Crowdsourcing Panel Studies and Real-Time Experiments in MTurk

Dino P. Christenson and David M. Glick
Boston University
DinoPC@BU.edu and DMGlick@BU.edu

While researchers conducting quick experiments and pilot studies currently appear to make the most use of Amazon's Mechanical Turk (MTurk) as a subject recruitment tool, it is capable of supporting more ambitious research designs, many of which would be otherwise infeasible for most researchers. Specifically, researchers with a modest budget can use MTurk to recruit participants for customized and flexible panel studies. Because it is a cloud based environment, it is easy to recontact respondents, which helps limit panel attrition. Moreover, when used in tandem with online survey software, panel waves can be quickly fielded around imminent and recent political events, rather than at constant intervals or other times determined well in advance.

Thus, MTurk's attributes allow researchers to affordably collect individual level data for pre-post comparisons that can be combined with real-time experimental treatments. In this piece we briefly discuss our own experience conducting panel studies in MTurk and provide some basic instructions for researchers looking to do the same. We utilize the design and data from one of our own recent studies to discuss how we took advantage of MTurk and suggest some avenues for future research.

Implementing a Panel Study in MTurk

Last summer we implemented a five wave panel study to capture public opinion about health care, the Supreme Court, and politics in general around the Supreme Court's Affordable Care Act (ACA) decision (Christenson and Glick, 2012). We collected two waves of data prior to the ACA decision, two waves shortly after, and then reconnected with our participants for a final wave around the November elections. In all, we were able to independently conduct five extensive political surveys for approximately $5,000 total. Employing an MTurk sample was not just inexpensive but, as we show below, also offered unparalleled flexibility for collecting panel data and implementing a real-time experimental treatment around a real world event.

We used MTurk for sample recruitment and re-contact while conducting the actual surveys using an online platform.[1] We have used both SurveyGizmo and Qualtrics in our work and found MTurk to be equally compatible with them. The first step in recruiting a panel is very similar to conducting a one-time survey in MTurk. You simply have to post a public "HIT" (MTurk jargon for an assignment or "Human Intelligence Task") to attract participants and direct them to the survey. For this HIT we offered participants $1 for a 15 minute survey in which they would "answer questions about politics and healthcare." Following Berinsky, Huber and Lenz (2012*a*), we restricted our survey to those in the United States with a 95% approval rating from other MTurk "requesters" (those who post assignments like us) on at least 50 previous assignments. Our initial HIT also told participants that "attentive and satisfactory completion of the initial survey will earn invitations to follow up surveys at higher pay rates." We also advised respondents that "we check responses to ensure that [they] read the questions and respond carefully." Such recruitment restrictions and clear wording regarding the length of the study lead to more reliable samples across the waves and less respondent drop-off on each survey.

The initial HIT must also include a link to take participants recruited in MTurk to online survey hosted elsewhere (e.g., Qualtrics or SurveyGizmo), the generic password for entering it, if password protected, and a box for participants to enter the unique code which serves as a receipt they will get at the end of the survey. We recommend explicitly telling participants both in the HIT and at the end of the survey to return to the HIT and enter the random number displayed at the end of the survey to claim their compensation and approval. This last part is critical because it is necessary for a) determining who satisfactorily completed the assignment for payment, and b) linking a participant's surveys to each other for the panel. This step is necessary in part because Amazon is the intermediary for all connections between researchers and participants, which provides anonymity to the participants. Researchers ("requesters") fund an Amazon account with a credit card and then Amazon deposits money in participants' accounts after assignments are approved. All the researcher knows about participants is their random ID number. Both SurveyGizmo and Qualtrics will produce a unique ID for the respondent on the last page of the survey, which also appears as a variable in the survey output. Respondents can then copy and paste or type this receipt into MTurk. In the ACA study, we recruited participants straight into our first wave. Alternatively, one can conduct a screener survey prior to the full survey to ensure that those taking the survey meet particular characteristics, which would be appropriate for oversampling from particular populations (e.g., based on demographic questions) or limiting the sample to particular populations (e.g., based on geographic regions) or launching side-by-side panels around different events without repeat participants. Such screener surveys generally cost only a few cents per respondent.

