Appendix to ‘Election Fraud: A Latent Class Framework for Digit-Based Tests’

Juraj Medzihorsky
Central European University
juraj.medzihorsky@gmail.com

Thursday 30th July, 2015 at 15:44

Replication materials are available online as Medzihorsky, Juraj, 2015, “Replication Data for: Election Fraud: A Latent Class Framework for Digit-Based Tests”, http://dx.doi.org/10.7910/DVN/1FYXUJ, Harvard Dataverse, V1 [UNF:6:FIWHvsHNzZgPStT0+kgbsQ==], and include the version of the R package pistar (Medzihorsky, 2015a) used in the analysis. The article uses data from Beber and Scacco (2012) which is available online also as (Beber and Scacco, 2013).

A.1 The Benford Distribution

Table A.1: Distributions of numerals under Benford's law for the first nine positions. Source: author's calculation.

	Position								
Numeral	1st	2nd	3rd	4th	5th	6th	7th	8th	9th
0		.119679269	.101784365	.100176147	.100017592	.100001759	.100000176	.100000018	.100000002
1	.301029996	.113890103	.101375977	.100136888	.100013681	.100001368	.100000137	.100000014	.100000001
2	.176091259	.108821499	.100972198	.100097673	.100009771	.100000977	.100000098	.10000001	.100000001
3	.124938737	.10432956	.100572932	.1000585	.100005862	.100000586	.100000059	.100000006	.10000000
4	.096910013	.100308202	.100178088	.100019371	.100001953	.100000195	.10000002	.100000002	.1
5	.079181246	.096677236	.099787576	.099980285	.099998044	.099999805	.09999998	.099999998	.1
6	.06694679	.093374736	.09940131	.099941242	.099994136	.099999414	.099999941	.099999994	.099999999
7	.057991947	.090351989	.099019207	.099902241	.099990228	.099999023	.099999902	.09999999	.099999999
8	.051152522	.087570054	.098641184	.099863284	.099986321	.099998632	.099999863	.099999986	.099999999
9	.045757491	.084997352	.098267164	.099824369	.099982414	.099998241	.099999824	.099999982	.099999998

A.2 Simulations

Figures A.1 and A.2 report the results for π^* and Δ, respectively. For the same pairs of means and standard deviations, the values of each index are strongly correlated across the simulated processes, with Pearson's $\rho > 0.88$ (N=400) for all pairs excluding Mebane's (2006) two mechanisms and $\rho > 0.45$ (N=169) for all possible pairs (shown in Figure A.3).

