1. Introduction

The 2008 Cooperative Campaign Analysis Project (CCAP) is a six-wave panel survey of the American electorate (registered voters only) preceding and immediately after the 2008 Presidential elections. This report describes the CCAP design, data collection procedures, weights, response rate, and proper analysis procedures.

CCAP 2008 Overview

The 2008 CCAP is a six-wave panel study of the registered voters fielded between December 2007 and November 2008. It includes "Common Content", which is the first 10 minutes of every respondent's survey. The total length of each wave is 20 minutes. After the common part of the survey respondents are routed to one of the many team studies, which make up the second half of the survey. Data collection is on the Internet using a representative sample generated from the YouGov opt-in panel. The sampling frame is U.S. registered voters over the age of 18.

CCAP 2008 at a Glance

Title:	2008 Cooperative Campaign Analysis Project
Main purpose:	Pooling resources to conduct a large representative panel survey prior to the 2008 presidential elections.
Population:	Adult registered voters in the U.S.
Sample:	Target sample generated from the American Community Study (ACS), and then matched with YouGov/Polimetrix participant pool. Participants from Battle-ground states are over-represented so that each strata (Battle-ground and non-Battle-ground) makes 50% of the sample.
Design:	Panel for common content, cross-sections with unique content for each wave
Mode:	Internet
Number of variables:	1,705
Field Period:	The baseline wave of CCAP was fielded on 17 December 2007. Subsequent waves were fielded in 2008 on 24 January, 21 March, 17 September, 22 October, and 5 November (post-election). Each wave was in the field for approximately 2 weeks.
Response rate:	The within-panel response rate (off of completed baseline interviews) for the matched sample is 82%, 92%, 87%, 83%, and 95% in each respective wave. The final matched sample contains 15,388 completed interviews in the baseline wave, 8,062 of whom completed each of the six waves of the project.
Number of cases on the file:	20,000 total
Interview length:	20 minutes – 10 minutes common content and 10 minutes for the separate team studies.
Weight:	All analyses that generalize to the population should employ weights. *bgweight* should be used to compare between Battle-ground and non Battle-ground states. *natweight* should be used for inferences about the general population.
Sponsor and design:	The project was a joint venture of 27 research teams around the world.

Data collection firm: YouGov

2. Survey questions

Variable labels indicate survey wave. Each contains a series of modules.

Waves: Baseline [CAP], January [JCAP], March [MCAP], September [SCAP], October [OCAP], November [OCAP].

Common Modules: Attentiveness, issues, candidate assessment, vote intentions, presidential nominee preferences, media consumption, political information, religion, demographics, panelist profile, administrative questions.

3. Data Collection (see also Jackman and Vavreck, 2010)

Step 1: Defining the Target Sample
YouGov/Polimetrix constructed a sampling frame for CCAP from the 2005–2007 American Community Study (ACS), including data on age, race, gender, education, marital status, number of children under 18, family income, employment status, citizenship, state, and metropolitan area. The frame was constructed by stratified sampling from the full 2005–2007 ACS sample with selection within strata by weighted sampling with replacements (using the person weights on the public use file). Data on reported 2008 voter registration and turnout from the November 2008 Current Population Survey Supplement was matched to this frame using a weighted Euclidean distance metric. Data on religion, church attendance, born again or evangelical status, news interest, party identification and ideology was matched from the 2007 Pew Religious Life Survey. The target sample was selected by stratifying on age, race, gender, education, and state (with battleground states double sampled) using simple random sampling within strata, excluding all non-registered persons.

Step 2: Matching to the Target to Generate the "Pool"
With the target defined, respondents were chosen from the YouGov/Polimetrix Polling-Point Panel and the MyPoints panel using a five-way cross-classification (gender $\times$ race (3 categories) $\times$ battleground state). At each wave, additional cases were added to deficient cells to achieve approximately 30,000 interviews. All respondents who had completed any prior wave were re-invited to subsequent waves. The final set of completed interviews (numbering approximately 48,000, after quality controls were applied) was then matched to the target frame, using a weighted Euclidean distances metric, scaled by standard deviation of the target variable (the Mahalanobis distance) with penalty matrices for categorical variables. This set of respondents is called the "pool" of completed interviews from which the final matched sample will be drawn.

   The variables in the distance function are the percentage waves completed out of possible completed waves, state, region, metropolitan statistical area, marital status, born again/evangelical status, income, employment, age, race (white, black, Hispanic, other), years of education, interest in news, gender, 5-point party identification, 3-point ideology, the interaction of news interest and ideology, turnout, and "don't know" response on ideology. For unordered variables, matrices of distances were used, as indicated above.

Step 3: Constructing the Matched Sample from the Pool

With 48,000 people in the pool, there are, on average, between two and three possible matches from the pool for each of the 20,000 respondents in the target sample. For example, if a 40 year-old Republican woman with a college degree is drawn for the target sample (off the ACS), Polimetrix uses nearest neighbor matching (using the variables above) to find the closest match to this woman in the pool of completed interviews. This reduces the pool from 48,000 to 20,000. The resulting sample is called the "matched sample".

Even though care has been taken to hit the targets before the final sample is constructed, the sample may still miss on some combinations of characteristics. In other words, no match is guaranteed to be perfect. Because of this, the final step in sample construction is to generate a set of post-stratification weights.