The authors' names are listed alphabetically.

[1] We were able to conduct our first panel studies with few complications in part by utilizing the work of Berinsky, Huber and Lenz (2012*a*) and their supplemental documentation online (Berinsky, Huber and Lenz, 2012*b*,*c*).

At the end of the first wave you will have two spreadsheets: one downloaded from the survey software with all of the responses, and one downloaded from MTurk. To get the latter, log in as "requester," then select "manage," then click on the "batch" you want, and then choose "download csv." The next step is deciding whom to approve and whom to keep in the panel. This requires matching rows in the survey data to rows in the MTurk participant file using the unique ID that participants entered into the MTurk HIT and which appears in the "answer" column at the far right of the MTurk export file. Matching with the unique identifier can be done in any statistical program or spreadsheet. We chose to compensate and approve anyone with a successful match and those who appeared to make a good faith effort to enter a proper unique ID (even those that may have made a typo or otherwise gotten confused). While you can compensate those without a perfect match, you can only keep those for whom you can match an MTurk entry to a survey entry in the panel.

In addition, you may want to include a few trick screener questions in the first survey to identify those who are not paying any attention and simply checking boxes (Berinsky, Margolis and Sances, 2012). Because some *Turkers* may be multitasking, expecting little oversight or exhausted from previous MTurk tasks, you are likely to have some survey satisficing.[2] For example, in the ACA study, embedded in a series of multiple choice political knowledge questions, such as "what government position does John Roberts hold," we asked: "what government position (Senator, Chief Justice, Speaker of the House, Secretary of State, None of the Above) do you currently hold?" We deemed participants who failed all three screeners unsatisfactory and dropped them from the data and from the recruitment list for subsequent waves, since they were likely not paying attention at various points in the survey.

At the end of this process you will be able to create a simple master panel file which comprises MTurk ID numbers in the first column and their corresponding wave one survey IDs in the second column. As the panel progresses you will simply tack on additional columns of survey IDs after subsequent waves by matching to the MTurk IDs in the corresponding MTurk output files. The final step in the first wave, though it is not absolutely necessary, is giving all of those we wanted to invite back for the second wave a custom MTurk "qualification," which we called "complete 1." You can create and bulk-assign a custom qualification by uploading a spreadsheet of MTurk IDs within the "manage" function. MTurk will allow you to make this qualification a requirement to view the second wave HIT which will prevent other *Turkers* from accidentally trying to join the panel.

The only difference between the first wave and subsequent waves is that instead of posting a public HIT to recruit, you will need to have Amazon send emails to your chosen participants inviting them to the next wave. You can do this automatically with a script. We followed Berinsky, Huber and Lenz (2012*c*) and used a simple Perl code to feed it a text file of MTurk IDs (all of the "live" participants from the first column of our master panel spreadsheet) and the contents of an email which will be sent from the address associated with your Amazon account to each participant.[3] You will want to include a subject line such as "invitation to follow up survey for $1.50" and include a link to the proper survey and the password to enter it in the text (and, probably, a friendly and gracious note). We placed our own MTurk ID (which you will get when signing up) at the top and bottom of text file so that we would get the first and last emails to verify that the Perl script was working correctly.

For each wave after the first, you will need to create a new HIT so that participants can enter the new code they will get at the end of the survey. This HIT is no longer for recruiting but is still needed for collecting survey IDs to compensate participants and for continuing the panel. On that note, we found that participants occasionally had trouble accessing the HIT, and thus we ended up creating a main HIT as well as one or two backups (with smaller Ns) for those who got locked out of the primary HIT. Doing so simply means that you will have to download and append multiple CSVs from MTurk for a given wave. It is likely that a few people will recognize that they made a mistake entering their unique ID and email you. As long as they send a MTurk ID and a survey code, you can update your master sheet manually. We recommend double-checking all non-matches manually, as some will not match due to formatting issues, like leading or lagging spaces (we encountered approximately five per wave like this).