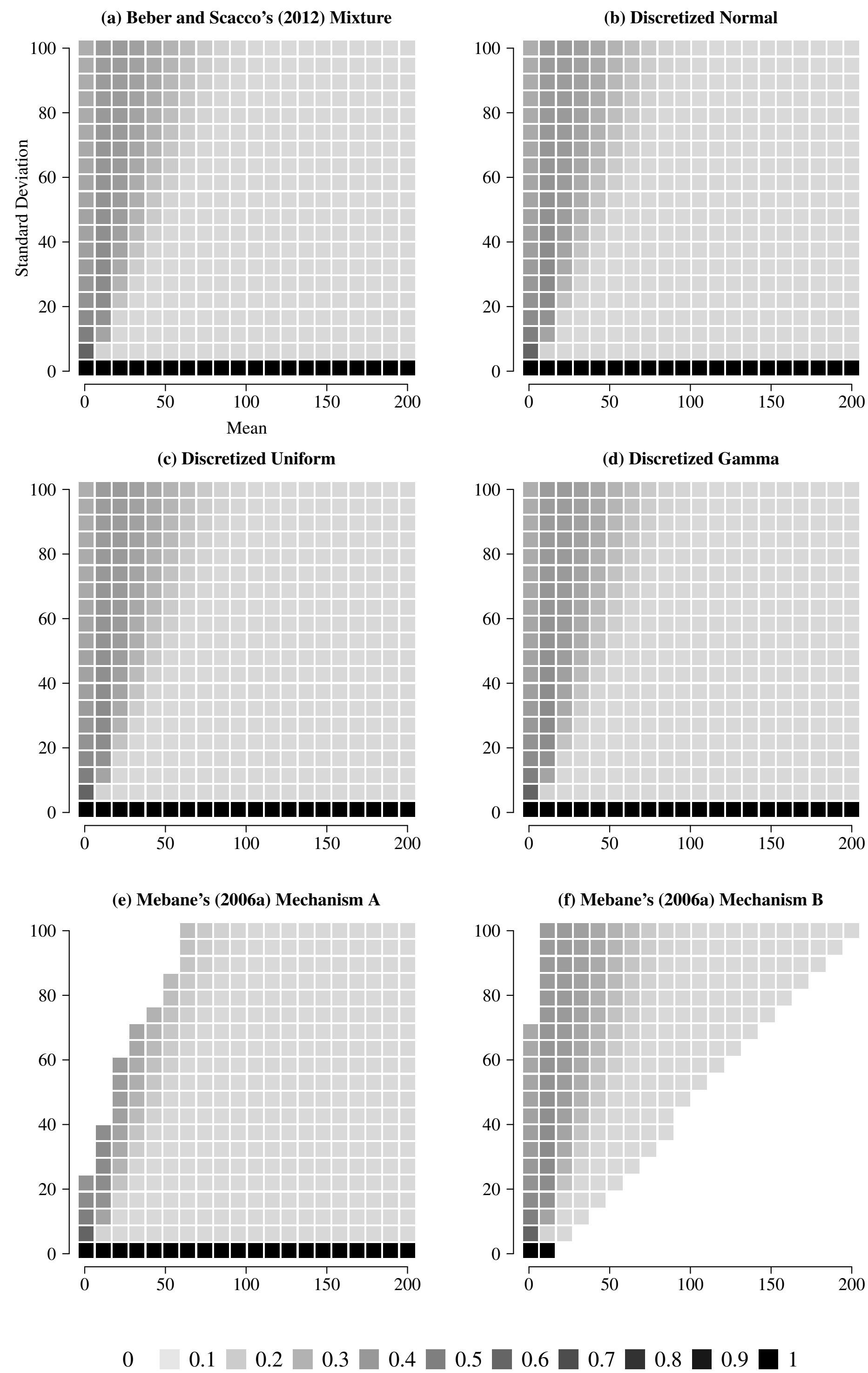

Figure A.1: Mean values of the π^* mixture index of fit for the model of uniformity for Beber and Scacco's (2012) simulated scenarios. Two of the six mechanisms are Mebane's (2006). For each combination of mechanism, mean, and standard deviation 1,000 samples of 1,000 numbers were simulated.

