Planned Missingness with Multiple Imputation: enabling the use of exit polls to reduce measurement error in surveys

Marco A. Morales	René Bautista
Wilf Family Department of Politics	Survey Research and Methodology (SRAM)
New York University	University of Nebraska-Lincoln
marco.morales@nyu.edu	rbautis1@bigred.unl.edu

This version: March 17, 2008.

Abstract

Exit polls are seldom used for voting behavior research despite the advantages they have compared to pre and post-election surveys. Exit polls reduce potential biases and measurement errors on reported vote choice and other political attitudes related to the time in which measurements are taken. This is the result of collecting information from actual voters only minutes after the vote has been cast. Among the main reasons why exit polls are not frequently used by scholars it is the time constraints that must be placed on the interviews, i.e., short questionnaires, severely limiting the amount of information obtained from each respondent. This paper advances a combination of an appropriate data collection design along with adequate statistical techniques to allow exit polls to overcome such a restriction without jeopardizing data quality. This mechanism implies the use of Planned Missingness designs and Multiple Imputation techniques. The potential advantages of this design applied to voting behavior research are illustrated empirically with data from the 2006 Mexican election.

Prepared for presentation at the Annual Conference of the Midwest Political Science Association, April 3-6, 2008, Chicago, IL. We owe a great debt of gratitude to Francisco Abundis at *Parametría* who allowed us to test PM-MI empirically in one the most complicated - and close - elections in Mexican history. This paper extends a method introduced in Bautista, Morales, Abundis & Callegaro (in press).

Scholars have made extensive use of pre and post-election surveys to study public opinion and voting behavior. An often overlooked problem with this data is the potential measurement error that is a function of the moment in time in which the survey is conducted. For example, voting behavior studies rely on measures of vote choice that are taken days, weeks or even months before or after an election which may differ from the votes cast because these measurements fail to account for several effects related to the difference between the time of the election and the time when interviews were conducted. Error is derived from timing alone. Interestingly enough, these measurement errors can be minimized with Election Day surveys - popularly known as exit polls - because they collect vote choice, along with other key variables, immediately after voters have cast their ballots.

Despite the fact that exit polls are able to collect information in a timely manner, there are still limitations that have led scholars to underestimate such an advantageous feature to study voting behavior. Arguably, the most critical limitation of exit polls is that short questionnaires are typically used. This paper discusses how to overcome such a constraint by combining data collection designs with statistical mechanisms. In doing so, this paper also illustrates a venue where survey methodologists, public opinion scholars and political scientists can look eye to eye in order to further the development of data collection designs that can improve the quality of the data. The exit polling design being discussed combines *Planned Missingness* and *Multiple Imputation* mechanisms; concepts that have been advanced previously in the literature but, to the best of our knowledge, never fully implemented in an exit polling context. This combined mechanism fits within the new paradigm emerging in survey research that focuses on the improving survey designs to minimize sources of errors, especially when they are under the control of the researcher (Weisberg 2005).

The utility of this *missing-by-design* exit poll is illustrated in the context of one of the most contested elections in recent times in Mexico, the 2006 Presidential election. The paper is divided into four parts. Section 1 details the advantages of using exit polling data and the potential reduction of measurement errors that can result from it. Section 2 details the Planned Missingness-

Multiple Imputation (PM-MI) mechanism advanced here. Section 3 illustrates empirically some of these advantages of using the PM-MI exit polling design. Lastly, section 4 summarizes this design and discusses its potential drawbacks and applicability.

1 Reducing measurement error through exit polls

Measurement error that results from the time at which a survey is taken can affect the typical dependent variable in vote choice studies as much as other variables in the analysis. Consider first the case of vote choice. There are, at the very least, two sources of bias on reported vote choice (Tourangeau, Rips & Rasinski 2000, Tourangeau & Yan 2007): the first relates to voters who did not turn out on Election Day reporting that they did vote and vice versa, and the second to "true" voters failing to reveal who they actually favored with their ballot.[1]

The first concern is addressed directly with the exit poll design. By definition, people surveyed in exit polls are actual voters, which is a direct result of asking questions to individuals as they leave the polling places. Therefore, the risk of having non-voters accounted for as voters in the analysis is effectively eliminated. In essence, this is a natural result of sampling from two different populations with each design. It is much harder to "filter" likely voters when sampling from a population of adults (as is the case in pre and post-election surveys) than simply using responses sampled from a universe of actual voters (as is the case in exit polls).[2]

The second concern - voters not giving an accurate account of their choice - might be due to voters forgetting who they voted for, social desirability, or simply campaign events happening after the interviews that changed a voter's choice. These effects can be minimized in exit polls. Reported vote choices could be more accurate as only minutes have passed since voters left the polls and

[1] Of course, there is always the risk that respondents do not provide truthful responses, but no survey design can effectively eliminate this problem.

[2] For instance, recently pre election polls conducted to gauge vote intention among Democrats for the primary election in New Hampshire predicted a sound victory of Barack Obama over Hillary Clinton; however, the actual winner turned out to be the latter, raising questions on the ability to collect information on actual voters using pre election polls. To our interest, exit polls successfully measured vote intention in New Hampshire providing better data on voters.

should more clearly remember their actual vote choice. There is no incentive for respondents to appear as having voted for the winner of the election as they are still uncertain about the outcome of the election. And, as vote choice is asked minutes after the vote is cast, there are no more campaign events that might modify vote choice later on.

Exit polls can also potentially contribute to reduce measurement error in other variables, asides from vote choice. A couple of interesting examples can illustrate potential improvements in measurements of attitudes: Party ID and economic performance evaluations. Perhaps one of the most studied variables in voting behavior is party identification. Since the 1950s, social psychologists and political scientists have been puzzled by the idea that voters have a relatively stable attachment to political parties that influences vote choice. But even the most stable accounts of party ID (Campbell et al. 1960) suggest that it might change over time. Furthermore, the most elaborated theoretical accounts conceptualize party ID as a dynamic variable that is constantly updated (Fiorina 1981, Achen 1992, 2002).[3] Hence, exit polls - by virtue of taking the closest possible measure to the time when the vote is cast - would produce the most reliable measurements of party ID assuming that it does change over time. No additional harm is done if it is stable over time.

A rich literature has developed on economic voting since Kramer's (1971) foundational empirical study. Evidence on the relationship between the state of the economy and vote choice has been consistently found for different countries over many years (Lewis-Beck & Stegmaier 2000, Anderson 2007). By now, nearly all academic surveys include questions on retrospective and prospective assessments of the state of the economy. Yet, there are reasons to be concerned about what these questions are measuring.[4] The point to stress here is that the closer these measurements are taken to the time in which the vote is cast, the more likely it is that the measures of the

[3] A related question is whether party ID is a cause or a consequence of vote choice. If party ID is causally related to vote choice, it would make little sense to produce inconsistent estimates - and biasing causal inference - by excluding it from the model specifications. In any case, this is not a matter to be addressed in survey design, and is ultimately a matter for researchers to decide based on their theoretical models.

[4] There is also a - valid - concern regarding the endogeneity of economic evaluations and vote choice (Wlezien, Franklin & Twiggs 1997, Anderson, Mendes & Tverdova 2004, Glasgow & Weber 2005, Ladner & Wlezien 2007, Lewis-Beck, Nadeau & Elias 2008). Unfortunately, this matter cannot be addressed through survey design as it requires theoretical underpinnings that have little to do with survey design.

perceived current and/or future state of the economy would capture the full assessment that voters are thought to use when selecting between candidates.[5]

Measurement errors reviewed above are a function of closeness to the time when vote choice is made and/or knowledge of the identity of the winner of the election. Fortunately, exit polls collect information minutes after the vote has been cast while the winner of the election is still unknown, thus minimizing these particular problems. In sum, exit polls pose advantages over other survey designs that make them better suited to analyze vote choice. While measurements taken in an exit poll might not eliminate all possible forms of measurement error, they are certainly not prone to increase it.

Surprisingly, exit polls are seldom used for voting behavior research despite the possible biases and measurement errors in pre and post-election surveys. Perhaps one of the most relevant reasons why exit polls are not commonly used by scholars analyzing elections is the time constraints that must be placed on the interviews. Evidently, this drastically reduces the amount of data that exit polls can collect from each voter. By the very nature of the survey process, exit poll interviews must be conducted rapidly. The usual three to five-minute interviews do not allow for lengthy questionnaires. If interviews were to last longer it is likely that the response rate might drop, that the sequence of the interviews might be altered if the n-th interviewee is missed, or that the interviewer might become tired sooner reducing the quality of the records of responses. All of these could potentially be sources of measurement error. For these reasons, the usual exit poll design limits the amount of information that can be collected as voters leave the polls.

[5] Researchers should be concerned though, with the vague phrasing of the prospective economic evaluation question. Are voters providing the assessment of the economy under the current incumbent? Under the party preferred by the voter? Or is it a party-free assessment? This is but another example of measurement error in the survey design, although this time due to the vague phrasing of the question.