You can make the HIT private by including a custom qualification (above) as a requirement for accepting the HIT. Importantly, you will want to make sure you can view the HIT from your own account, so that you can copy the unique URL and include it on the last page of the survey and/or in the email, which points participants to the URL where they can enter their codes. In the design phase you will have a choice of whether to allow those who do not meet the requirements (including yourself) to still be able to view a HIT. The procedure for subsequent waves is the same, but with a new custom qualification, new private HIT, and new list of MTurk IDs for the Perl script to send the emails. By the end of the process you should have a master participant

[2] We recommend reading little, if anything, into the "time to completion" data from MTurk. This number simply represents how long it took for respondents to enter the survey receipt after opening the HIT. It is an accurate indicator of time only for those who open the HIT, go straight to the survey, and then go straight back to MTurk. If one refreshes or reopens the MTurk HIT for example, their time to completion could be a few seconds. This becomes even more relevant later when participants access the survey via an email and only go to the HIT at the end.

[3] You can also update your ID file partway through a wave to send a reminder email.

file matching MTurk IDs with each survey they completed and output files for each wave from the survey program, i.e., panel data.

Panel Demographics and Attrition

The primary concern in conducting any study with a sample of *Turkers* is representativeness. As with any online environment, one expects a younger, more liberal, and more technology savvy participant pool than the population at large. These demographic traits and other features of MTurk samples are thoroughly described and analyzed in Berinsky, Huber and Lenz (2012*a*), and our initial ACA study sample largely bolsters their cross-sectional findings. Before turning to the evolution of our sample over the panel waves, we briefly describe our first wave sample demographics with an eye towards the potential for unrepresentativeness.

Table 1: Sample Demographics and Comparison with Other Surveys

		Internet		**Face to Face**	
	Our	**BHL**	**ANES-P**	**CPS**	**ANES**
Variable	**Sample**	**MTurk**	**2008-09**	**2008**	**2008**
% Female	54.4	60.1	57.6	51.7	55.0
% White	79.0	83.5	83.0	81.2	79.1
% Black	7.9	4.4	8.9	11.8	12.0
% Hispanic	5.0	6.7	5.0	13.7	9.1
Age (years)	33.4	32.3	49.7	46.0	46.6
Party ID (mean 7 pt.)	3.2	3.5	3.9		3.7
Ideology (mean 7 pt.)	3.3	3.4	4.3		4.2
Education	50% Coll Grad 37% Some Coll	14.9 yrs	16.2 yrs	13.2 yrs	13.5 yrs
Income (median)	30-49K	45K	67.5K	55K	55K

Traits for our sample from wave 1 (N=1242), "BHL MTurk" = Berinsky, Huber and Lenz (2012a),, ANES-P = American National Election Panel Study (Knowledge Networks), CPS = Current Population Survey, ANES = American National Election Study), CPS and ANES are weighted. Data from all columns other than "Our Sample" reproduced from Table 3 in Berinsky, Huber and Lenz (2012a)

Table 1 compares our (wave one) MTurk sample to that of another MTurk sample, a high quality internet panel study, the 2008-2009 American National Election Panel Study (ANES-P) conducted by Knowledge Networks, and two gold standard traditional surveys, the 2008 Current Population Survey (CPS) and the 2008 ANES. Overall, our sample appears to closely resemble the population with a few expected deviations, especially age and education. While not as representative as the field's best national probability samples, it outperforms the typical convenience samples. Indeed, other than being a slightly younger sample, it is comparable to the high quality ANES-P population.

Panel studies, however, introduce additional opportunities to reduce the sample representativeness with every wave. Panel attrition may be the result of participants choosing to opt out of future surveys or difficulties faced by the researcher in re-contacting them. In either case, if those in particular demographic groups or with certain attitudes leave the panel, the reduction in sample size can affect representativeness and pose a problem for inference. Because it utilizes cloud computing, MTurk makes panel attrition less likely. For one, regardless of whether panelists move—a typical problem for long-term panel studies—they are still easy to contact through their MTurk accounts. In addition, because we can screen for Turkers with a history of carefully completed (but unrelated) projects, we can sample those who are more likely to complete new tasks.

