Bolsa Família and the Shift in Lula's Electoral Base, 2002-2006: A Reply to Bohn

Cesar Zucco & Timothy Power

WEB-APPENDIX: Further Details on Replication Efforts

The sample:
In an effort to make our analysis and replication attempt as transparent as possible, we detail here how we arrived at the sample that we effectively analyzed. Most of our focus is on the dependent variables, because these are the less straightforward to measure. Since vote recall in each round of the 2002 and 2006 elections was recorded in slightly different formats by LAPOP 2007, it was necessary to follow different procedures regarding each question in order to adopt the same criteria.

Table A1 reports the questions in the survey, and corresponding variables in the data set, that are relevant to the case at hand. The shaded cells contain the actual voting recall questions, while non-shaded cells are auxiliary questions. We also report the number of missing observations in the original data file, and the number of missing observations when we ran the analysis.

Observations that were "missing in file" mean different things in each question, as we detail below. This is because each of the four recall questions were asked in different ways. For some, there was a preceding question that filtered out "don't remembers" or "didn't votes," whereas other dealt with all or some of these possibilities in the same question. The main purpose of this section is to explain how we got from the values in the third column to the values in the last column.

Table A1 - Vote Recall Related Questions in LAPOP 2007

Variable Name	Approximate Question (Presidential Elections)	Missing in file	Substantive Missing
VB2	Did you vote in 2006?	2	
BRAVB3	Who did you vote for in 2006, 1ST round?	269	284
BRAVB4	Who did you vote for in 2006, 2nd round?	2	284
VOL1	Did you vote in 2002?	24	
VOL2	Do you remember who you voted for in 2002, 1st?	290	
VOL2A	Who did you vote for in 2002 1st?	688	691
VOL3	Did you vote in 2002, 2nd?	335	
VOL3A	Who did you vote for in 2002 2nd?	370	652

As we explain in the paper, we determined that useable recall observations are the ones that can be linked to a candidate that was on the ballot, or to having cast an invalid ballot (either null or blank). As the analysis is one of support for the incumbent, we believe going to the polls and casting a blank or null ballot can be read as non-support for the incumbent, especially given that electronic voting in Brazil has reduced invalid votes due to error. To cast a blank or null vote, one has to do it on purpose. Furthermore dropping invalid votes would further increase Lula's vote shore in the sample to close to 80%.

All other observations are substantively missing, in that they do not convey information about how the respondent recalls voting. We recognize that abstentions are different from nonresponses or bad responses, and we report in Table 1 of the paper the share of the total sample that declared to have not voted. We do not, however, take those into consideration when we compute vote shares. This is simply to maintain symmetry with the way in which we report actual electoral results. As the share of abstentions in the survey is close to the true share of abstentions, including these in the computations of vote shares in *both* electoral results and survey recall would not change the comparisons we presented in Table 1 of the survey.

Table A2 - Recoding of Second Round Recall Questions

Variable		Raw Freq	Final sample	Dummy
BRAVB4 (2006 2nd round)	Lula	678	681	Coded as 1
	PT	3		
	Alckmin	210	212	Coded as 0
	PSDB	2		
	Nenhum (branco/nulo)	37	37	Coded as 0
	Não Lembra	42	No/Bad Info	
	Outro	12	No/Bad Info	
	NR	21	No/Bad Info	
	Não Votou	207	Abstention	
	Missing in File	2	No/Bad Info	
VOL3A (2002 2nd round)	Lula	429	429	Coded as 1
	Serra	120	120	Coded as 0
	Branco/Nulo	13	13	Coded as 0
	Não Lembro	272	No/Bad Info	
	Não Votou	2	Abstention	
	Outro	8	No/Bad Info	
	Missing in File	370	Abstention	

Observations that were "missing in file" for the second round of the elections in 2002 were those that had answered a previous question saying they had not voted. On the other hand, for 2006 those observations are really simply missing data, as abstentions were captured by the main vote recall question. For the first round of both elections, "missing in file observations" include different sets of respondents, whose true meaning is obtained from looking at their answers to the other support questions listed in Table 1.