2 Planned missingness with multiple imputation: overcoming limitations in exit polls

As described earlier, exit polls are a potentially exploitable instrument to analyze elections since they can overcome some limitations of other survey designs when measuring attitudes. The main challenge to enable exit polls to become a part of a researcher's tool kit is to design them so that they can collect substantial amounts of information without jeopardizing data quality. If the principal channel to alter data quality is the time it takes to conduct the interviews, then the design must *not* alter the length of the interviews. How is this to be achieved? By combining an appropriate data collection design with adequate statistical techniques. One such combination implies the use of Planned Missingness (PM) where various versions of the same size length questionnaires are applied in interviews, maintaining a set of common questions in all of them to produce databases with certain information common to all respondents and certain pieces of information common to subsets of respondents. It also implies the use of Multiple Imputation (MI) techniques to "fill in" the missing values and generate plausible values for the missing information, which will produce "completed" data set. In other words, the Planned Missingness design enables the collection of a plethora of information by using variations in the questionnaires, and Multiple Imputation allows for the simultaneous use of all the pieces of information *as if* it had been collected simultaneously. This notion builds on Warren Mitofsky's (2000) idea to implement different questionnaires on exit polls as means to gather more information on general descriptions of opinion and attitudinal variables. Even when PM was implemented on exit poll designs, the data was never combined - nor multiply imputed - to obtain larger data sets to be used to analyze voting behavior more thoroughly.

2.1 Planned Missingness: enhancing data collection

Planned Missingness is the name commonly given to survey designs in which the same target population is queried to answer different sets of questions, thus generating a controlled item non-response.[6] Briefly, different questionnaire versions are randomly administered to different subsamples of the population. As every individual in the sample answers a different questionnaire version, planned missingness generates various small-n data set that contain certain pieces of information from a given population subset. Theoretically, as a result of random assignment, each data set should reflect the parameters of interest for the target population with a given level of uncertainty.

Applied to an exit poll, planned missingness permits the collection of information on vote choice for all surveyed individuals, and different pieces of information relevant for modeling voting behavior from different subsets of voters. The key feature of this design is that data-missingness is not related to unobservables - or, alternatively, that conditional on the observed data, missingness is random. The survey design, then, permits the unbiased estimation of parameters of interest. For this feature to hold, it is important to assign the different questionnaire versions randomly to each population subset. Figure 1 below gives a graphical description of the data collected using this planned missingness design.

[Figure 1 about here]

For all practical purposes, the data-missingness in the design produces various small-n data sets that can be analyzed as if they were unrelated groups of observations. Naturally, using them in this manner might produce biased and inconsistent estimates as a result of excluding relevant information from the analysis. So the necessary additional step after collecting more information from voters is to enable its simultaneous use.

[6] Planned missingness has been used for research purposes in various settings, see for example Graham, Hofer & Piccinin (1994), Graham, Hofer & MacKinnon (1996) or Littvay & Dawes (2007).

2.2 Multiple Imputation: enabling simultaneous use of data

Multiple Imputation is one among the available procedures devised to deal with missing data. Originally proposed by Rubin (1977, 1987), multiple imputation is a model-based approach to assign plausible values for missing data conditional on observed data. It is not deterministic - in the sense that only one value is attributed to each respondent that can be replicated later on - but stochastic - in the sense that it produces a set of $m > 1$ plausible values (that cannot be exactly replicated) for each missing observation that are generated using the available information and taking into account the covariation among variables in the full data matrix. Briefly, the process consists of generating $m > 1$ data sets that make no changes to the observed data, but assigns different plausible values for the missing data. Figure 2 illustrates an intuitive way to think about this process.

[Figure 2 about here]

For this design to work, *ignorability* in the missing data mechanism must be assumed (Rubin 1976). That is, missingness must not be conditional on unobservables and can be ignored given the appropriate conditioning on the observed data.[7] In essence, the random assignment of the different questionnaire versions embedded in planned missingness guarantees *ignorability*, which should not be surprising as missingness is the result of research design so that "almost all of the missingness

[7] It is important, at this point, to distinguish between data-missingness generated by the survey design, and item non-response that is independent of the survey design. In the first case, data-missingness results from questions not being asked to subsets of the population thus being Missing Completely at Random (MCAR), in the sense that missingness cannot be predicted by the observed data (or by any unobserved variable). Formally,

$$P(R|D) = P(R)$$

where R is an indicator for missing data, D_{obs} is the observed data, and $D \in \{D_{obs}, D_{miss}\}$ is all data, both observed and missing. In contrast, item non-response results from individuals failing to provide information on certain items and is Missing at Random (MAR) in the sense that the probability of missingness depends exclusively on the observed data (by assumption). Formally,

$$P(R|D_{obs}) = P(R|D)$$

For the purpose addressed here, we are concerned with the data generating mechanism to the extent that it allows us to predict the missing data given observed data. That is, we care that ignorability can be assumed in both cases so that multiple imputation techniques are suitable (Rubin 1977, 1987). In our particular case, the observed data produced by planned missingness - and the estimated covariances for the observed data - allow us to estimate plausible values for individuals who were not asked particular questions.

is due to unasked questions" (Gelman, King & Liu 1998, 847). We also assume that non-missing values in the data set are good predictors for the missing values. Furthermore, the imputation is "clean" since all survey responses share sampling methodology, were collected on the same date, by the same polling organization, and all respondents answer a series of questions that are presumably adequate predictors of the missing values.

Certain features of planned missingness make the data particularly adequate for a proper imputation. Data-missingnes, for example, is governed by the same mechanism across the data set: planned missingness that is random (and hence MCAR). Also, item-nonresponse is governed by the same mechanism on each variable (which is MAR), and conditioning on the appropriate covariates can be made ignorable. Similarly, these features suggest that a covariance matrix common to all respondents can be reasonably assumed as they come from the same population surveyed at the same point in time. This is an issue that has worried researchers when dealing with imputation across surveys since incorrectly assuming a common covariance matrix could bias the imputations (Brehm 1998, Judkins 1998). Yet this should not be an issue in planned missingness on an exit poll, for the reasons sketched above. In sum, given that the data are collected over the same day, right after each one of the voters cast their ballot, absent ongoing campaigns, and while the outcome of the election is unknown with missingness being random (or arguably governed by a MAR process), the imputation can be confidently carried out.

Usual analyses can be performed on each on of these $m > 1$ completed data sets, and the results combined to produce asymptotically consistent and efficient estimates. Since the imputations carry a degree of uncertainty with them, it must also be incorporated in the estimates of the model. Therefore, the appropriate estimation involves three steps (Rubin 1987). First, impute m values for each of the missing observations, producing m data sets where observed values do not change but missing ones take different plausible values that reflect the uncertainty on the imputation. Second, perform the usual analysis (e.g., regression analyses) on each of the m data sets - that is, on the data set that are at this point are already imputed. And third, use these m estimates to compute point estimates and variances for the parameters of interest. By virtue of this procedure, we are able to use all the available data set from the exit poll *as if* all questions had been asked to all respondents (Gelman, King & Liu 1998) producing consistent and efficient estimates of the parameters of interest (Rubin 1987, King et al. 2001).

2.2.1 Other possible methods

Many possible methods have been proposed to deal with missing data, that could effectively also be applied to Planned Missingness situations: hot-deck imputation, cold-deck imputation, deductive imputation, conditional mean imputation, unconditional mean imputation, hot-deck random imputation, stochastic regression imputation, regression imputation, deductive imputation, exact match imputation, to name a few (Little 1992, Weisberg 2005). One recognized advantage of multiple imputation over other types of methods to deal with missing data is its ability to reflect the estimation uncertainty. This is a problem that needs to be addressed as single-value imputations would lead to underestimated variances (Rubin 1987). That is, instead of having one imputed value *as if it were the true value*, we can have $m > 1$ values from the predictive posterior distribution of the missing data. So, uncertain estimates will have high dispersion, while more certain imputations will lie tightly around its expected value. With this information, variances are computed taking into account the within-estimates variance and the between-estimates variance (see eq. A.3 on Appendix A) producing efficient estimates with a limited number of imputations (Rubin 1987, King et al. 2001).

Franklin (1989) devised a method to deal specifically with the type of problem at hand: some information available in one data set, but not in another data set, although both share additional auxiliary information. By treating both data sets as samples from the same population, the information in one set is used to obtain estimates of parameters that are used in the second data set to generate predicted values to "fill in" the missing values.[8] While this estimator is shown

[8] More formally, Franklin's (1989) 2SAIV estimator relies on the structural equations

$$\begin{aligned} y_1 &= x_1\beta = u \\ x_1 &= \mathbf{z}_1\gamma + \epsilon_1 \\ x_2 &= \mathbf{z}_2\gamma + \epsilon_2 \end{aligned}$$

to be consistent, it is not always efficient. But for these results to hold, the estimated parameters that are used to generate the predicted values and the variance of the error terms must be the same across data sets, which might not be the case if there is missingness in the auxiliary data set that is not MCAR. This might simply be a difficult sell in survey data, but a much easier problem to address with multiple imputation techniques than with any other.

Ultimately, multiple imputation is an operation performed on the data itself, which allows traditional econometric models to be applied with minimal additional complications (other that simple computations to estimate parameters of interest and variances). Franklin's estimator, on the other hand, would require to be developed in situations that depart from OLS. But perhaps only for simplicity and ease of application of well-known econometric models, multiple imputation seems to be a preferable alternative.

3 Testing PM-MI: Mexico 2006

Planned missingness was implemented in an exit poll conducted by *Parametría* - one of the largest independent Mexican polling firms - for the 2006 Mexican Presidential election. *Parametría*'s exit poll collected information from 7,764 voters, which yields an approximate sampling error of +/- 1.1% with 95% of statistical confidence. It is the result of a multistage sampling design where primary sampling units (*i.e.* precincts or "Electoral Sections") were selected with probability proportionate to size. The relative size of each cluster was the number of registered voters as determined by the Federal Electoral Institute (IFE). A total of 200 precincts were drawn as a nationwide sample. The number of interviews collected ranged from 7 to 70 voters depending on the precinct.[9]

where y_1 is a dependent variable of interest, x_1 and x_2 are predictors of y_1 in different data sets such that y_1 and x_1 are not observed in the same one; $\mathbf{z}_1$ and $\mathbf{z}_2$ are predictors of x_1 and x_2, respectively, found in the same data set. The estimator takes advantage of the fact that x_2 and $\mathbf{z}_2$ can be used to estimate γ by OLS, which is assumed to be the same across data sets. $\hat{\gamma}$ is in turn used to estimate $\hat{x}_1$, that is used to estimate β, thus solving the problem of missing data across data sets.

