Two imperfect surveys: Crowd-sourcing a diagnosis?

John M. Carey, Dartmouth College
Brendan Nyhan, Dartmouth College
Thomas Zeitzoff, American University

January 18, 2016 – v.3

Abstract
We have two surveys with similar population samples, similar questions, and largely similar results, but each with a distinct, and apparently serious, flaw. We seek feedback on what could account for the flaw in each survey, and what, if any, reliable data can be salvaged from them.

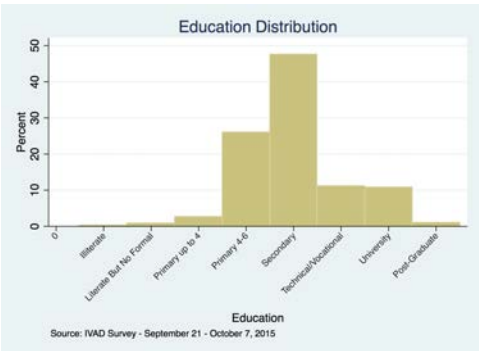
2 surveys
To measure political conspiracy theory beliefs and their covariates in Venezuela, we commissioned surveys with two different firms in the fall of 2015:

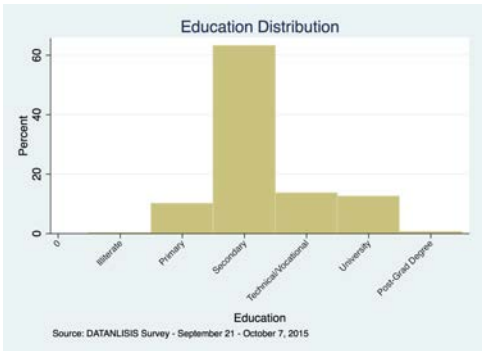
Instituto Venezolano de Analisis de Datos (IVAD)
We commissioned a 25-question survey that was in the field September 17 to October 7, 2015. The survey had 1,200 respondents, interviewed face-to-face at their homes.

Datanalisis
We attached 11 questions to an omnibus survey that was in the field October 13-23, 2015. The survey had 1,000 respondents, interviewed face-to-face at their homes.

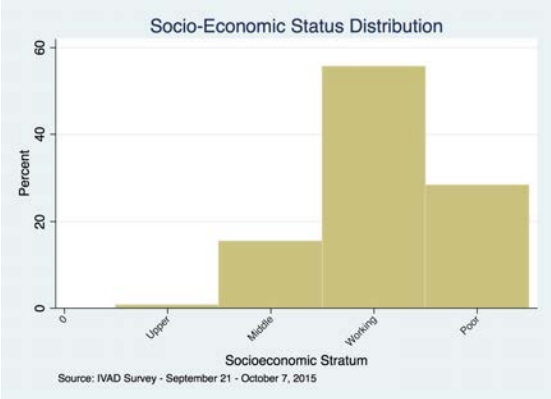
Key Point #1: The responses on identical or closely related questions are remarkably similar.
This is the case for questions that are standard on surveys, such as:

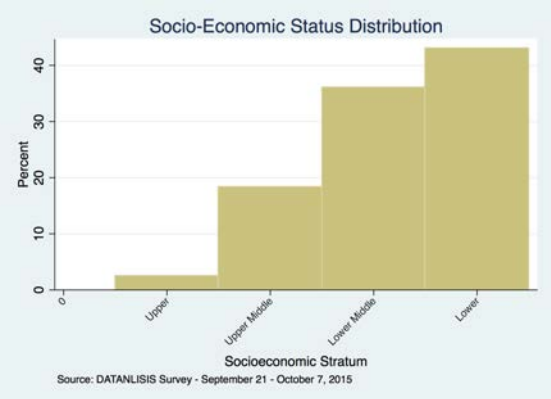
Education



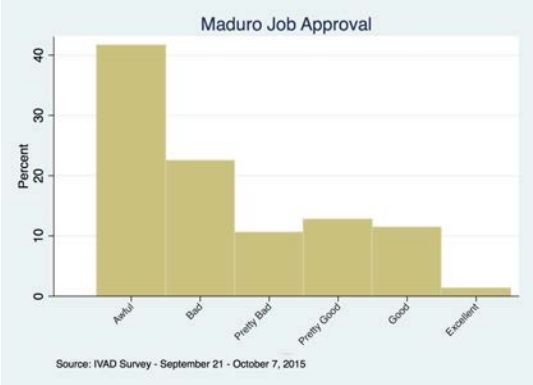


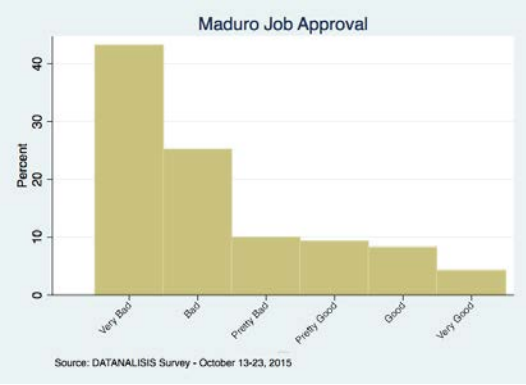
Socio-Economic Stratum



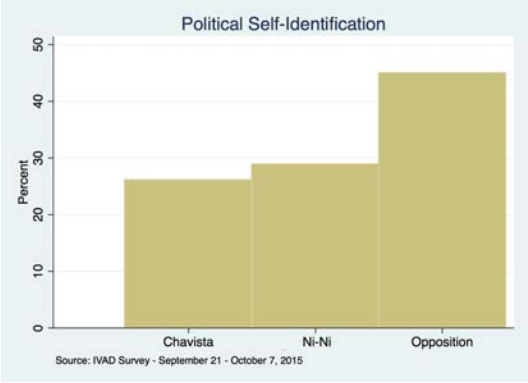


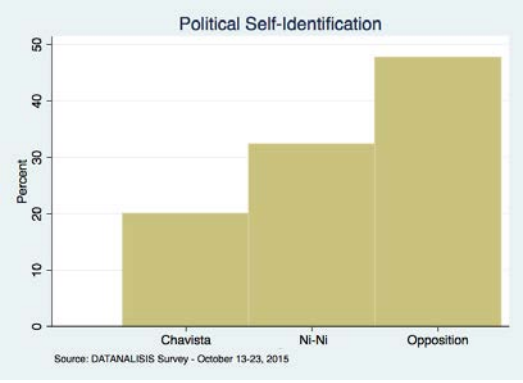
President Maduro Job Performance





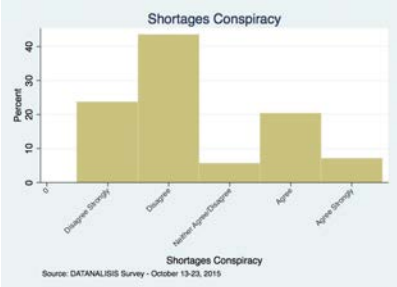
Political Self-Identification

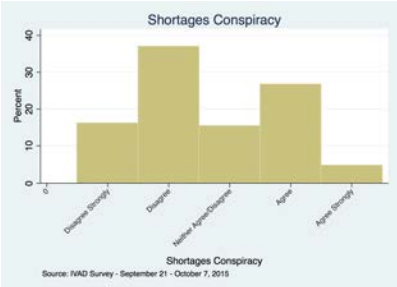




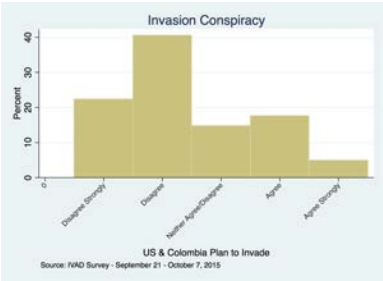
It is also the case for the questions that were original to our surveys, on belief in conspiracy theories.