Table A3 - Recoding of First Round Recall Questions

Variable		Raw Freq	Final sample	Dummy
BRAVB3 (2006 1st round)	Lula	657	657	Coded as 1
	Alckmin	195	195	Coded as 0
	Helena	32	32	Coded as 0
	Buarque	9	9	
	Bivar	0		
	Rangel	0		
	Eymael	1	1	Coded as 0
	Pimenta	0		
	Nenhum (branco/nulo)	36	36	Coded as 0
	Outro	15	No/Bad Info	
	Missing in File	269	No/Bad Info (74)	
			Abstention (193)	
VOL2A (2002 1st round)	Lula	395	395	Coded as 1
	Serra	93	93	Coded as 0
	Ciro	13	13	Coded as 0
	Garotinho	9	9	Coded as 0
	Rui Pimenta	1	1	Coded as 0
	Eymael	1	1	Coded as 0
	NS/NR	4	No/Bad Info	
	Outras respostas	10	No/Bad Info	
	Missing in sample	688	Blank/Null (11)	Coded as 0
			Abstention (264)	
			No/Bad Info (413)	

The shaded cells in Tables A2 and A3 represent the actual valid responses for vote recall. The sum of these values with the "substantive missing" column in Table A1 is 1214, which is the number of people interviewed in the survey. All models presented in our paper used the 2nd round results as the dependent variable, because this is what we suppose was done in the original piece. Below, in this appendix, we provide results for the 1st round of the 2006 elections. The N reported in the regression table (Table 1) in our paper is slightly smaller than what is reported in Tables A2 and A3 because of missingness in other variables, most notably income.

Additional Results

In Table 1 of the main body of the paper, we report an attempt to replicate the specification that was presented in Bohn's Table 3, as well as results for a similar specification in which we recoded age and education into a continuous scale and entered income in the regression as the categorical variable it is. However, as we report in Table A4, below, the recoding itself is almost immaterial.

Table A4 - Results Using Bohn's Coding But Treating Categorical Variables as Such

	2002	2006
Bolsa Familia (Beneficiary=1)	0.531	0.351
p-value	0.028	0.056
Gender (Male=1)	0.014	0.125
	0.91	0.193
Age (25-34)	0.187	-0.181
	0.406	0.242
Age (35-44)	0.154	-0.263
	0.495	0.088
Age (45-59)	0.103	-0.362
	0.649	0.021
Age (More than 59)	0.137	-0.38
	0.616	0.048
Income (2-5 MW)	0.028	-0.05
	0.863	0.689
Income (5-10 MW)	0.143	-0.239
	0.493	0.121
Income (10-20 MW)	0.28	-0.208
	0.29	0.315
Income (20 MW or more)	0.031	-0.409
	0.949	0.267
Education (1st-8th grade)	0.64	0.122
	0.04	0.665
Education (High-school)	0.478	-0.235
	0.148	0.428
Education (College)	0.264	-0.55
	0.471	0.091
Center-West	-0.257	0.239
	0.265	0.186
North	0.438	0.431
	0.143	0.064
Northeast	0.301	0.42
	0.063	0.001
South	0.053	-0.123
	0.78	0.362
Intercept.	-0.151	0.841
	0.706	0.009
N	531	878
Probit models were estimated on LAPOP 2007 data, as described earlier. P-values are shown below the estimates.		

If we stick to the coding that Bohn describes in the paper, and simply enter the variables age, income, and education as a series of dummies for each category, the results are all but identical to what we report in the paper. It is noteworthy that in this specification, the older age group does not seem to vote for Lula at a higher rate, as would be suggested by Bohn's argument about the Benefício de Prestação Continuada (BPC). We are not, it should noted, making any thorough evaluation of the role the BPC played in the election, and it is an issue that deserves further

attention. However, LAPOP 2007 data provide no basis of support for the claims made in the paper regarding the BPC.

Although we believe it makes substantive sense to consider blank/null votes as votes cast against the incumbent, we report, in Table A5 results for the exact same models presented in the main paper after dropping blank/null vote recall. The coefficient for BF is very similar to what we found before, but it falls just shy of conventional levels of statistical significance.