[9] The number of primary sampling units (precincts) were determined based on costs, being 200 primary sampling units in this particular exit poll. Although the expectation is to conduct at least ten interviews per cluster (i.e., precinct), the final number of interviews cannot be determined in advance.

In particular, a mix mode data collection method was implemented.[10] First, the interviewer approached the selected respondent in order to ask demographic questions and presidential approval.[11] Then, a blank facsimile of the official ballot was handed to the respondent who would deposit it in a portable ballot box.[12] Next, a final set of questions - which varied across interviewees - were administered to the respondent. Four different versions of the last portion of the questionnaire were administered rotatively and each version differed on the additional information that was asked. Hence, Version "A" asked respondents their recollection of having watched Fox administration's campaign advertisements, and whether respondents are beneficiaries of several social policy programs.[13] Version "B" asked respondents to assess which candidates and parties had produced the most negative or positive campaigns.[14] Version "C" asked respondents to place candidates, parties and themselves on a 7-point ideological scale, as well as their party ID and economic performance evaluations.[15] Version "D" did not gather any additional information relevant to this analysis.[16]Given the high missingness in the Planned Missingness-Multiple Imputation data, $m = 10$ data sets were imputed that "filled-in" the missing values on 37 variables. (See Appendix A for details.)

3.1 Potential bias in the projection of election results

Exit polls are implemented with two aims in mind: i) projection of election results, and ii) analysis of information on voters. One natural concern for exit pollsters is that modifications in the questionnaire might affect the quality of information collected on vote choice and alter the accuracy of the projections of election results. This was a salient concern for the 2006 Mexican election,

[10] A mixed mode method combines self- and interviewer- administered methods, which is the most suitable data collection option for populations with low levels of education, such as Mexico. For further details on mixed mode methods applied to the Mexican case see Bautista et al. (2007)

[11] In sum, information was collected for 7,764 voters on gender, age, presidential approval, vote choice for President and Congress, income, and education.

[12] Each ballot facsimile contained a control number that would allow matching reported vote choice with the information collected in the questionnaires.

[13] 2,032 voters received this version leaving 5,732 answers to be imputed

[14] 1,859 voters answered this version leaving 5,905 answers to be imputed.

[15] 1,795 voters replied to this version, leaving 5,969 answers to be imputed.

[16] This exercise was the result of a syndicated exit poll; unfortunately version "D" was not released for academic research.

which subjected the design to a particularly stringent test. Three months before the election, poll after poll confirmed that the race would be centered on the National Action Party (PAN) and the Party of the Democratic Revolution (PRD) candidates - Felipe Calderón and Andrés Manuel López Obrador, respectively - but also that it might be too close to call on election night. As can be seen in figure 3, exit poll estimates of the outcome of the election were highly accurate and within the margin of error, suggesting that the variations in the questionnaires did not generate any additional biases in vote choice estimates.[17]

[Figure 3 about here]

Typically, vote choice analyses are performed on pre and post-electoral data alike. But researchers should be wary of the potential error that measured vote has on these surveys that is related to the time in which the surveys are taken and knowledge about the identity of the winner of the election. This is potentially a problematic feature of the data, especially when it can generate biased and inconsistent estimates. Take the well-known case of OLS, where measurement error in the dependent variable can be ignored - as it would still produce unbiased and consistent, although less precise, estimates (Hausman 2001) - as long as it remains independent from other regressors.[18] But it is highly unlikely that certain biases - related to underdog or bandwagon effects - in reporting vote choice would be uncorrelated with unobservables that are also correlated with other regressors. This alone justifies advocating better measurements of vote choice, and analysts should be aware of this fact when using post-electoral surveys and deriving conclusions from analyses performed on

[17] Exit polling figures released on Election night were weighted to represent the population from which the sample was drawn. In this case, population-based weights were used: urbanicity, number of registered voters and number of actual voters as of the previous general election held in 2003.

[18] That is, if we define $y^* = y + \nu$ with ν as measurement error, y as the true value of the variable of interest and y^* as the variable measured with error, it is straightforward to show that OLS estimates of β are consistent as ν dissolves in the disturbance

$$\begin{aligned} y^* &= x\beta + \epsilon \\ y &= x\beta + (\epsilon - \nu) \end{aligned}$$

This does not cause any problems as long as plim $\frac{1}{n}\sum_{i=1}^{n} x_i\epsilon_i = 0$ and plim $\frac{1}{n}\sum_{i=1}^{n} x_i\nu_i = 0$. Unfortunately, if ν is correlated with x, then plim $\frac{1}{n}\sum_{i=1}^{n} x_i\nu_i \neq 0$, and plim $(\hat{\beta}) = \beta + \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\nu_i\right) \neq \beta$ thus generating inconsistent estimates (Greene 2003).

this data.

In order to show the particular advantage of exit polls - which extends to planned missingness-multiple imputation (PM-MI)- on this regard, a simple comparison of vote estimates is performed with pre and post-electoral measurements for this same election from the Mexico 2006 Panel Study.[19] The first estimate (Pre-election) corresponds to the estimate generated by the Mexico 2006 Panel study over the month prior to the election. The second estimate (PM-MI) is produced by the exit poll data. The third estimate (Post-election) corresponds to the raw estimate of the post-electoral survey of the Mexico 2006 Panel Study. The fourth estimate (Post-election rev) corresponds to the same Panel estimate but “corrected” to exclude non-voters. Since registered Mexican voters are issued a special ID or “electoral card” that is marked every time an individual casts a ballot, the survey included an indicator for those cases where the interviewer could directly verify the existence of the mark on the voter’s ID.[20] Figure 4 compares the point estimates and their associated theoretical sampling error for each of these estimates. The actual election results are denoted by the vertical line with the official percentage of vote marked above.

[Figure 4 about here]

It becomes obvious from figure 4 that the realized vote shares are always within the margin of error of the exit poll estimates. Unfortunately, the same cannot be said of pre and post-electoral survey data. While the estimates were accurate for the PAN candidate - and winner of the election - in the post-election wave, the estimates and margins of error were notoriously off for the PRI and PRD candidates whose votes were under and overestimated respectively in the post-election wave. The pre-election wave only produced a good estimate for the PRD candidate, but was notoriously

[19] Senior Project Personnel for the Mexico 2006 Panel Study include (in alphabetical order): Andy Baker, Kathleen Bruhn, Roderic Camp, Wayne Cornelius, Jorge Domínguez, Kenneth Greene, Joseph Klesner, Chappell Lawson (Principal Investigator), Beatriz Magaloni, James McCann, Alejandro Moreno, Alejandro Poiré, and David Shirk. Funding for the study was provided by the National Science Foundation (SES-0517971) and *Reforma* newspaper; fieldwork was conducted by *Reforma* newspaper’s Polling and Research Team, under the direction of Alejandro Moreno. `http://web.mit.edu/polisci/research/mexico06`.

[20] This is by no means a perfect measure, as some actual voters might not carry their ID with them at the time of the interview, thus potentially discarding actual voters from the sample. Doing so, reduces the sample nearly by half, but this only ensures that the voters included in the sample are actually voters and cannot affect the accuracy of reported vote choice.

off for the other two candidates. This is not unexpected since exit polls and other surveys sample from different populations: exit polls sample from actual voters, while pre and post-electoral surveys sample from potential voters and try to screen voters from among them. We would naturally expect exit polls to be more accurate than post-electoral surveys simply as a result of survey design.

3.2 Enhancing academic research

As more information is available from each voter, we are able to explore simultaneously more potential explanations for vote choice. Were we to use the typically limited number of variables in an exit poll for vote choice analyses, we would produce less efficient estimates that result from smaller-n data sets, and perhaps substantially different conclusions that might result from biased estimates due to omitted variables in the analysis.

Over the course of the 2006 campaign, a series of journalistic claims were advanced to explain the results of the election. First, the Presidency aired a promotional campaign that focused on the achievements of the President Fox administration - the so-called "continuity campaign" - that was said to have boosted the PAN candidate. Second, the PAN camp initiated an early negative campaign against the PRD candidate, who would later retaliate; it was said that the negative campaign affected the PRD candidate. Third, it was said that a series of cash-transfer and social spending programs by the Fox administration favored the PAN candidate. Fourth, it was said that the relatively good state of the economy - which implied no end-of-term economic crisis - had a positive impact on the PAN candidate. Finally, it was said that the high approval numbers of the outgoing President Fox produced coattails that helped the PAN candidate. Needless to say that these claims have not been thoroughly investigated empirically, and perhaps could not be if the relevant information has not been collected.

The PM-MI design applied to *Parametría*'s exit poll allowed the collection of sufficient information to evaluate the plausibility of the main journalistic claims advanced to explain the results of the 2006 election. Readers are spared of the lengthy table of coefficients produced by the multinomial probit analysis (which can be found in Appendix B). For ease of exposition, we present these estimates graphically (Gelman, Pasarica & Dodhia 2002) on figure 5 which summarizes our simulations of changes in probabilities - first differences (King 1998) - of a typical individual voting for candidate j given variations on a particular variable (see Appendix C for details on these simulations).