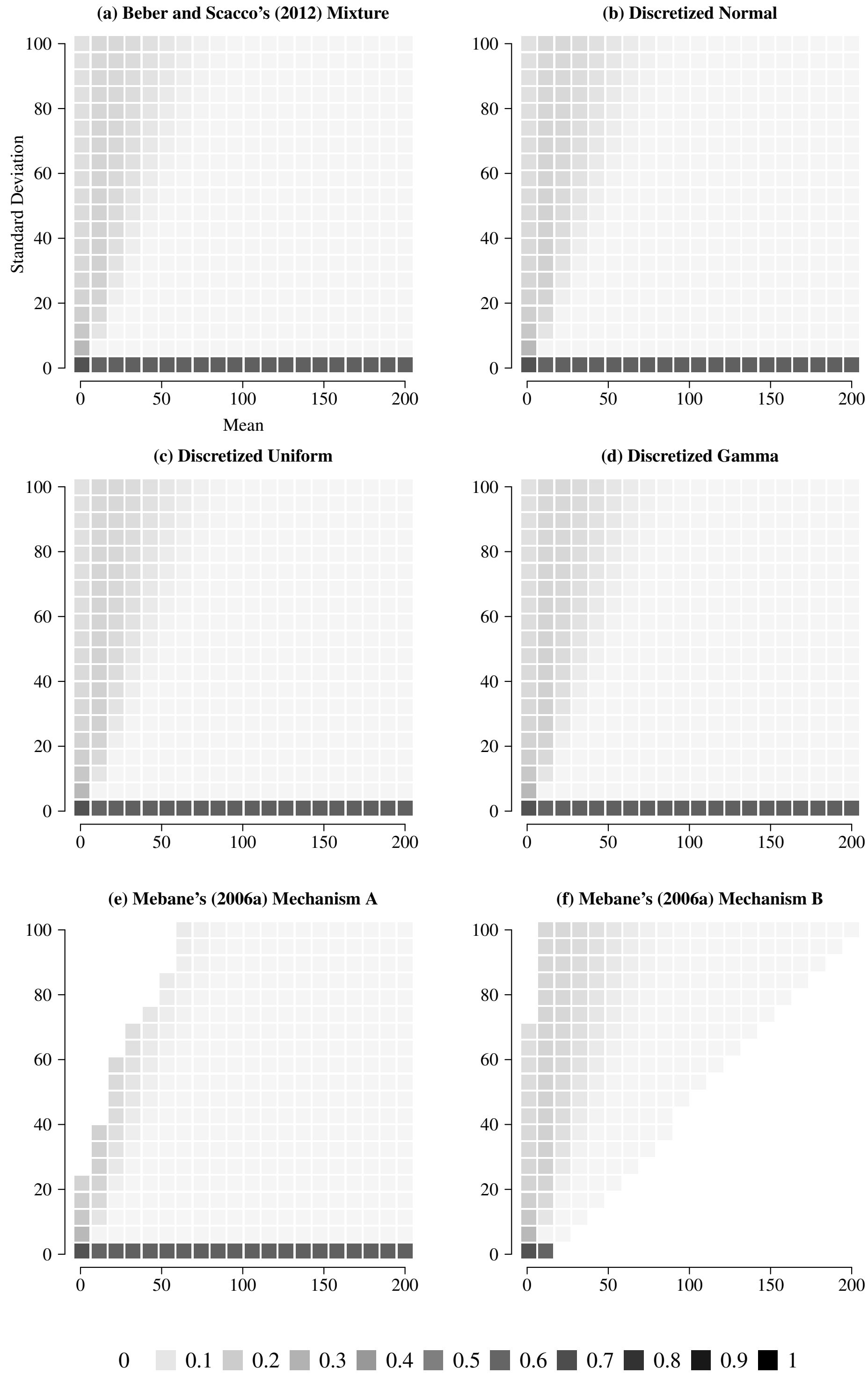

Figure A.2: Mean values of the Δ dissimilarity index for the model of uniformity for Beber and Scacco's (2012) simulated scenarios. Two of the six mechanisms are Mebane's (2006). For each combination of mechanism, mean, and standard deviation 1,000 samples of 1,000 numbers were simulated.

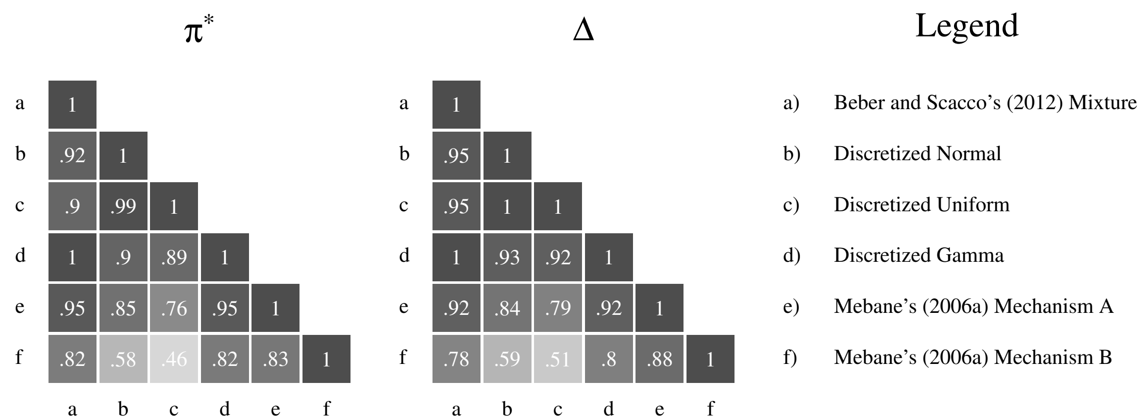

Figure A.3: Pearson correlations for the two indexes for pairs of the six mechanisms simulated by Beber and Scacco (2012). Mechanisms (e) and (f) devised by Mebane (2006). For pairs of the first four scenarios, N=400, for the rest N=169.