Step 4: Weighting the Matched Sample
The matched cases were weighted to the sampling frame using propensity scores. The matched cases and the frame were combined and a logistic regression was estimated for inclusion in the frame. The propensity score function included age, years of education, gender, black and Hispanic race indicators, news interest, turnout, saying "don't know" on ideology, party identification, and interactions of age and gender, and turnout and gender. Weights were constructed by quantiling the propensity scores into 10 cells. The final weights were then post-stratified by battleground status, gender, and race. Weights were not trimmed. The largest weight is 10.26. The final weights were normalized such that their sum equals the sample size.

4. **Waves and Response Rates**

Table 1: Interview Disposition by Wave

Interview Disposition	Baseline	January	March	September	October	Post-election
Completed	15,388	13,595	11,527	10,221	8,458	8,062
Started – not complete	195	823	559	447	346	323
Never started	4417	5759	1,648	1,313	1,084	1,031

The baseline wave of CCAP was fielded on 17 December 2007. Polimetrix has a steady stream of panelists taking surveys every day, and as people hit the survey servers, they are sent to the survey that needs their "match" the most. The CCAP baseline wave was completed by 43,739 panelists. These people make up the pool of respondents from which the final matched sample will be drawn. Subsequent waves were fielded in 2008 on 24 January, 21 March, 17 September, 22 October, and 5 November (post-election). Each wave was in the field for approximately 2 weeks. The within-panel response rate off the baseline pool for each wave is roughly 52% of the initial set of completed baseline interviews. The final matched sample contains 15,388 completed interviews in the baseline wave. The within-panel response rate (off of completed baseline interviews) for the matched sample is 82%, 92%, 87%, 83%, and 95% in each respective wave. The current release of CCAP data contains a total of 20,000 respondents, 15,388 completed the baseline survey, 195 started but did not complete, and 4,417 never started. 8,062 participants completed each of the six waves of the project.

Fresh cross was added in every wave except September and the Post-election survey. The re-interview rate of fresh cross tended to be lower than in the initial invitation group. Of ultimate interest, however, is the final matched and weighted sample. Not all of the completed interviews (the pool) are used in the final matched sample – the point of sample matching is to choose the closest match for each target given a set of possible matches.

Table 2: Panel retention by wave started
(Numbers indicate complete answers in each subsequent wave)

Wave started	Baseline	January	March	September	October	Post-election
Baseline (Dec)	15,388	12,595	11,527	10,221	8,458	8,062
January		823	559	447	346	323
March			1,648	1,313	1,084	1,031
October					1,946	1,604

5. **Weights and Variance Estimation (see also Vavreck and Rivers, 2008)**

The data are designed to be analyzed with weights. Weights need to be used to generalize to the general population.

In Web survey panels, all methods of recruitment (including those that start with some form of probability sampling) will inevitably involve some degree of self- selection. Without adjustment, survey estimates based upon such samples will be biased. Conventional methods of adjustment, such as quota sampling or post-stratification based upon a few demographic variables, are inadequate to address these biases.

The 2008 CCAP uses sample matching to construct a representative sample of the general population. The method of sample matching simultaneously reduces bias and improves efficiency. The availability of large amounts of auxiliary information from consumer and voter databases make it feasible to select a sample that is approximately balanced on a large set of variables. Sample matching is a cost-effective method for constructing samples with minimal bias. With sample matching, a population frame that includes large amounts of auxiliary information is used to select a target sample using known probabilities of selection. For each element of the target sample, the closest matching element from the panel is selected for interviewing. Because of imperfect matching, the resulting sample still needs to be weighted, but the weights are much smaller than would be needed for either a random sub-sample or a quota sample.

The 2008 CCAP sample is balanced between subjects from "Battleground states" (competitive states in the 2008 Presidential election: CO, FL, IN, MO, NV, NH, NM, NC, OH, PA, VA) and non-Battleground states. Comparisons between battleground and non-battleground states should use the weight *bgweight*. Analyses aiming at inferences about the general population should use the weight *natweight*.

6. Missing Data

Missing data are assigned codes or to indicate the reason that the data are missing. The codes in use are as follows:

8: SKIPPED (for categorical variables with less than 10 categories)
9: NOT ASKED (for categorical variables with less than 10 categories)
- 8: SKIPPED (for categorical variables with more than 10 categories)
- 9: NOT ASKED (for categorical variables with more than 10 categories)
998: SKIPPED (for continuous variables)
999: NOT ASKED (for continuous variables)
. SYSTEM MISSING

"SKIPPED": The respondent did not respondent to the question (item non-response).
"NOT ASKED": Respondent completed survey but was not administered the specific question.
"SYSTEM MISSING": Respondent was not administered the question because they did not start or complete the survey.

References

Jackman, Simon, and Lynn Vavreck. "Primary politics: Race, gender, and age in the 2008 democratic primary." *Journal of Elections, Public Opinion and Parties* 20, no. 2 (2010): 153-186.

Vavreck, Lynn, and Douglas Rivers. "The 2006 cooperative congressional election study." *Journal of Elections, Public Opinion and Parties* 18, no. 4 (2008): 355-366.