[Figure 5 about here]

From figure 5, we learn that the single most important predictor of vote for PAN's Felipe Calderón Hinojosa (FCH) was presidential approval, and that having a good evaluation of the state of the economy made voters more likely to favor him, as well as perceiving him as running a positive campaign. But also, that perceiving him to have run a negative campaign, and being a recipient of the poverty reduction program *Oportunidades* made voters less likely to vote for Calderón. We also learn that the single most important predictor of vote for PRD's Andrés Manuel López Obrador (AMLO) was perceiving him to have ran a positive campaign. But we fail to detect any effects of the state of the economy, presidential approval or even *Oportunidades* as the journalistic claims suggested. Finally, we learn that voting for PRI's Roberto Madrazo Pintado (RMP) was affected positively by disapproving President Fox, being perceived as running a negative campaign and, to our surprise, begin a beneficiary of *Oportunidades* despite the fact that this program was not controlled by PRI during the outgoing administration.

In sum, only one of the journalistic accounts find empirical support, namely that the single most important predictor for voting for Calderón was approval of President Fox. At the time we write, only two analyses of the 2006 Mexican election using survey data have been published (Moreno 2007, Estrada & Poiré 2007), although neither of them fully addresses the most common factors that are thought to have influenced the election. This is mostly due to the limited data available from the exit poll employed by these studies. The analysis reviewed here provides a more general and informative overview of the determinants of vote choice in the 2006 election. The results might deserve further discussion, but that falls out of the scope of this paper, which is to present and justify the use of PM-MI on exit polls.

4 Discussion and conclusions

To the best of our knowledge Planned Missingness coupled with Multiple Imputation has not been applied to exit polls - or to voting behavior research - previously. As the illustration from the Mexican 2006 election shows, the design does not seem to generate particular problems with the quality of the data being collected, but reduces measurement error caused by the survey design while it also enables a much richer data analysis. The point we want to stress is that, by mere design, measurement error derived from the time in which the surveys are taken and knowledge about the winner of the election is effectively minimized, hence producing estimates that are less likely to be biased and inconsistent as a result of this type of measurement error.

We recognize that an imputation is as good as the correlation between the observed and the missing covariates: the better the correlation across these variables, the more accurate - and efficient - the imputation will be (Brehm 1998, Binder 1998). Hence, an obvious topic in the agenda is to improve the design to enhance correlations across variables. This question was not explicitly addressed in the paper, but it is useful to briefly discuss it here.

One obvious way to incide in the quality of the imputation is through the patterns of the planned missingness. In our implementation of PM-MI in the 2006 Mexican election, we chose to create planned missingness with question blocks. That is, *questions that were not asked to all respondents* were only included in one questionnaire version (a block), with no overlaps across questionnaires. Other split-block designs where questions overlap in “Swiss-cheese” missing-data patterns (Judkins 1999) are also possible. Graham, Hofer & McKinnon (1996) show that estimates that use data from unique block designs are as efficient as those generated with split-block designs, although efficiency might be better in split-block designs depending on the correlations between and across questions in a block. They also show that estimates using data from question block designs become more efficient as the correlation of the questions within the block increases. Similarly, the efficiency of the estimates based on split-block design data depend highly on the correlations between blocks of questions; a finding that is corroborated by Raghunathan & Grizzle (1995). Therefore, it seems to be good practice to group blocks of questions in a way such that correlation is enhanced: between the questions in a block if grouping questions in unique blocks, or across blocks of questions if using split-block designs.

In view of the potential limitations of our design, it is useful to recount our reasons for choosing it. The first, and most obvious one, is that grouping blocks of questions by version is *logistically* much simpler to implement. From a practitioner stand point, it is paramount to avoid adding sources of confusion to the data-collection method. The second reason is a matter of custom in exit polling: it was Mitofsky's solution to enable exit polls to collect as much information as possible to meet various clients' needs using the same survey design. Instead of fielding different exit polls, different versions of a questionnaire were fielded out keeping a set of variables common to all questionnaires for consistency-verification purposes. Furthermore, a missing-by-design exit poll, as the one being discussed, is more likely to be encountered in real-life settings such as syndicated exit polls.[21]

An additional variable that can affect the quality of the imputations is the number of questionnaires that would be optimal to aim for in the planned missingness design and still get "good" (*i.e.* efficient and consistent) imputations. Alternatively, the same question can be rephrased as the number of questions to include on each questionnaire version in order to get "good" imputations.[22] There are two possible ways to answer these questions. On one end, holding sample size constant, the number of questions and/or questionnaires is related to the algorithm employed to impute. In other words, what is the lower bound for the number of variables to be used in the imputation that still produce efficient imputations? Simulations might provide useful guidance on this matter.

[21] All things considered, it might have been a better alternative to use a split-block design, although this is a more challenging alternative to implement for logistic reasons. That said, the distribution of questions within each questionnaire version does enhance the correlation across variables, as questions are grouped by topic.

[22] Not all questionnaires must have the same sample size, and it might make sense to have a particular block with a larger relative sample size if the question under investigation justifies this choice.

On the other end, the number of questions and/or questionnaires is closely related to sample size in the exit poll. The larger the sample size, the more questions and/or questionnaires could be included. Yet there is no standard "optimal" sample size for exit polls, as it is typically determined on a case-by-case basis as long as the final number of sampling observation units (*i.e.* voters) may vary as a function of turnout and response rates.

Interestingly enough, planned missingness has been recently implemented on national exit polls is the U.S., at least over the last four elections. This has certainly been a common practice of the National Election Pool (NEP), the consortium of news media and broadcasters responsible for carrying out exit polls on Election Day and providing tabulated data on vote choice. Potentially, this is a very rich source of information except for the fact that the questions asked do not always resemble the information needed for political scientists to model vote choice. Ideally, academics might be able to introduce a block of questions or a questionnaire version that inquiries on what is directly relevant to scholarly research. This is precisely the approach taken by *Parametría*'s exit poll in Mexico. In the U.S. case, such a possibility of course would depend on the board of the NEP.

An encouraging development is that exit polls seem to be more popular throughout the world. Over the past few years, an increasing number of countries in Latin America and Eastern Europe have had exit polls fielded on Election Day. If this trend continues, we may come to a situation where at least one exit poll is fielded for every election in a substantial number of countries. There is also a trend for exit polls to be syndicated. That is, given the higher cost of exit polls relative to pre and post-election polls, many stakeholders (*i.e.* political parties, media organizations, universities, think tanks, and others) share the costs of an exit poll. They do so on the condition that each one of them gets their own portion of "exclusive" questions, and access to the general pool of non-exclusive questions (*i.e.* vote choice, demographic characteristics, among others). This opens an interesting venue for researchers to become one of the syndicated "clients", fielding their own questionnaires for academic purposes and perhaps collecting additional information from other "clients" to be used for academic analysis.

To summarize, this paper has attempted to focus on a particular problem with the pre and post-election survey data that is commonly used by scholars interested in voting behavior: measurement error associated with the time in which the surveys are taken. In a nutshell, measures of political behavior and attitudes taken before and after the election - even when accurate - might capture information that is different from that which would be obtained if the surveys were taken as individuals cast their votes. Measures can also be faulty due to other contextual considerations, such as desirability biases derived from knowing the identity of the winner of the election. These errors in measurement might produced estimates that are biased and inconsistent relative to the "true" parameters that could be estimated using measures taken almost immediately after the act we seek to explain has taken place.

The solution we propose is simple: rely on measures that are taken right after voters cast their ballots in order to minimize this particular type of measurement error. This is precisely what exit polls do. But in order to implement this solution, one major problem must be overcome: increase the amount of data collected from every individual without jeopardizing the quality of the collected data.

We believe that PM-MI is a reasonable alternative to collect data to analyze voting behavior given the empirical restrictions typically faced by the researcher: interviews must remain short so that every *n*-th voter can be interviewed. Ideally, we would want long interviews on voters after they leave the polls, so that all questions are asked from all voters. If this were possible, the need to rely on planned missingness and multiple imputation to "fill in" missing values would be moot. Unfortunately, carrying out exit polls in this manner would require a substantial increase in manpower so that hour-long interviews can be conducted with each voter, leaving sufficient interviewers available to approach each *n*-th voter coming out of every sampled polling station. This would enormously increase the cost of the already high cost of conducting exit polls. Hence, our proposed solution - PM-MI - seems to be a plausible alternative to collect better data, less affected by some sources of measurement error, given the restrictions in the field.

References

Abayomi, Kobi, Andrew Gelman & Marc Levy. forthcoming. "Diagnostics for Multivariate Imputations." *Applied Statistics* .

Achen, Christopher H. 1992. "Social Psychology, Demographic Variables, and Linear Regression: Breaking the Iron Triangle in Voting Research." *Political Behavior,* 14(3):195–211.

Achen, Christopher H. 2002. "Parental Socialization and Rational Party Identification." *Political Behavior,* 24(2):151–170.

Alvarez, R. Michael & Jonathan Nagler. 1995. "Economics, Issues and the Perot Candidacy: Voter Choice in the 1992 Presidential Election." *American Journal of Political Science* 39(3):714–744.

Alvarez, R. Michael & Jonathan Nagler. 1998. "When Politics and Models Collide: Estimating Models of Multiparty Elections." *American Journal of Political Science* 42(1):55–96.

Anderson, Christopher J. 2007. "The End of Economic Voting? Contingency Dilemmas and the Limits of Democratic Accountability." *Annual Review of Political Science* 10:271–296.

Anderson, Christopher J., Silvia M. Mendes & Yuliya V. Tverdova. 2004. "Endogenous Economic Voting: Evidence from the 1997 British Election." *Electoral Studies* 23(4):687–708.

Bautista, René, Marco A. Morales, Mario Callegaro & Francisco Abundis. in press. Exit polls as valuable tools to understand voting behavior: Using an advanced design in Mexico (Excerpts of unpublished manuscript). In *Elections and Exit polling*, ed. Wendy Alvey & Fritz Scheuren. New York, NY: John Wiley & Sons.