Table A5 Results Excluding Blank/Null Votes

	Our Attempt at Replication	Correction for Categorical Vars.	Balanced data-set
Bolsa Familia (Beneficiary=1)	0.303	0.300	0.325
p-value	0.100	0.108	0.132
Gender (Male=1)	0.152	0.141	-0.091
	0.125	0.157	0.671
Age*	-0.078	-0.008	-0.015
	0.060	0.024	0.078
Income (5-point).	-0.117		
	0.045		
Income (02-.05.MW)		0.056	0.101
		0.813	0.734
Income (05-10 MW)		0.103	0.249
		0.684	0.532
Income (More than 10 MW)		-0.182	0.112
		0.475	0.875
Schooling*	-0.270	-0.054	-0.017
	0.001	0.001	0.602
Center-West	0.113	0.088	0.388
	0.536	0.632	0.408
North	0.321	0.329	0.412
	0.181	0.168	0.509
Northeast	0.286	0.299	0.388
	0.028	0.022	0.112
South	-0.250	-0.235	0.472
	0.063	0.082	0.182
Intercept.	1.737	1.332	1.238
	0.000	0.000	0.027
N	846	846	238

* Age and schooling are measured as Bohn proposed in the models in first two columns, but were recoded into a true continuous variable in the other models. *P*-values are shown below the estimates. Data are from LAPOP 2007.

Table A6 reports estimates for the recall in the 1st round of the 2006 election. The BF effect estimated here is smaller than what we found for the second round, and the coefficients are on the border of statistical significance for the first two specifications, and plainly not significant when the data set is balanced through matching.

Table A 6 Results for 1st Round or 2006 Elections

	Our Attempt at Replication	Correction for Categorical Vars.	Balanced data-set
Bolsa Familia (Beneficiary=1)	0.273	0.278	0.175
p-value	0.107	0.107	0.384
Gender (Male=1)	0.079	0.064	-0.421
	0.401	0.497	0.038
Age*	-0.040	-0.005	-0.007
	0.308	0.128	0.351
Income (5-point).	-0.100		
	0.070		
Income (5-point).		0.043	0.139
		0.846	0.613
Income (02-.05.MW)		0.136	0.624
		0.564	0.112
Income (05-10 MW)		-0.067	-0.232
		0.778	0.708
Schooling*	-0.332	-0.072	-0.044
	0.000	0.000	0.144
Center-West	0.050	0.030	0.594
	0.769	0.858	0.147
North	0.383	0.422	0.909
	0.090	0.061	0.132
Northeast	0.377	0.426	0.812
	0.002	0.001	0.001
South	-0.231	-0.206	0.541
	0.074	0.113	0.097
Intercept.	1.597	1.150	0.895
	0.000	0.000	0.070
N	880	880	255

* Age and schooling are s measured as Bohn proposed in the models in first two columns, but were recoded into a true continuous variable in the other models. P-values are shown below the estimates. Data are from LAPOP 2007.

Table A7 reports results for an alternative specification of the main independent variable (BF participation). While Bohn provides no details of how she operationalized participation in the BF program, we assumed in the results presented in the main body of the paper that it entailed simply observing the answers to the question of whether the respondent participated in the program.

Table A 7 Results with Alternative Operationalization of the BF Participation

	Our Attempt at Replication	Correction for Categorical Vars.	Balanced data-set
Bolsa Familia (Beneficiary=1)	0.149	0.144	0.261
p-value	0.349	0.372	0.197
Gender (Male=1)	0.112	0.107	-0.059
	0.238	0.263	0.772
Age*	-0.097	-0.010	-0.009
	0.014	0.003	0.262
Income (5-point).	-0.112		
	0.046		
Income (02-.05.MW)		0.123	0.053
		0.579	0.848
Income (05-10 MW)		0.113	0.168
		0.634	0.655
Income (More than 10 MW)		-0.105	0.082
		0.661	0.905
Schooling*	-0.311	-0.063	-0.006
	0.000	0.000	0.849
Center-West	0.237	0.202	0.566
	0.187	0.259	0.219
North	0.415	0.429	0.173
	0.073	0.062	0.740
Northeast	0.416	0.426	0.593
	0.001	0.001	0.010
South	-0.117	-0.096	0.676
	0.372	0.463	0.049
Intercept.	1.761	1.303	0.782
	0.000	0.000	0.120
N	878	878	243