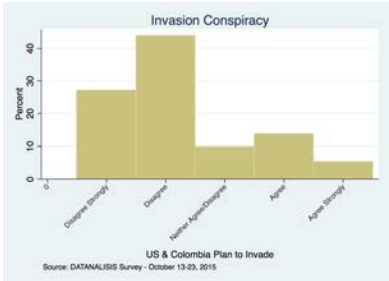
"Shortages of basic goods are caused by merchants to create chaos in the economy."



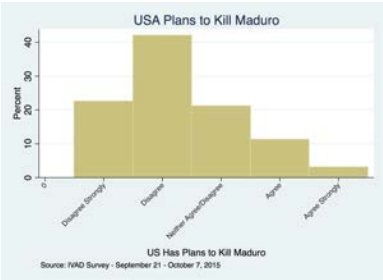


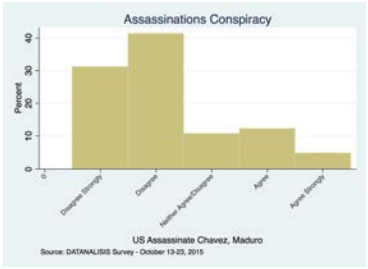
"The United States and Colombia have plans to invade Venezuela to seize its natural resources."



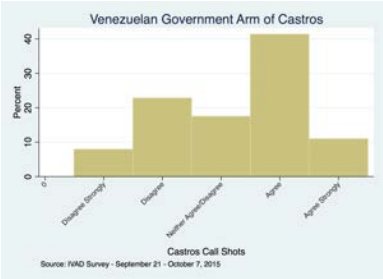


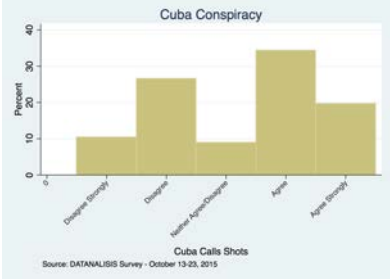
"The United States played an active role in killing Hugo Chavez, and has well developed plans to assassinate President Maduro."





"The Venezuelan government is, by now, just an arm of the Castro regime in Cuba. President Maduro does not make a major decision without first consulting with Havana."[1]



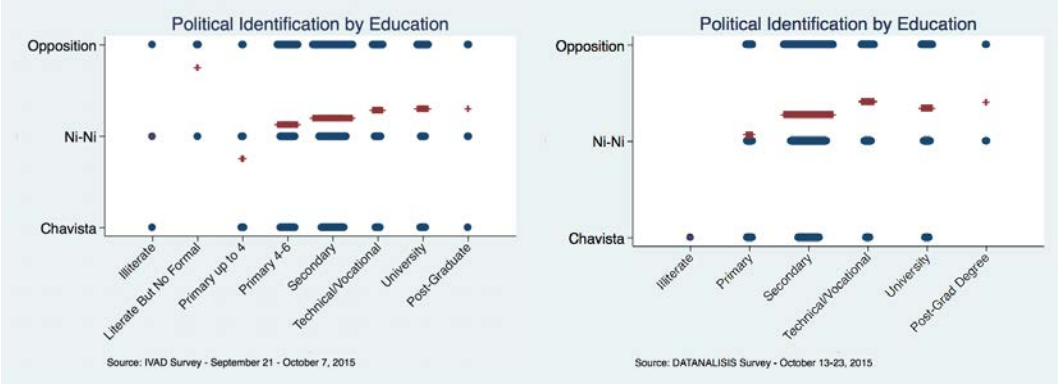


[1] Note that on the IVAD instrument broke both the Assassinations statement and the Cuba Controls Venezuela into two separate questions. We show one in each comparison with the Datanalisis survey. The distributions on the second question from each pair are the same.