Bautista, Rene, Mario Callegaro, José A. Vera & Francisco Abundis. 2007. "Studying Nonresponse in Mexican Exit Polls." *International Journal of Public Opinion Research* 19(4):492–503.

Binder, David A. 1998. "Not Asked and Not Answered: Multiple Imputation for Multiple Surveys: Comment." *Journal of the American Statistical Association* 93(443):858–859.

Brehm, John. 1998. "Not Asked and Not Answered: Multiple Imputation for Multiple Surveys: Comment." *Journal of the American Statistical Association* 93(443):859–860.

Campbell, Angus, Phillip E. Converse, Warren E. Miller & Donald E. Stokes. 1960. *The American Voter.* Unabridged ed. New York, NY: University of Chicago Press.

Estrada, Luis & Alejandro Poiré. 2007. "Taught to protest, learning to lose." *Journal of Democracy* 18(1):73–87.

Fiorina, Morris P. 1981. *Retrospective voting in American national elections.* New Haven, CT: Yale University Press.

Frankin, Charles H. 1989. "Estimation across Data Sets: Two-Stage Auxiliary Instrumental Variables Estimation (2SAIV)." *Political Analysis* 1(1):1–23.

Gelman, Andrew, Cristian Pasarica & Ralph Dodhia. 2002. "Let's Practice What We Preach: Turning Tables into Graphs." *The American Statistician* 56(2):121–130.

Gelman, Andrew, Gary King & Chuanhai Liu. 1998. "Not Asked and Not Answered: Multiple Imputation for Multiple Surveys." *Journal of the American Statistical Association* 93(443):846–857.

Glasgow, Garrett & Roberto A. Weber. 2005. "Is there a relationship between election outcomes and perceptions of personal economic well-being? A test using post-election economic expectations." *Electoral Studies* 24(4):581–601.

Graham, John W., Scott M. Hofer & Andrea M. Piccinin. 1994. Analysis with missing data in drug prevention research. In *Advances in data analysis for prevention intervention research*, ed. L. M. Collins & L. Seitz. Washington, DC: National Institute on Drug Abuse.

Graham, John W., Scott M. Hofer & David P. MacKinnon. 1996. "Maximizing the Usefulness of Data Obtained with Planned Missing Value Patterns: An Application of Maximum Likelihood Procedures." *Multivariate Behavioral Research* 31(2):197–218.

Greene, William H. 2003. *Econometric Analysis*. Fifth ed. New York, NY: Prentice Hall.

Hausman, Jerry. 2001. "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left." *Journal of Economic Perspectives* 15(4):57–67.

Hausman, Jerry A. & David A. Wise. 1978. "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogenous Preferences." *Econometrica* 46(2):403–427.

Honacker, James & Gary King. 2006. "What to do about Missing Values in Time Series Cross-Section Data." Ms. UCLA-Harvard University.

Honaker, James, Gary King & Matthew Blackwell. 2007. "AMELIA II. A program for missing data. Version 1.1.27." Cambridge, MA: Harvard University [GKing.Harvard.edu/amelia/].

Horton, Nicholas J. & Ken P. Kleinman. 2007. "Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models." *The American Statistician* 61(1):79–90.

Imai, Kosuke, Gary King & Olivia Lau. 2006. "Zelig: Everyone's Statistical Software." Cambridge, MA: Harvard University [GKing.Harvard.edu/zelig].

Judkins, David R. 1998. "Not Asked and Not Answered: Multiple Imputation for Multiple Surveys: Comment." *Journal of the American Statistical Association* 93(443):861–864.

Judkins, David R. 1999. "Imputing for Swiss cheese patterns of missing data." Proceedings of Statistics Canada Symposium 97, New Directions in Surveys and Censuses.

King, Gary. 1998. *Unifying Political Methodology. The Likelihood Theory of Statistical Inference.* Ann Arbor, MI: University of Michigan Press.

King, Gary, James Honaker, Ann Joseph & Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An alternative algorithm for multiple imputation." *American Political Science Review* 95(1):49–69.

King, Gary, Michael Tomz & Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2):341–355.

Kramer, Gerald H. 1971. "Short-Term Fluctuations in U.S. Voting Behavior: 1896-1964." *American Political Science Review* 65(1):131–143.

Ladner, Matthew & Christopher Wlezien. 2007. "Partisan Preferences, Electoral Prospects, and Economic Expectations." *Comparative Political Studies* 40(5):571–596.

Lewis-Beck, Michael S. & Mary Stegmaier. 2000. "Economic Determinants of Electoral Outcomes." *Annual Review of Political Science* 3:183–219.

Lewis-Beck, Michael S., Richard Nadeau & Angelo Elias. 2008. "Economics, Party, and the Vote: Causality Issues and Panel Data." *American Journal of Political Science* 52(1):84–95.

Little, Roderick J.A. 1992. "Regression with Missing X's: a Review." *Journal of the American Statistical Association* 87(420):1227–1237.

Littvay, Levente & Christopher T. Dawes. 2007. "Alleviation of Context Effects in Attitude Questions Through Computer Assisted Interviewing and a Planned Missingness Data Design." Prepared for presentation at the 2nd European Survey Research Association Conference, Prague, Czech Republic June 25-29, 2007.

Mitofsky, Warren J. 2000. "At the polls: Mexican democracy turns the corner." *Public Perspective* 11(5):37–40.

Moreno, Alejandro. 2007. "The 2006 Mexican Presidential Election: The Economy, Oil Revenues, and Ideology." *PS: Political Science & Politics* 40(1):15–19.

Raghunathan, Trivellore E. & James E. Grizzle. 1995. "A Split Questionnaire Survey Design." *Journal of the American Statistical Association* 90(429):54–63.

Reiter, Jerome P., Trivellore E. Raghunathan & Satkartar K. Kinney. 2006. "The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data." *Survey Methodology* 32(2):143–149.

Rubin, Donald B. 1976. "Inference and Missing data." *Biometrika* 63(3):581–592.

Rubin, Donald B. 1977. "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys." *Journal of the American Statistical Association* 72(359):538–543.

Rubin, Donald B. 1987. *Multiple imputation for nonresponse in surveys.* New York, NY: Wiley & Sons.

Rubin, Donald B. 1996. "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association* 91(434):473–489.

Tomz, Michael, Jason Wittenberg & Gary King. 2000. "CLARIFY: Software for Interpreting and Presenting Statistical Results, (Versions 1998-2002)." Cambridge, MA: Harvard University [GKing.Harvard.edu].

Tourangeau, Roger, Lance J. Rips & Kenneth Rasinski. 2000. *The Psychology of Survey Response.* New York, NY: Cambridge University Press.

Tourangeau, Roger & Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133(5):859–883.

von Hippel, Paul. 2007. "Regression with Missing Ys: An Improved Strategy for Analyzing Multiply-Imputed Data." *Sociological Methodology* 37(1):83–117.

Weisberg, Herbert F. 2005. *The Total Survey Error Approach. A Guide to the New Science of Survey Reserach.* Chicago, IL: University of Chicago Press.

Wlezien, Christopher, Mark Franklin & Daniel Twiggs. 1997. "Economic Perceptions and Vote Choice: Disentangling the Endogeneity." *Political Behavior* 19(1):7–17.

Appendix

A Multiple Imputation

We use *Amelia II* (Honaker, King & Blackwell 2007) - which employs the bootstrapping-based Expectation-Maximization (EMB) algorithm - to make the imputations (Honacker & King 2006). Briefly, missing values are imputed linearly from the model:

$$\tilde{D}_{ij} = D_{i,-j}\tilde{\beta} + \tilde{\epsilon}_i \tag{A.1}$$

where the tildes denote random draws from appropriate posterior distributions for parameters (β) and the random term (ϵ), and the imputations are a function of the observed data ($D_{i,-j}$). Table A.1 reports the statistics on missingness in our data set.

[Table A.1 about here]

We generated $m = 10$ imputed data set on which the analysis was performed.[23] A good imputation for survey data should account for the sample design (Rubin 1996, King et al. 2001), otherwise risking producing inconsistent estimates (Reiter, Raghunathan & Kinney 2006). To account for this fact, the imputation included "cluster effect" dummy variables for each precinct in the sample. Similarly, the survey-design weights were included in the imputation to account for the variables used to select the precincts (Judkins 1998). Were these variables irrelevant to the imputation, we would be producing inefficient but not inconsistent estimates (Reiter, Raghunathan & Kinney 2006). The point estimates are computed as suggested by Rubin (1976, 1996). For simplicity, we use the notation by King *et al.* (2001) and define q as the quantity of interest, for which we calculate a point estimate as:

$$\bar{q} = \frac{1}{m}\sum_{j=1}^{m} q_j \tag{A.2}$$

and the variance of the point estimate as the sum of the *within* and the *in-between* imputations variance:

$$\begin{aligned} SE(q)^2 =& \bar{w} + b \\ =& \frac{1}{m}\sum_{j=1}^{m} SE(q_j)^2 + \left(1 + \frac{1}{m}\right)\frac{\sum_{j=1}^{m}(q_j - \bar{q})^2}{m-1} \end{aligned} \tag{A.3}$$

The quantity of interest ($\bar{q}$) is distributed t with degrees of freedom equal to:

$$d.f. = (m-1)\left[1 + \frac{1}{m+1}\frac{\bar{w}}{b}\right]^2 \tag{A.4}$$

[23] Graphic diagnostics from the imputations as suggested by Abayomi, Gelman & Levy (forthcoming) are available upon request.