A.3 Empirical Demonstration

A.3.1 Sweden

Table A.2 reports the last digits classified by numeral, party, and result from the Swedish parliamentary elections of 2002 for the two parties with the largest national vote totals.

Table A.2: Last digits of ward-level vote counts with three or more digits in the 2002 Swedish parliamentary elections for two parties with the largest national vote counts. 5,963 wards inspected, 13 wards where the two parties tied excluded. Source: author's calculation. N=8,940. Data source: Beber and Scacco (2012).

	SAP		MSP	
	Won	Lost	Won	Lost
0	538	54	39	285
1	492	57	43	292
2	533	62	33	305
3	527	61	37	292
4	526	57	41	305
5	496	52	46	268
6	528	62	33	278
7	488	55	35	294
8	514	43	34	270
9	518	46	35	266

As reported in Table A.3, both NHST and latent class methods indicate that uniformity describes all inspected subsets of numerals well. Under the strong distributional assumption this is evidence of no fraud.

Table A.3: Evaluation of uniformity for ten subsets of last digits from the Swedish data. Fraction sizes in %. Reference distributions of tests statistics obtained with one million simulations.

	N	χ^2	p	π^*	Δ
SAP Victory	5,160	5.63	0.78	5	1
SAP Loss	549	6.87	0.65	22	4
SAP	5,709	6.75	0.66	5	2
MSP Victory	376	4.85	0.85	12	5
MSP Loss	2,855	6.66	0.67	7	2
MSP	3,231	6.32	0.71	7	2
Victory	5,536	4.50	0.88	6	1
Loss	3,404	10.15	0.34	8	2
Registered	5,962	7.75	0.56	9	1
Total	5,959	3.54	0.94	4	1

Under the relaxed distributional assumption the numeral subsets can be compared using a series of loglinear models that represent different processes by lifting a different set of restrictions. The independence model allows the number of last digits to vary across parties and ward result, and some numerals to be more common than others. The second model further allows for one party to win/lose more wards than the other. Under this model the same probability distribution describes the numerals of both parties wards or only those won/lost by the party. This corresponds to equal kind and amount of fraud for the inspected subsets. The remaining three models further allow numeral probabilities to vary across parties, ward results, and both, respectively, corresponding to different amounts of fraud across the inspected subsets.

Table A.4: Fit of the five loglinear models to the Swedish data. N=8,940. Fraction sizes in %. Jackknifed confidence intervals.

	χ^2	df	p	π^*	95% ci		Δ	95% ci	
Independence	5,439.38	28	<0.01	34	(27,	41)	36	(36,	37)
Party-Result, Numeral	15.98	27	0.95	4	(1,	6)	1	(1,	2)
Party-Result, Numeral-Party	11.00	18	0.89	2	(0,	4)	1	(0,	1)
Party-Result, Numeral-Result	9.53	18	0.95	2	(0,	6)	1	(0,	1)
Party-Result, Numeral-Party, Numeral-Result	4.51	9	0.87	1	(0,	3)	1	(0,	1)

As shown in Table A.4, the assessment is similar under NHST and the latent class approach. Unsurprisingly, since the two parties won different numbers of wards, the independence model fits badly. The second model delivers a near perfect fit with a π^* of 4% and Δ of 1%, and is not rejected by the χ^2 test of fit at the 5% level. Lifting further assumptions is left with little room to improve the fit.

A.3.2 Nigeria

Table A.5 reports the data–last digits of numbers with three or more digits in polling station returns from the Plateau state for the two electorally strongest parties–by numeral, party, and polling station result, excluding the seven tied polling stations.