The numbers of successfully completed responses for the four waves conducted in the weeks surrounding the ACA decisions were 1242, 944 (76% of the successful completes from the previous wave), 856 (91%), and 751 (88%). We consider these numbers to be especially high given the tight time windows we were working with. They are in line with rates in other "carefully conducted" high quality surveys (Bartels, 1999). Because we wanted to make sure all of our responses in a respective wave had the same opportunities to be exposed to relevant information, our surveys were only in the field for a few days. In fact, our second wave, the one with the largest attrition, was only in the field for 48 hours, because we wanted to stop it prior to Supreme Court decision, the date of which was not known until moments before the announcement. Four months after our initial four waves, we were able to collect responses from 472 participants (63% response rate).

While the number of responses and the speed with which we were able to collect them are impressive in their own rights, the seemingly random attrition may be even more important for statistical inference in a panel study. Figure 1 shows the evolution of our panel's demographics. Particularly in the first four waves, collected in a one-month period around the Supreme Court decision (highlighted as the event in the figure), our panel demographics remained stable as people fell out of the panel. While there were some deviations in the fifth wave, collected four months later, our data suggest that panel attrition was essentially more or less equal across the categories of respondent traits, such as race, gender, partisanship, and income. Interestingly, slight trends are evident in age and education, with 18 and 19 year olds and those with some college education consistently falling out slightly more often than the older and more educated, respectively. It is somewhat intuitive to expect that across five months, the older and more educated would be more reliable survey respondents. In all, however, we believe that the results of our study suggest that there is little danger in the panel attrition in MTurk samples across the typical demographics.

Leveraging the Cloud

Turkers respond to survey requests incredibly quickly, which is especially valuable when conducting surveys intended to explore how individuals respond to real events in an observational setting. Indeed, the pre-post event design has the unique potential to capture any micro-level change in this context, which is one of several typical reasons to collect panel data (see Sharot, 1991). Moreover, because events in the real world unfold quickly and often change with the discovery of new information, related events or media coverage, it is important that the panels surrounding the event are tightly situated around it. MTurk makes it possible to launch a survey and collect hundreds of responses within hours. Even in latter waves of our panel, we registered dozens of responses in the first several minutes of posting the HIT. In our case, this allowed us to end waves in anticipation of the health care decision and start a new one right after it was announced. While we did not need to, it also would have allowed us to quickly go back into the field with another wave if, for example, President Obama had made a health care speech a few days after an adverse decision. Thus, along with MTurk's low cost and ease of use, the low panel attrition and quick response times of the MTurk sample provide unusual flexibility for researchers to quickly adapt their research design to real world events. In our study of the ACA decision, scheduled weekly panels would have been much less efficient than panels quickly and easily adaptable to the event's uncertain timing. Our case is not unique; for example, one who is interested in public opinion in response to natural disasters could time surveys around forecasts and collect data in tight windows as events unfold.

A related benefit of doing MTurk panel studies around events is the ability to do what we are calling *real-time* survey experiments, in which one randomizes participants to realistic treatments related to the event as they are simultaneously getting political information about it in their natural environments. Relative to conducting artificial experiments with captive audiences, this approach dramatically increases external validity and offers a unique opportunity to combine experimental treatments with the unfolding events in the real world, thereby reconciling some of the known discrepancies in survey and experimental studies (Hovland, 1959). Indeed, there has been some noteworthy attention to the implications of captive audiences in political science (e.g., Arceneaux and Johnson, 2010; Gaines and Kuklinski, 2011; Gaines, Kuklinski and Quirk, 2007; Levendusky, 2011) and even to realistic experiments of attitude change in the longitudinal context (Druckman, Fein and Leeper, 2012). However, outside of using an MTurk sample with an online survey platform or a convenience sample, combining a large N panel design around imprecisely timed real world events with realistic experimental treatments would likely be too costly and/or would require sacrificing external validity due to its advanced planning demands.