Readers interested in further details on Multiple Imputation and the algorithms developed to implement it are referred to Rubin (1987), King *et al.* (2001) and Horton & Kleinman (2007). Readers interested in MI applications to political science are referred to King *et al.* (2001) and Honaker & King (2006).

B Econometric analysis

The the desirable properties of our multiply imputed data set would be partially wasted if we do not use an econometric model that more closely resembles the assumptions made by the theoretical model we are testing. So we advocate the use of en econometric model that makes the most efficient use of all available information, instead of discarding it by assuming it irrelevant. We do so based on two main concerns: not imposing an unwarranted Independence of Irrelevant Alternatives (IIA) assumption on voters, and failing to take advantage of readily available specifications that take into account individual and candidate-specific characteristics.

Common accounts of the 2006 election assume that the presence of a third candidate altered the probability of voting for either of the remaining two candidates. As the campaign was reaching its end, pollsters tried to forecast PRI's voting share knowing that it would modify the distribution of votes for PAN and PRD's presidential candidates. Thus, it was not uncommon to read that "had Madrazo been a better candidate" or "if Madrazo drops from the race" we would have observed a different outcome. If we were to ignore this feature, we would need to assume that Madrazo was indistinguishable from Calderón or López Obrador in the voter's mind (Hausman & Wise 1978). This would imply that the ratio of the probabilities *of an individual* voting for Calderón relative to López Obrador does *not* change whether Madrazo appears as a candidate or not (Alvarez & Nagler 1998). Most likely an unrealistic assumption, or at least one in need of empirical verification. Incorrectly assuming IIA may lead to inconsistent estimates and to incorrect conclusions on the 2006 election (Alvarez & Nagler 1998). An additional point is that some of the arguments advanced to explain vote choice are related to features of the candidates, such as better image, negative ads, and the like. To address these problems and explicitly accounting for both candidate and individual-specific features while relaxing the IIA assumption, we present estimates from a multinomial probit model.

The multinomial probit is motivated as a Random Utility Model (RUM) where utility is determined by a *systemic component* that reflects the average behavior of individuals given a set of observed characteristics related to individuals and choices, and by a *stochastic (random) component* that accounts for deviations from the average behavior and is assumed to be determined by unobserved differences in tastes across individuals as well as unobserved characteristics of the alternatives. Note that on Eq. B.1 below, $\beta X_{ij} + \psi_j a_i$ is the systemic component and ϵ_{ij} the random component.

To apply the model to our case, we assume that individuals seek to maximize the utility they obtain from a candidate and choose from the set of available alternatives according to this criterion. The utility (U_{ij}) that each voter derives from the alternatives is defined as:

$$U_{ij} = \beta X_{ij} + \psi_j a_i + \epsilon_{ij} \tag{B.1}$$

where a_i contains characteristics of individual i, X_{ij} contains characteristics of candidate j according to individual i, ϵ_{ij} is the random component. Note that β is a vector of candidate-specific parameters and ψ_i is a vector of individual-specific parameters to be estimated. We estimate a set of β and two sets of ψ_j. ϵ_{ij} are assumed to be distributed multivariate normal, which requires specifying the correlation between each alternative's random components:

$$\epsilon_{ij} \sim MVN(\mathbf{0}, \mathbf{\Sigma}) \tag{B.2}$$

The IIA assumption is overcome by allowing the covariance matrix $\mathbf{\Sigma}$ to have non-zero correlations terms between the ϵ_{ij} and estimating it. To identify the estimation of parameters, the coefficients for PRI are normalized to zero, thus producing coefficients for PAN and PRD *relative* to PRI. To identify and facilitate the estimation of the elements in $\mathbf{\Sigma}$, the disturbances are assumed to be homoscedastic ($\sigma^2_{PAN} = \sigma^2_{PRI} = \sigma^2_{PRD} = 1$) and the correlation between PAN and PRD's random component is assumed to be zero ($\sigma_{PAN,PRI} = 0$). This leads to the estimation of the covariance matrix:

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & & \\ 0 & 1 & \\ \sigma_{PAN,PRD} & \sigma_{PRI,PRD} & 1 \end{bmatrix} \tag{B.3}$$

Note on table B.1 that both estimated correlations - $\sigma_{PAN,PRD}$ and $\sigma_{PRI,PRD}$ - are statistically significant, suggesting that assuming IIA would be inappropriate.

[Table B.1 about here]

Given the high missingness in our data, and in order to avoid a higher estimation error derived from using imputed vote choices, we discard them from the analysis. Therefore the the presidential election analysis is performed with $n = 6,455$. All available information from these cases was used for the imputation process. Discarding imputed y's from the analysis has been shown to produce at least as good estimates as those produced when using all - observed and imputed - y's, but discarding imputed y's produces more efficient estimates with high missingness or a low m (von Hippel 2007). That is because cases with missing y's contain no information about the parameters we are trying to estimate in the models.

For further details on the multinomial probit model, readers are directed to Hausman & Wise (1978), and Greene (2003). For specific applications to multicandidate elections in political science, readers are directed to Alvarez & Nagler (1995, 1998) and examples cited therein.

C Simulating vote probabilities

Following Hausman & Wise (1978) and Alvarez & Nagler (1995), the probabilities of voting for a given candidate for three choices are given by:

$$P_{i,PAN} = \Phi\left(\frac{(\bar{U}_{i,PAN} - \bar{U}_{i,PRI})}{\sqrt{\sigma^2_{PAN} + \sigma^2_{PRI} - 2\sigma_{PAN,PRI}}}\right) \tag{C.1}$$

$$P_{i,PRD} = \Phi\left(\frac{(\bar{U}_{i,PRD} - \bar{U}_{i,PRI})}{\sqrt{\sigma^2_{PRD} + \sigma^2_{PRI} - 2\sigma_{PRD,PRI}}}\right) \tag{C.2}$$

$$P_{i,PRI} = 1 - P_{i,PAN} - P_{i,PRD} \tag{C.3}$$

where:

$$\bar{U}_{ij} = \beta X_{i,j} + \psi_j a_i, \quad j \in \{PAN, PRI, PRD\} \tag{C.4}$$

Note that $\sigma^2_{PRD} = \sigma^2_{PRI} = \sigma^2_{PRD} = 1$ since we assumed homoscedasticity, and that $\sigma_{PAN,PRI} = 0$ also by assumption to identify the parameters in $\mathbf{\Sigma}$. Similarly, we normalized $\bar{U}_{i,PRI} = 0$ to identify the estimation of the parameters.

Neither Zelig (Imai, King & Lau 2006) nor Clarify (Tomz, Wittenberg & King 2000) support multinomial probit models, so we replicated "by-hand" the procedure set forth in King *et al.* (2000) to produce simulations. Briefly, the algorithm consists of:

a) obtain estimates for β and its covariance matrix Var(β) from the m models using Eqs. A.2 and A.3. Generate $d = 8000$ draws from the distribution

$$\tilde{\beta} \sim MVN(\hat{\beta}, Var(\hat{\beta})) \tag{C.5}$$

which reflects the estimation uncertainty that derives from not having infinite observations for the estimation.

b) Determine a value for each explanatory variable (X_{ij}, a_i), compute the utility $(\bar{U}_{ij})$ as defined in Eq. C.4 using a draw from $\tilde{\beta}$ in Eq. C.5. Plug this value into Eqs. C.1 or C.2 to obtain the probability of voting for a given candidate (P_{ij}). Repeat the process d times and compute the expected value $\tilde{E}[P_{ij}] = \sum_{k=1}^{d} P_{ij}/d$.

c) *First differences* (King 1998) require computing the probabilities of voting for a given candidate as defined in b), with the particularity that it is computed twice. Once with the variable of interest set at the low value (P^L_{ij}) and once at the high value (P^H_{ij}). The first difference is simply $\mathcal{D}_i = (P^H_{ij} - P^L_{ij})$ and its point estimate $\tilde{E}[\mathcal{D}] = \sum_{k=1}^{d} \mathcal{D}_i/d$.

To define our "typical" individual we set all continuous variables at their means and categorical variables at their mode rendering a 41 year-old, primary-educated, middle class, urban resident of the southwest, who does not remember seeing any ads and is not a beneficiary of government programs, who evaluates positively the president as well as the economy in the past year and the next year, strongly identifies with PRI, thinks López Obrador generated the most negative campaign and Calderón the most positive one, and has the mean distance to all candidates as well as the mean uncertainty levels about candidates' positions.