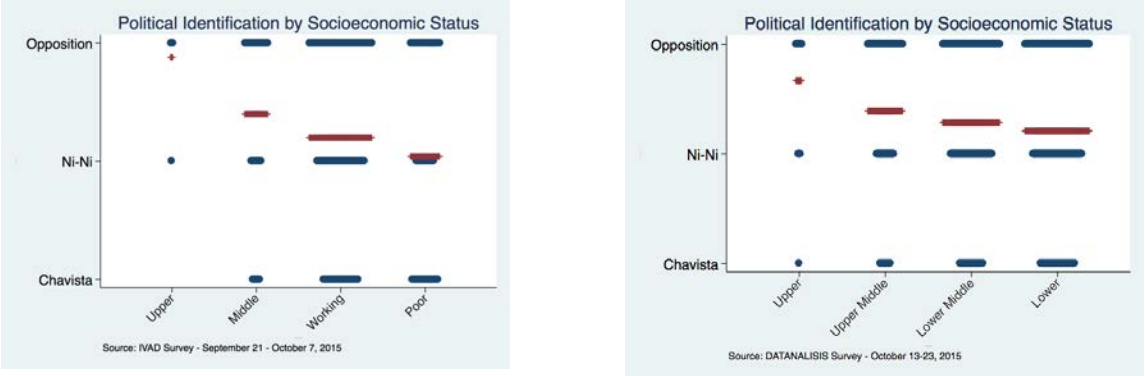
Note: For questions with identical response options across the two surveys (e.g., beliefs in conspiracy theories), we ran Wilcoxon-Mann-Whitney tests which, for most responses, reject the null hypothesis that the responses are drawn from an identical distribution of respondents. Given minor differences in the sampling methods used by IVAD and DATANALISIS, and in the timing of the two surveys, and the large numbers of responses, this is perhaps not surprising. Nevertheless, the patterns are sufficiently similar that the two surveys appear to be tapping a common strain in Venezuelan public opinion.

Key Point #2: The relationships between most key variables are also similar.

Political Self-ID by Education

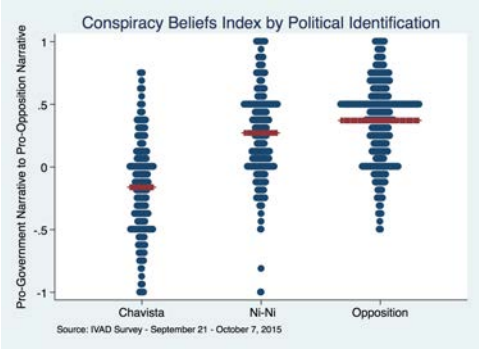


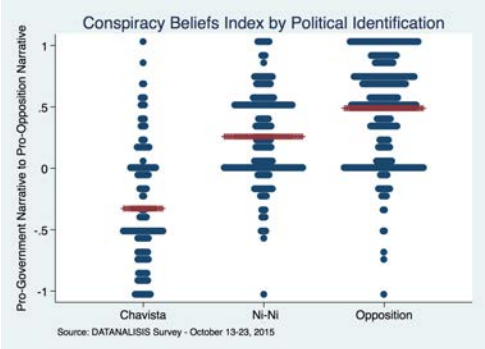
Political Self-ID by Socio-Economic Stratum



The relationships between Political Self-ID and belief in the various conspiracy theories are also strikingly similar across surveys. Because we asked about a variety of conspiracy theories, and respondent beliefs were correlated across them, we simplify by first generating an overall ***Conspiracy Index*** that runs from -1 (fully accepts all GOV-endorsed narratives & rejects all OPP-endorsed narratives) to 1 (fully accepts all OPP-endorsed narratives & rejects all GOV-endorsed narratives). We can then compare how the Conspiracy Index correlates with either:

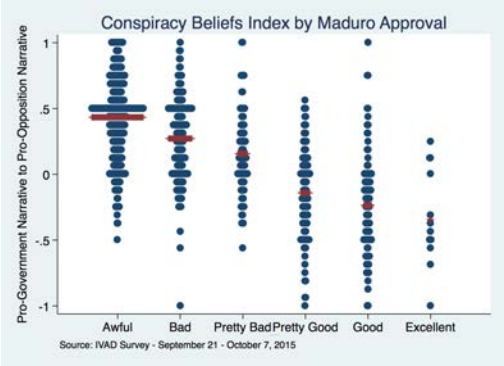
Political Self-ID

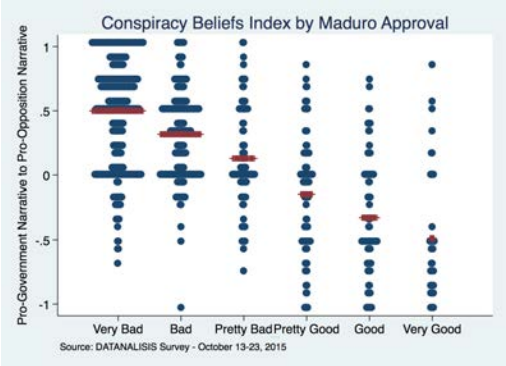




or with

Approval of President Maduro:





So far, then, the surveys appear to yield remarkably similar results.

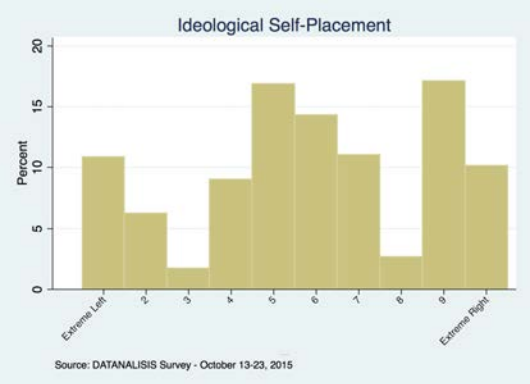
BUT …

Key Point #3: The behavior of one key variable, *Ideological Self-Placement*, is fundamentally different across the surveys.

First, it is important to note that the distributions of responses on Ideology are quite similar, despite the fact that:

- IVAD used a 7-point Left-Right scale and labeled each option with words, as indicated on the X-axis of the IVAD histogram, below; whereas
- Datanalisis used a 10-point Left-Right scale and supplied word labels only for the end-point values.

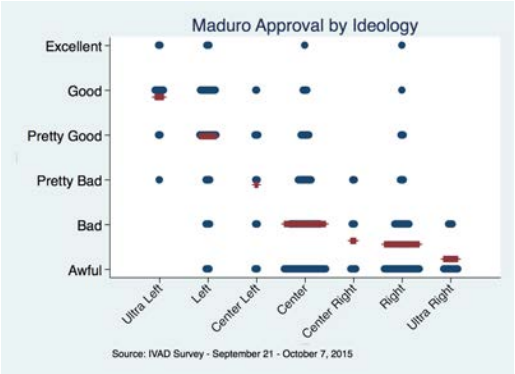
Ideology

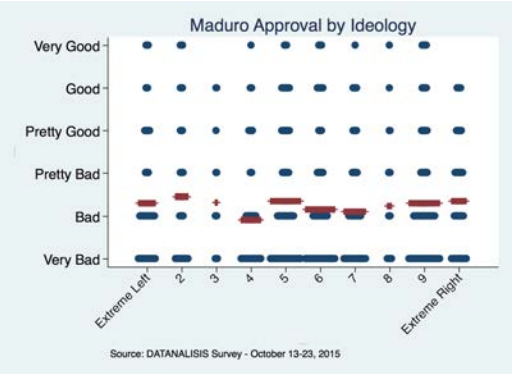


Both response distributions are clearly tri-modal, with the largest grouping clustered around the center of the scale and the cluster on the extreme right slightly larger than the one on the extreme left.

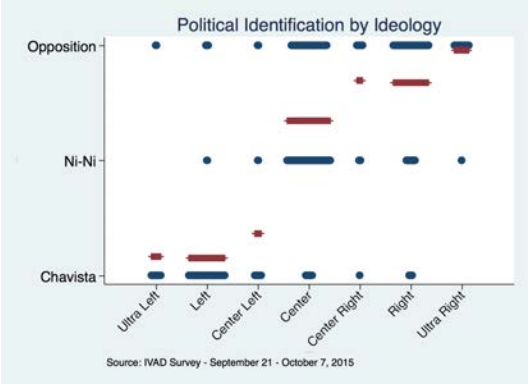
But the relationship between Ideology and the variables measuring other political beliefs and attitudes differ dramatically across the surveys.