Our approach using MTurk allows one to not only capture the micro-level change around the event, but also to leverage any effects of the event with a related experiment. That is, we might be interested in not only attitude change as a result of an event, like the ACA decision, but also with factors associated with the event that can be experimentally manipulated, such as media frames. In our case, we took advantage of news reports that Chief Justice Roberts switched sides for political reasons that broke within days of the decisions. While a typical experiment might expose participants to this story, and this story only, weeks or months after a decision to investigate the effect of a story about politicians in robes, MTurk enabled us to expose some participants to it as it was unfolding and as they were also self-selecting other news and information about the decisions. An example of a similar application would be studying the effects of media frames leading up to and/or after presidential debates. MTurk would allow a researcher to easily conduct a pre-wave and then expose participants to different media frames in the days after the debate to estimate their effect on people who are also exposed to other uncontrolled information.

To be sure, our health care study is but one example of using MTurk to conduct an inexpensive and flexible panel study around a foreseeable political event. It enabled us to ask dozens of questions of our choosing in each wave, to adopt the timing of our surveys to events as they unfolded, and to embed experiments by exposing participants to a little-known yet important news story as it broke. Supreme Court cases are but one potential application. As we mentioned above, whether one wants to collect pre-post data around precisely scheduled events, like presidential debates or the release of economic data, or those that can only be forecast a few days in advance, like natural disasters or major legislative votes, MTurk makes it easy to interact political events with experimental treatments in real-time. It also allows one to simply conduct a standard panel design at a workable price. In all, MTurk makes sophisticated research designs possible for those with constrained research budgets and introduces little in the way of inferential bias.

References

Arceneaux, Kevin and Martin Johnson. 2010. "Does Media Fragmentation Produce Mass Polarization?" Temple University. Unpublished paper.

Bartels, Larry M. 1999. "Panel Effects in the American National Election Studies." *Political Analysis* 8(1):1–20.

Berinsky, Adam J., Gregory A. Huber and Gabriel S. Lenz. 2012*a*. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3):351–368.

Berinsky, Adam J., Gregory A. Huber and Gabriel S. Lenz. 2012*b*. *Mechanical Turk Instructions*. Available at `https://docs.google.com/document/pub?id=1Bjqp0FSTmEG83S-vtbr0uFrdesq1OSC27Xbp_iCI8Eg.`

Berinsky, Adam J., Gregory A. Huber and Gabriel S. Lenz. 2012*c*. *Recontacting Mechanical Turk Workers*. Available at `http://docs.google.com/View?id=dd4dxgxf_9g9jtdkfc.`

Berinsky, Adam J., Michele. Margolis and Michael W. Sances. 2012. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Internet Surveys." Presented at NYU CESS 5th Annual Experimental Political Science Conference .

Christenson, Dino P. and David M. Glick. 2012. "Roberts's Health Care Decision Disrobed: The Microfoundations of the Court's Legitimacy." Presented at the Meeting of the Pacific Northwest Political Science Association, November 8–10, 2012. Portland, OR.

Druckman, James N., Jordan Fein and Thomas J. Leeper. 2012. "A Source of Bias in Public Opinion Stability." *American Political Science Review* 106(2):430–54.

Gaines, Brian J and James H. Kuklinski. 2011. "Experimental Estimation of Heterogeneous Treatment Effects Related to Self Selection." *American Journal of Political Science* 55(3):724–736.

