354–363.
- Hyde, S., Marinov, N., 2008. *Does Information Facilitate Self-enforcing Democracy? The Role of International Election Monitoring*. Unpublished Manuscript. Yale University.
- Kitschelt, H., Wilkinson, S., 2007. Citizen-politician linkages: an introduction. In: Kitschelt, H., Wilkinson, S. (Eds.), *Patrons, clients, and policies: patterns of democratic accountability and Political Competition*. Cambridge University Press, Cambridge.
- Mebane, W.R., 2006. *Election forensics: Vote counts and Benford's law*. Summer Meeting of the Political Methodology Society, UC-Davis, July.

- Mebane, W., 2008. Election forensics: the second-digit Benford's law test and recent American presidential elections. In: Alvarez, Michael, Hall, Thad E., Hyde, Susan D. (Eds.), *Election Fraud: Detecting and Deterring Electoral Manipulation R*.
- Mebane, W., 2010a. Election fraud or strategic voting? Can second-digit tests tell the difference?. In: Prepared for Presentation at the 2010 Summer Meeting of the Political Methodology Society. University of Iowa.
- Mebane, W., 2010b. Fraud in the 2009 presidential election in Iran? *Chance* 23, 6–15.
- Mebane, W., 2011. Comment on "Benford's law and the detection of election fraud". *Polit. Anal.* 19, 269–272.
- Morrow, J., 2010. Benford's Law, Families of Distributions, and a Test Basis. Unpublished Manuscript.
- Mozaffar, S., Schedler, A., 2002. The comparative study of electoral governance—introduction. *Int. Polit. Sci. Rev.* 23 (1), 5–27.
- Myagkov, M., Ordeshook, P.C., Shaikin, D., 2008. On the trail of fraud: estimating the flow of votes between Russia's elections. In: Alvarez, M., Hall, T.E., Hyde, S.D. (Eds.), *Election Fraud: Detecting and Deterring Electoral Manipulation R*. Brookings Institution Press, Washington DC.
- Myagkov, M., Ordeshook, P.C., Shakin, D., 2005. Fraud or fairytales: Russia and Ukraine's electoral experience. *Post-Soviet Aff.* 21 (2), 91–131.
- Newcomb, S., 1881. Note on frequency of use of different digits in natural numbers. *Am. J. Math.* 4 (1/4), 39–40.
- Nuspliger, K., 2011. Bericht zur Vernichtung von Stimmzetteln. Staatskanzlei Kanton Bern, Bern.
- OSCE, Organization for Security and Co-operation in Europe, 2012. *Swiss Confederation Federal Assembly Elections 23 October 2011*. Office for Democratic Institutions and Human Rights, Warsaw.
- Papke, L.E., Wooldridge, J.M., 1996. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *J. Appl. Econ.* 11 (6), 619–632.
- Pettitt, A.N., Stephens, M.A., 1977. The Kolmogorov-Smirnov goodness-of-fit statistic with discrete and grouped data. *Technometrics* 19 (2), 205–210.
- Raimi, R.A., 1969. The peculiar distribution of first digits. *Sci. Am.* 221, 118–120.
- Raimi, R.A., 1976. The first digit problem. *Am. Math. Mon.* 83 (7), 521–538.
- Sansoni, M., 2011a. A: Stimmzettel wurden geschreddert. *Grenchner Tagblatt*. August 19, p. 29.
- Sansoni, M., 2011b. B: Plötzlich tauchen die Stimmzettel wieder auf. *Grenchner Tagblatt*. September 9, p. 28.
- Schedler, A., 2002. The menu of manipulation. *J. Democr.* 13 (2), 36–45.
- Shikano, S., Mack, V., 2011. When does 2nd Digit Benford's law-test signal an election fraud? Facts or misleading test results. *Jahrbücher für Natl. Stat.* 231 (5 + 6), 719–732.
- Snedecor, G.W., Cochran, W.G., 1989. *Statistical Methods*. Iowa State University Press.
- Stephens, M.A., 1970. Use of the Kolmogorov-Smirnov, Cramér-Von Mises and related statistics without extensive tables. *J. R. Stat. Soc. Ser. B Methodol.*, 115–122.
- Wasserfallen, F., 2011. Vereinfachung von Variantenabstimmungen mit Stichfrage. *Grosser Rat des Kantons Bern*.
- Westfall, P.H., 2005. Combining P Values. *Encyclopedia of Biostatistics*.
- Westfall, P.H., Wolfinger, R.D., 1997. Multiple tests with discrete distributions. *Am. Statistic.* 51 (1), 3–8.
- Wyler, S., 2011. Die Wahlfälschung von Wimmis. *Der Bund*. December 31.
- Ziblatt, D., 2009. Shaping democratic practice and the causes of electoral fraud: the case of nineteenth-century Germany. *Am. Polit. Sci. Rev.* 103 (1), 1–21.