Table A.1: Descriptive statistics on MI variables

Variable	Obs	missing	Mean	Std. Dev.	Min	Max
Female	7763	1	0.483	0.500	0	1
Presidential approval	7609	155	0.822	1.839	-3	3
Split-ticket vote	7379	385	0.479	0.500	0	1
Education	7761	3	3.032	1.287	1	5
Income	6782	982	4.028	1.801	1	7
Social class	7581	183	2.149	0.945	1	5
Vote for President	6694	1070	2.082	0.970	1	5
Vote for Deputies	6670	1094	2.113	0.974	1	5
Remembers scholarship ad	2022	5742	0.253	0.435	0	1
Remembers schools ad	2023	5741	0.215	0.411	0	1
Remembers social insurance ad	2024	5740	0.284	0.451	0	1
Remembers housing ad	2021	5743	0.253	0.435	0	1
Remembers Oportunidades ad	2022	5742	0.283	0.451	0	1
Scholarship beneficiary	1977	5787	0.212	0.409	0	1
School program benef.	1975	5789	0.119	0.324	0	1
Popular insurance benef.	1976	5788	0.198	0.399	0	1
Housing program benef.	1977	5787	0.133	0.340	0	1
Oportunidades benef	1977	5787	0.281	0.449	0	1
Negative campaign	1470	6294	4.040	2.850	1	9
Positive campaign	1471	6293	3.789	2.846	1	9
Respondent ideology	1374	6390	4.749	2.077	1	7
PAN ideology	1411	6353	4.820	2.197	1	7
PRI ideology	1410	6354	4.361	2.217	1	7
PRD ideology	1397	6367	3.147	2.183	1	7
Econ retrospective eval	1771	5993	0.120	1.346	-3	3
Econ prospective eval	1505	6259	1.054	1.269	-3	3
Party ID	1648	6116	3.927	2.374	1	8
Expected winner	1354	6410	1.969	0.909	1	5
Strategic voter	1699	6065	0.068	0.252	0	1
Political interest	1790	5974	3.075	0.822	2	4
Mexico democracy	1657	6107	0.769	0.421	0	1
FCH ideology	1409	6355	4.820	2.179	1	7
RMP ideology	1405	6359	4.397	2.115	1	7
AMLO ideology	1403	6361	3.264	2.219	1	7
Prefers balance of power	1788	5976	0.195	0.396	0	1
Voted for change	1788	5976	0.634	0.482	0	1

Table B.1: Multinomial Probit results for Presidential election

	PAN/PRI	**PRD/PRI**
Candidate distance	-0.004***	
	(0.001)	
Negative ad	-0.161***	
	(0.039)	
Positive Ad	0.369***	
	(0.050)	
Ad scholarship	0.062	0.055
	(0.066)	(0.053)
Add schools	0.010	0.014
	(0.072)	(0.050)
Ad insurance	0.032	0.012
	(0.089)	(0.047)
Ad housing	0.004	-0.050
	(0.080)	(0.047)
Ad Oportunidades	0.016	-0.005
	(0.079)	(0.042)
Scholarship	0.055	-0.035
	(0.077)	(0.048)
Schools	-0.043	-0.059
	(0.079)	(0.046)
Insurance	0.091	-0.067
	(0.131)	(0.080)
Housing	0.092	0.304***
	(0.104)	(0.066)
Oportunidades	-0.269***	-0.166***
	(0.069)	(0.046)
ID PAN	0.174***	0.023
	(0.033)	(0.022)
ID PRI	-0.239***	-0.192***
	(0.030)	(0.019)
ID PRD	0.002	0.142***
	(0.032)	(0.021)
Econ Retro Good	0.111**	0.007
	(0.047)	(0.030)
Econ Retro Bad	-0.073*	-0.006
	(0.040)	(0.024)
Econ Prosp Good	0.117	0.132**
	(0.088)	(0.059)
Econ Prosp Bad	-0.106	-0.100**

Continued on next page

	PAN/PRI	**PRD/PRI**
	(0.073)	(0.046)
Pres Approval Good	0.847***	0.052
	(0.074)	(0.048)
Pres Approval Bad	-0.730***	0.000
	(0.089)	(0.049)
Uncertainty FCH	-0.028***	-0.015***
	(0.006)	(0.004)
Uncertainty AMLO	-0.014**	0.014***
	(0.006)	(0.004)
Uncertainty RMP	0.005	0.003
	(0.008)	(0.005)
Age 18-24	-0.047	-0.062
	(0.107)	(0.072)
Age 25-40	0.048	-0.055
	(0.067)	(0.045)
Age 41-60	-0.027	-0.051
	(0.055)	(0.037)
Ed primary	0.171***	0.099
	(0.096)	(0.064)
Ed secondary	0.415***	0.301***
	(0.070)	(0.047)
Ed highschool	0.513***	0.362***
	(0.076)	(0.052)
Ed college	0.801***	0.487***
	(0.073)	(0.050)
Female	0.043	-0.066*
	(0.053)	(0.035)
Low class	-0.433***	0.201**
	(0.125)	(0.091)
Middle Class	-0.372***	0.085**
	(0.057)	(0.039)
Northwest	0.226***	-0.186***
	(0.085)	(0.057)
Northeast	0.037	-0.409***
	(0.071)	(0.051)
Southeast	-0.278***	-0.167
	(0.064)	(0.043)
Southwest	0.511***	0.492
	(0.081)	(0.050)
Urban	0.160	0.081

Continued on next page

	PAN/PRI	**PRD/PRI**
	(0.108)	(0.076)
Rural	-0.161***	-0.008
	(0.058)	(0.039)
Intercept	-0.196***	-0.205***
	(0.048)	(0.020)
$\sigma_{PAN,PRD}$	0.297**	
	(0.152)	
$\sigma_{PRD,PRI}$	0.338***	
	(0.140)	
Log-Likelihood	-5824.371	
LR-test	$\chi^2_{[79]}$=901.421***	
n	6,455	
MI sets	10	

Significance: 1% *** / 5% ** / 10%* two-tailed.

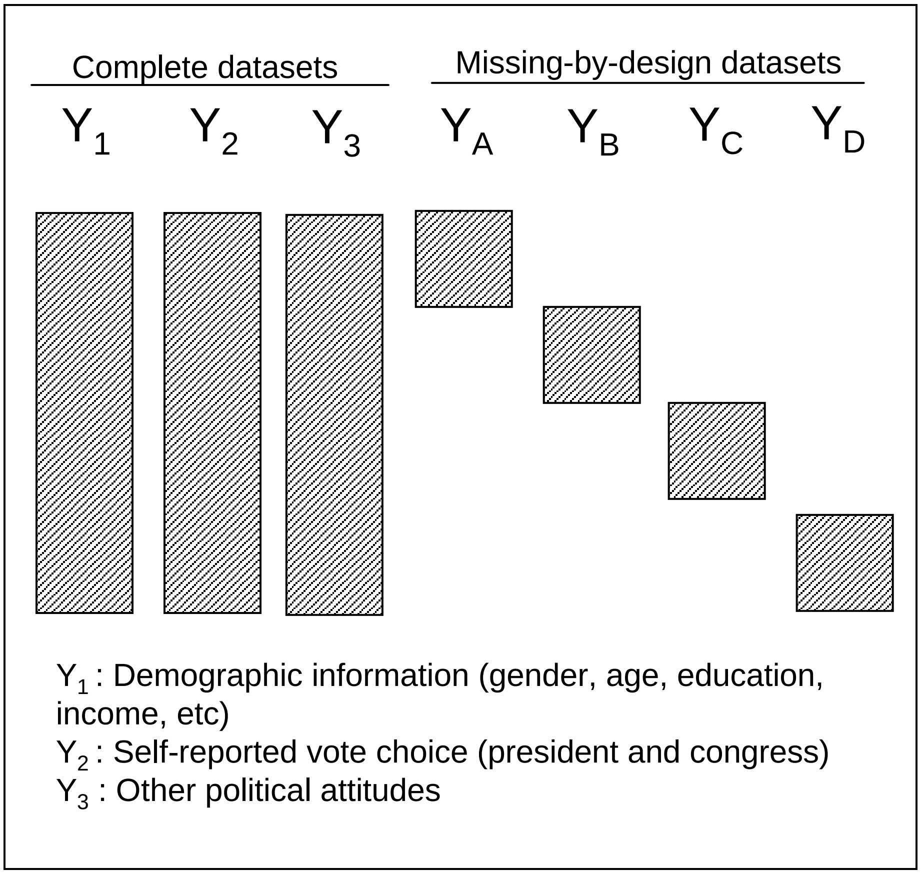
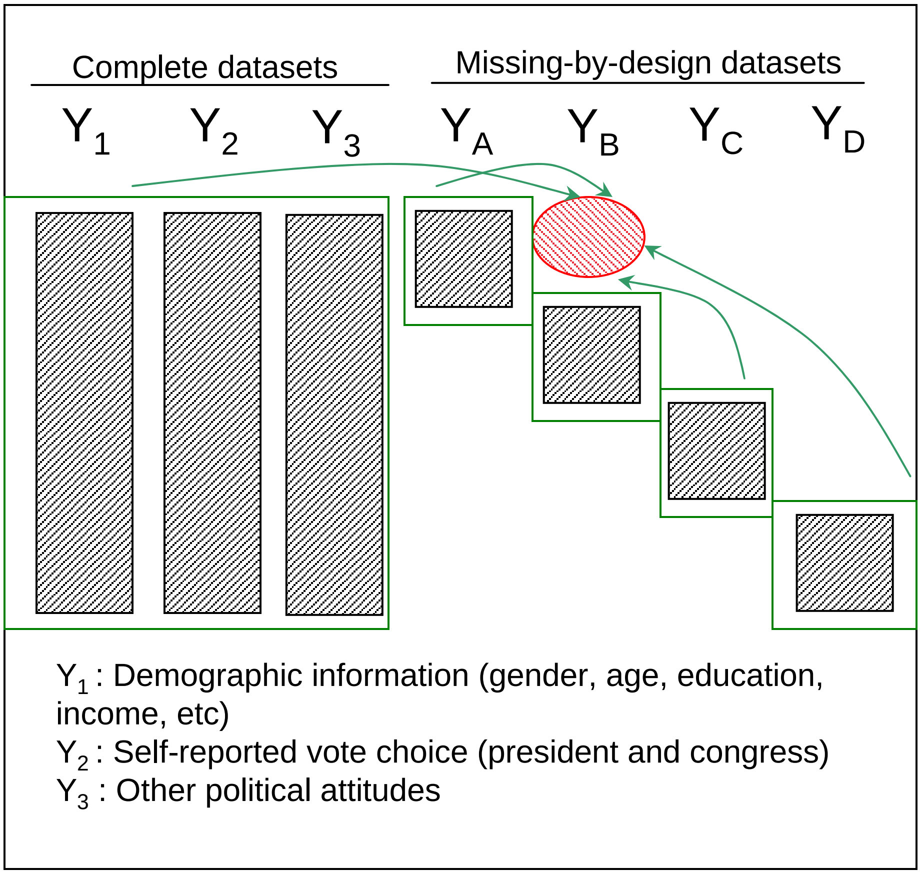
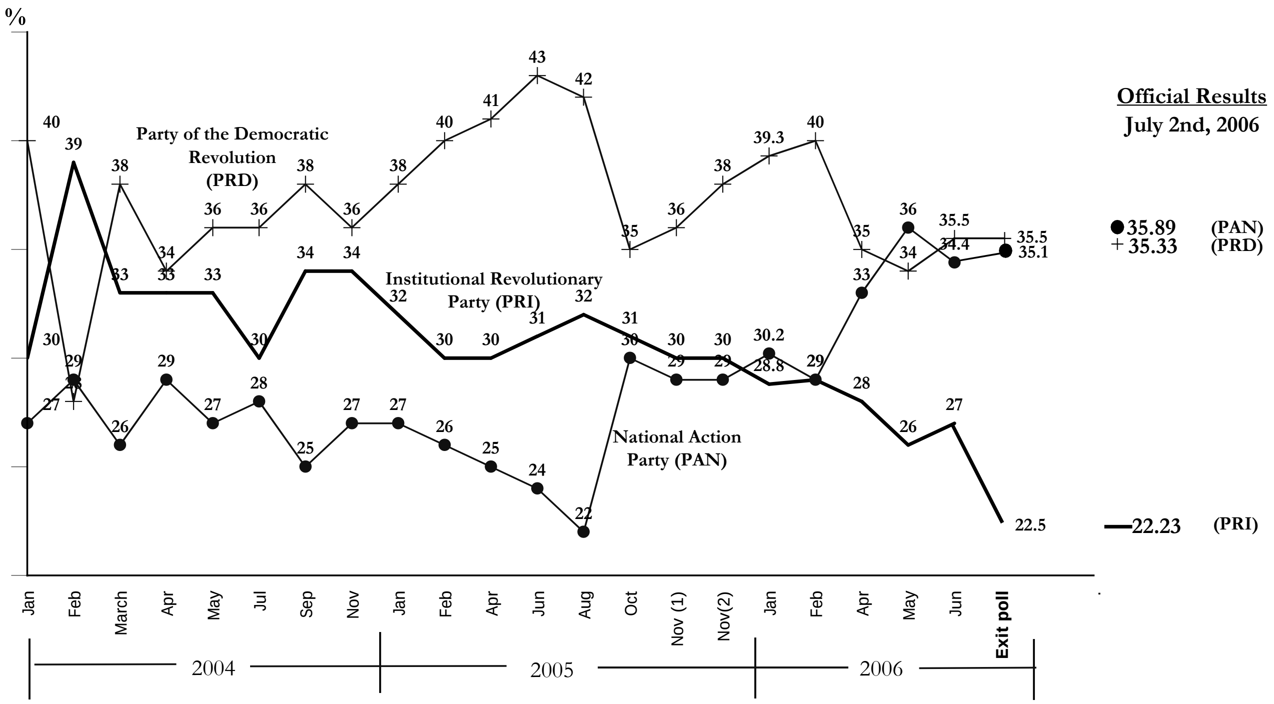