Table A.5: Last digits of polling station level vote counts with three or more digits in the 2003 Nigerian presidential elections in the Plateau state for two parties with largest vote counts. Seven polling stations where the two parties were tied are excluded. N=3,045. Source: author's calculation. Data source: Beber and Scacco (2012).

	ANPP		PDP	
	Won	Lost	Won	Lost
0	78	30	195	27
1	59	35	171	27
2	58	36	183	25
3	88	38	198	27
4	52	32	200	23
5	50	26	169	21
6	58	25	170	17
7	69	32	189	21
8	86	25	178	28
9	78	23	170	28

Fit of uniformity to several subsets of the data is reported in Table A.6. Several sets of digits depart noticeably from uniformity, including some of the parties' numbers as well as numerals from the counts of registered and total voters. Uniformity is rejected for ANPP's returns where it won by the χ^2 test at the 5% significance level. The π^* and Δ distances from uniformity are similar for votes for ANPP where it won, lost, or both, and for all votes for PDP as well as for votes where it lost.

Table A.6: Evaluation of uniformity for ten subsets of last digits from the Nigerian data. Fraction sizes in %. Jackknifed confidence intervals. Reference distributions of tests statistics obtained with one million simulations.

	N	χ^2	p	π^*	95% ci	Δ	95% ci
ANPP won	676	26.40	<0.01	26	(6, 46)	9	(5, 13)
ANPP lost	302	8.20	0.52	24	(0, 54)	7	(2, 13)
ANPP	978	20.53	0.02	22	(6, 39)	6	(3, 9)
PDP won	1,823	7.64	0.57	7	(0, 21)	3	(1, 5)
PDP lost	244	5.18	0.82	30	(0, 62)	6	(1, 12)
PDP	2,067	8.06	0.53	10	(0, 22)	3	(0, 5)
Lost	546	8.14	0.52	23	(1, 45)	5	(1, 9)
Won	2,499	16.08	0.07	12	(1, 23)	3	(1, 5)
Registered	2,565	113.54	<0.01	17	(7, 28)	7	(5, 8)
Total	2,546	292.28	<0.01	21	(10, 31)	10	(9, 12)

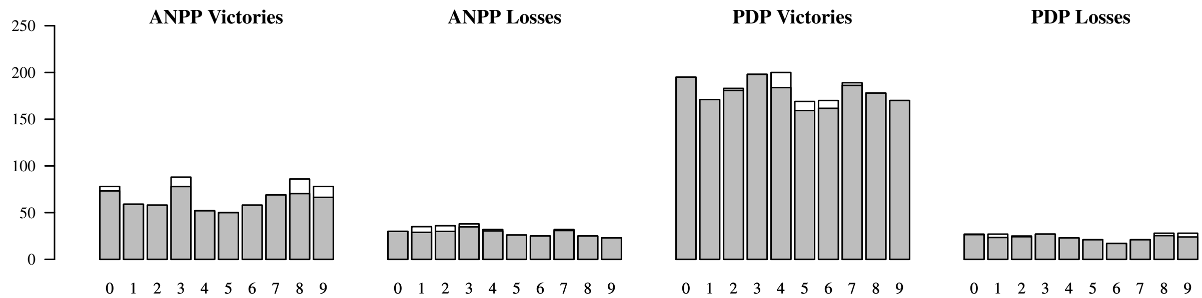

Figure A.4: Model fit under Δ of the second loglinear model (result-party, numeral) to the Nigerian data. Observations that do not need to be reallocated in grey, residuals in white.

Table A.7: Fit of the five loglinear models to the Nigerian data. N=3,045. Fraction sizes in %. Jackknifed confidence intervals.