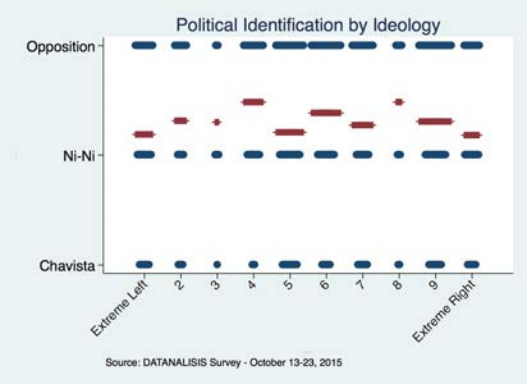
Maduro job performance by Ideology



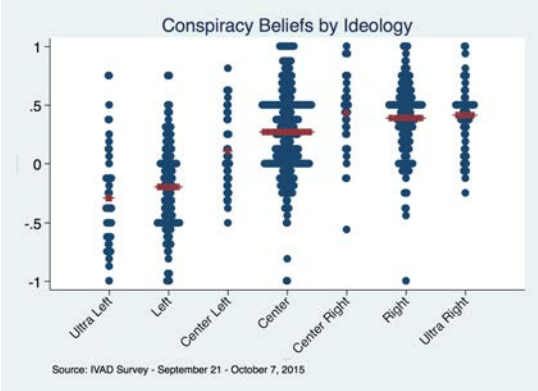


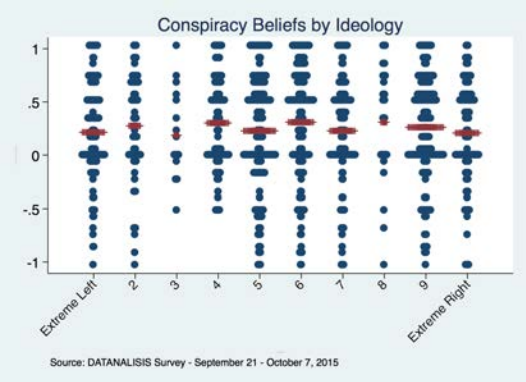
Political Self-Identification by Ideology





Conspiracy Index by Ideology





In the IVAD data, Ideology works as expected. Those who self-locate further left are more likely to identify as chavistas, more likely to approve of President Maduro's performance, more likely to believe in pro-chavista conspiracies and less likely to buy into the anti-chavista conspiracy narrative. In the Datanalisis data, by contrast, none of these relationships applies. Ideology is unrelated to the political beliefs with which one might normally expect it to covary. In multivariate regressions using the IVAD data, Ideology is consistently a powerful predictor of other political beliefs, whereas in the Datanalisis data, Ideology is never a significant predictor of other beliefs.

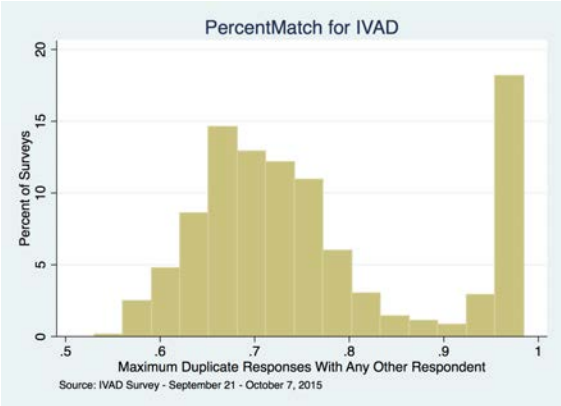
So far, then, the long list of similarities across surveys produced independently by the two companies would appear to be reassuring. Based just on this information, we might conclude that both surveys produced generally valid representations of overall Venezuelan opinion, but that some problem specific to the Datanalisis Ideology variable – perhaps a coding or data management miscue – accounts for its disconnection from expected covariates. There is, however, another key piece of information about the reliability of the data that complicates matters further.