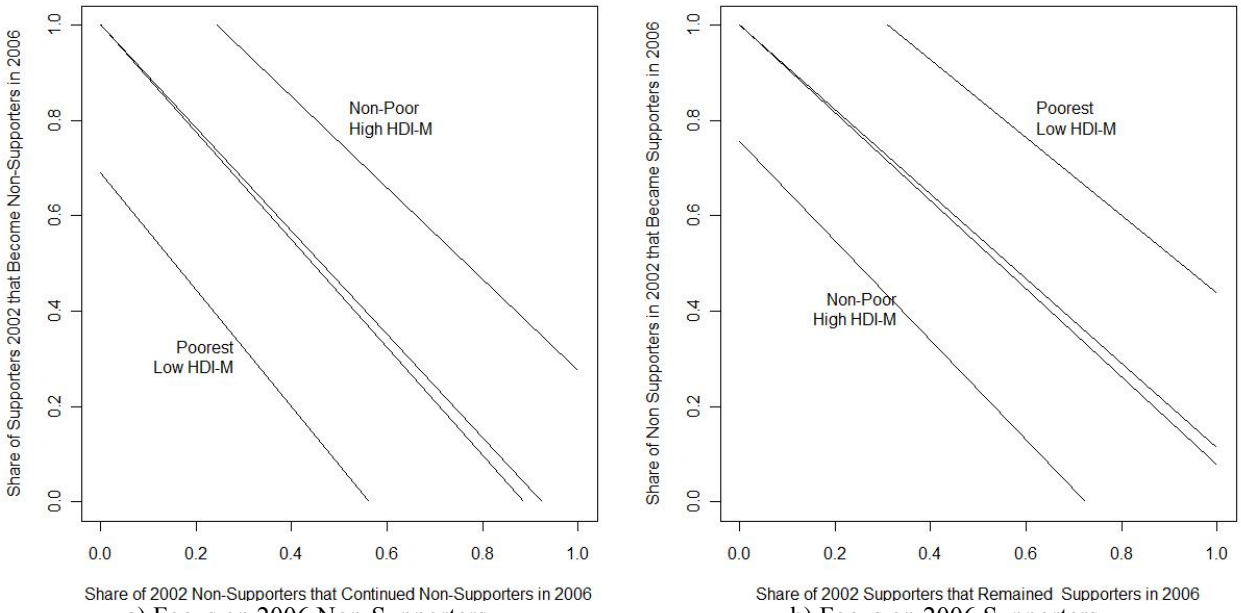
* Age and schooling are s measured as Bohn proposed in the models in first two columns, but were recoded into a true continuous variable in the other models. P-values are shown below the estimates. Data are from LAPOP 2007.

LAPOP 2007, however, also asked whether respondents participated in Bolsa Escola, Bolsa Alimentaçao and Vale Gás. By 2006, almost all beneficiaries of these other programs had been absorbed by BF, but it is plausible that people still referred to them by their original names. So we experimented with an alternative conceptualization of BF participation, where we coded the observation as participating in the program (1) if there was an positive answer to any of these questions. This shifted 27 respondents from non-beneficiaries to beneficiaries, and though beneficiaries in the sample still vote for Lula at a much higher rate than non-beneficiaries (85% to 72%), this difference is smaller than what we found for the other conceptualization of the dependent variable (88% to 72%), and the regression returns coefficient on BF that are not significant.

To wrap up, though BF seems to be associated with an increase in the probability of voting for Lula in most of the specifications we ran, in some specifications this coefficient was not statistically significant. This was the case when we looked at first round results and if when we expanded definition of BF participation. There is no indication in Bohn's paper, however, that this is what was done. In no specification did we find a negative coefficient on BF. Much more importantly, this exhaustive search for a model that resembles what was reported by Bohn is quite futile in light of the much larger problems of external validity of this particular voter recall data. Plainly put, no analysis of these recall data can be believed.

Bounds on Share of Voters that Switched to/from Lula

In the main body of the paper, we make brief reference to the information conveyed about switches to/from support for Lula in the analysis of support for Lula for four different sub-groups of voters. The statements we made are backed-up by the following tomography plots that show the possible ranges of switches for each group. We present two complementary graphs for ease of interpretation, but the information in one is implicitly contained in the other. For two of the groups (non-poor in less developed places and poorest in more developed places) the bounds lines are very similar and span almost the entire range of the plots in both dimensions. This indicates that almost any share of these groups might have switched (or not). For the other two groups in the Figure, the fact that the lines span only a limited share of the figure reveals that the bounds the possible number of switchers convey important information. C Consider the case of the poorest voters in the less developed places, in panel b. The projection of the line onto the y-axis indicates that between just under 50% and 100% of non-supporters of Lula in 2002 became supporters in 2006. The projection of this same line onto the x-axis indicates that between just under 30% and 100% of Lula supporters in 2002 remained Lula supporters in 2006. Though we cannot know for sure what happened, we do know that Lula attracted at least half of previous non-supporters in this group.



a) Focus on 2006 Non-Supporters b) Focus on 2006 Supporters

Figure A1 – Tomography Plots for Four Groups of Voters