Complete datasets
Missing-by-design datasets
Y_1
Y_2
Y_3
Y_A
Y_B
Y_C
Y_D
Y_1 : Demographic information (gender, age, education, income, etc)
Y_2 : Self-reported vote choice (president and congress)
Y_3 : Other political attitudes

Figure 1: Missing data pattern generated with Planned Missingness



Y_1 : Demographic information (gender, age, education, income, etc)
Y_2 : Self-reported vote choice (president and congress)
Y_3 : Other political attitudes

Figure 2: Missing data completion using Multiple Imputation



Source: Parametría SA de CV

Figure 3: Pre-election presidential race poll trends 2004-2006

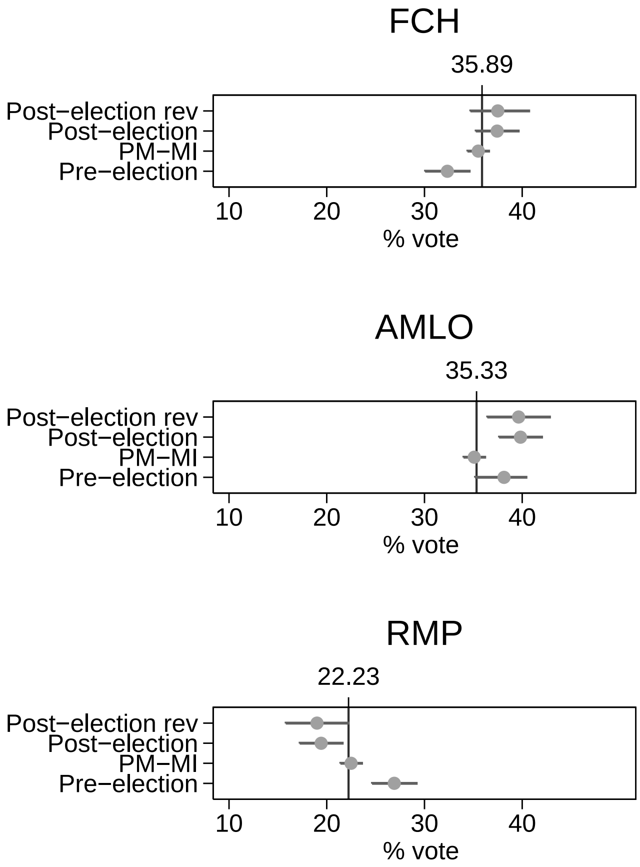


Figure 4: Point estimates - and associated 95% confidence intervals - of vote shares for each candidate generated with data from each one of the row sources. *Pre-election* estimates come from the Mexico 2006 Panel Study pre-election wave. *PM-MI* estimates come from the *Parametría*'s 2006 exit poll. *Post-election* estimates come from the Mexico 2006 Panel Study post-election wave. *Post-election rev* estimates come from the Mexico 2006 Panel Study pre-election wave, corrected for verified voters. Titles on each graph correspond to the acronym for each candidate: FCH for Felipe Calderón Hinojosa, the PAN candidate; AMLO for Andrés Manuel López Obrador, the PRD candidate; and RMP for Roberto Madrazo Pintado, the PRI candidate.

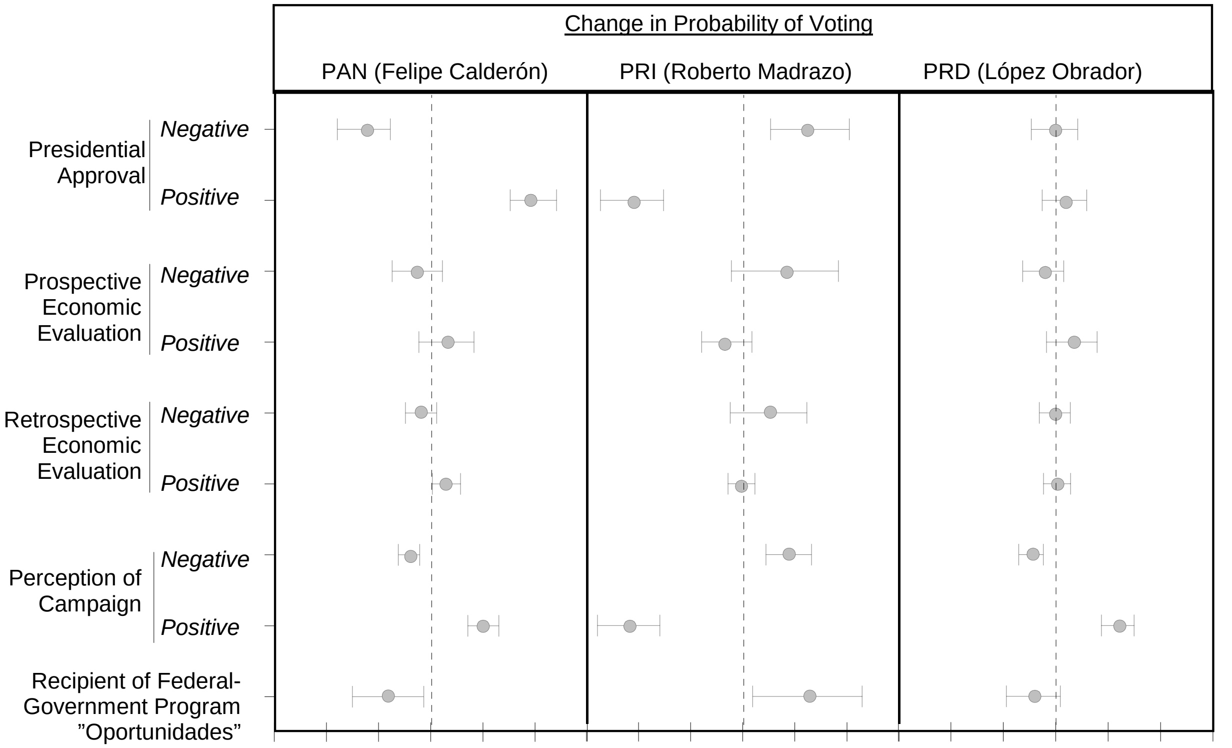


Figure 5: First differences - and their associated 95% confidence intervals - on the simulated probability of voting a party given a change in the row variable. Simulations are generated for the probability of voting for the party candidate denoted on each graph comparing a “typical” individual with the lowest value for the row variable to a “typical” individual that has the highest value on that same variable, holding all other variables constant.