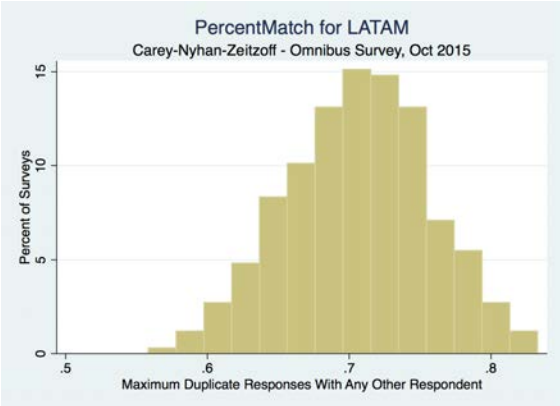
Key Point #4: The IVAD data are fundamentally dubious due to a high rate of duplicate responses. The Datanalisis data contain no duplicates and no indication of suspicious patterns of near-duplicates.

Upon receiving the data from each firm, we ran them through the PercentMatch software that Kuriakose and Robbins (2015) have developed to test for duplicate and near-duplicate observations in survey data. Kuriakose and Robbins find that a high proportion of survey datasets they examined, particularly those produced by firms outside the United States, contain high rates of duplicate, or near-duplicate response vectors – a pattern that suggests fabrication of responses, either by enumerators or by firms, to increase their number of responses without surveying more respondents.

PercentMatch calculates, for each observation, a variable indicating the maximum proportion of identical responses shared with any other observation in the data. The key scatterplots produced by PercentMatch, first for the IVAD data then for the Datanalisis data, are below. 18% of the observations in the IVAD data shared were identical (apart from the Response ID variable, which simply ran from 1 to 1,200) with at least one other observation, raising a huge red flag about the possibility of fabricated responses in the IVAD survey. There are no duplicates, or

even near-duplicates, in the Datanalisis data, which is much more in keeping with expectations for surveys of this nature (Kuriakose and Robbins 2015).





On October 15, shortly after receiving the IVAD data, we sent an email to the firm's director, Felix Seijas, informing him of the strange pattern we had detected. We requested that IVAD provide us with a dataset that included variables to identify the survey enumerator, the start time, and end time, for each survey response (as these not included in the data IVAD sent). We also asked for more information on the process by which IVAD assigns enumerators to locations, on how call-backs are conducted in the event of non-responses, and on how IVAD monitors enumerator performance.

Seijas initially responded immediately to request that we send him "all the observations that [we] consider duplicate for an exhaustive study on [IVAD's] part." We provided the list of duplicate observations. But Seijas and IVAD subsequently failed to provide any coherent explanation or to indicate any further willingness to investigate the problem.

For the purposes of exploring the IVAD data, we posited that the unique responses in the IVAD data were generated by a valid survey process, whereas the duplicates were fabricated (whether intentionally or not) from a subset of the original, valid responses. So we simply identified all duplicate response vectors and, for each pair (or triplet), we deleted one (or two) of the observations.

Summing up

- The data produced by each firm are, in most respects, consistent with each other and with other information about Venezuelan public opinion (including the December 6, 2015 election results).
- The apparent randomness of the Ideology variable in the Datanalisis survey is puzzling, but we have found no other dubious patterns in the Datanalisis data.
- The IVAD dataset is marred by the high rate of duplicate response vectors, and by IVAD's non-responsiveness in explaining or addressing this issue. That said, if we posit that the unique response vectors in the IVAD data were generated by a valid process and simply delete the duplicates, we still have 1,089 unique responses. The response patterns in those data are consistent with those from Datanalisis, with the exception that the Ideology variable from IVAD covaries as expected with other political beliefs.

We return, then, to the questions posed at the outset:

- What could account for the flaw in each survey?
- What, if any, reliable data can be salvaged from either or both?

Reference

Kuriakose, Noble and Michael Robbins. 2015. "Falsification in survey research: Detecting near-duplicate observations." Presented at the Annual Conference of the American Political Science Association.