	χ^2	df	p	π^*	95% ci	Δ	95% ci
Independence	193.53	28	<0.01	16	(11, 20)	8	(7, 10)
Party-Result, Numeral	28.03	27	0.41	10	(4, 15)	4	(2, 5)
Party-Result, Numeral-Party	18.87	18	0.40	5	(2, 9)	2	(1, 4)
Party-Result, Numeral-Result	23.11	18	0.19	7	(3, 10)	3	(2, 5)
Party-Result, Numeral-Party, Numeral-Result	13.26	9	0.15	2	(0, 4)	2	(1, 3)

Under the relaxed distribution assumption the data can be inspected by the same loglinear models as in the Swedish case. Table A.7 reports their fit. The independence model does not fit well, again due to the different numbers of territories carried by the parties. The second model fits markedly better, especially in terms of χ^2 and Δ, and only the two least restrictive models fit near perfectly according to both latent class indexes. The modeled and residual frequencies under Δ are shown in Figure A.4. Substantively, the extent of contamination by fraud appears roughly similar across all the inspected subsets. The subset that most stands out are PDP's numbers from wards carried by it, which might be considerably more or less contaminated then the other subsets.

A.4 Calculating π^* of Loglinear Models with `pistar`

The mixture index for the loglinear models were computed with the `pistar` package (Medzihorsky, 2015a) for the R language (R Core Team, 2014). Its version used here is included in the replication archive (Medzihorsky, 2015b). The development version is available at `https://github.com/jmedzihorsky/pistar`, and can be installed using the `devtools` library (Wickham and Chang, 2015):

```
library(devtools)
install_github("jmedzihorsky/pistar")
```

The package allows to compute the mixture index for a variety of model families. For loglinear models, the following can be used

```
p0 <- pistar(proc="ll", data=d, margin=m, jack=FALSE, eps=1e-1, iter=1e3)
```

where `proc="ll"` indicates a loglinear model, `d` is an object of class `array` with the data, `m` the model specification as for `loglin()`, `jack` indicates whether jackknife should be performed, and `eps` and `iter` are arguments passed to `loglin()` discussed in more detail below. The `loglin()` function, which is included in the standard `R` distribution, fits loglinear models using the Iterative Proportional Fitting algorithm (Haberman, 1972). In `pistar`, it is used within an EM algorithm (Dempster et al., 1977) that fits a two-point mixture of a loglinear model and an unrestricted component (Rudas et al., 1994).

The `eps` and `iter` arguments, shown above in their default values, might require adjustments depending on the problem. The first indicates the largest admissible deviation between the observed and the fitted margins, and the second the maximum number of iterations allowed in each call of `loglin()`. In general, lower (stricter) values of `eps` should be used with larger `iter` values, and will increase computation times. The default values shown above were chosen as providing an appealing speed-precision trade-off. In many settings, the defaults produce estimates of π^* identical to those under more strict (lower) `eps` when transformed to rounded percent, while making the computation several times faster. However, in some cases they can produce suboptimal results. For that reason, it is recommended to test the sensitivity of the results by using also lower values of `eps`, which should be accompanied by an increase of `iter`. The analyses reported here used 10^{-6} as the value of `eps` and 10^4 as the value of `iter`, which produced results identical in rounded percent to those under a variety of other settings.

References

Beber, B. and A. Scacco (2011). Replication data for: What the numbers say: A digit-based test for election fraud. `http://hdl.handle.net/1902.1/17151`. Harvard Dataverse, V2.

Beber, B. and A. Scacco (2012). What the numbers say: A digit-based test for election fraud. *Political Analysis 20*(2), 211–234.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

Haberman, S. J. (1972). Algorithm AS 51: Log-linear fit for contingency tables. *Applied Statistics*, 218–225.

Mebane, W. R. (2006). Election forensics: the second-digit Benford's law test and recent American presidential elections. In *Election Fraud Conference*.

Medzihorsky, J. (2015a). pistar: Rudas, Clogg and Lindsay mixture index of fit. `https://github.com/jmedzihorsky/pistar`. R package version 0.5.2.5.

Medzihorsky, J. (2015b). Replication data for: Election fraud: A latent class framework for digit-based tests. `http://dx.doi.org/10.7910/DVN/1FYXUJ`. Harvard Dataverse, V1 [UNF:6:FIWHvsHNzZgPStT0+kgbsQ==].

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rudas, T., C. Clogg, and B. Lindsay (1994). A new index of fit based on mixture methods for the analysis of contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological) 56*(4), 623–639.

Wickham, H. and W. Chang (2015). devtools: Tools to make developing R code easier. `http://CRAN.R-project.org/package=devtools`. R package version 1.5.