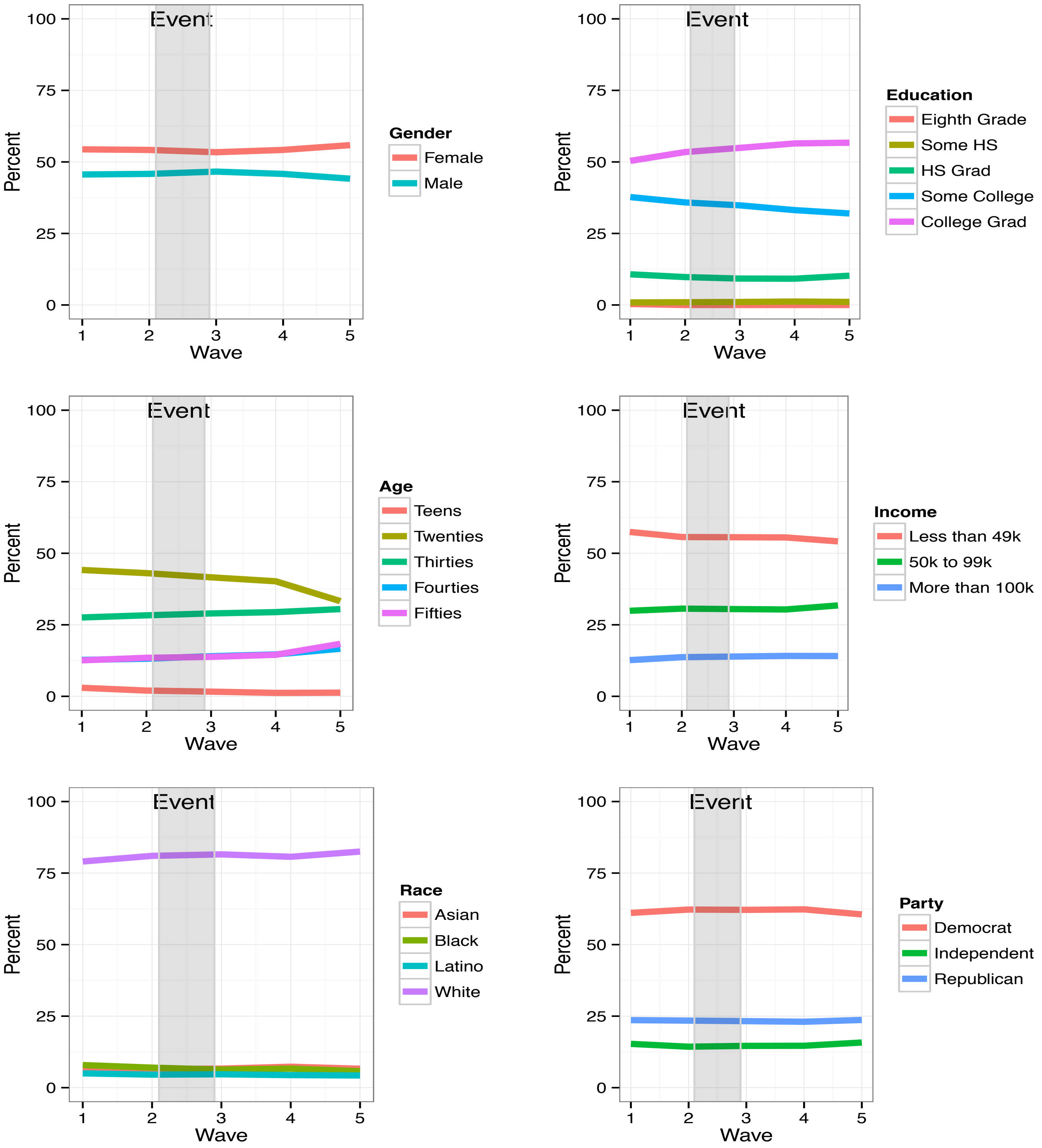
Gaines, Brian J., James H. Kuklinski and Paul J. Quirk. 2007. "The Logic of the Survey Experiment Reexamined." *Political Analysis* 15(1):1–20.

Hovland, Carl I. 1959. "Reconciling Conflicting Results Derived From Experimental and Survey Studies of Attitude Change." *American Psychologist* 14(1):8.

Levendusky, Matthew S. 2011. "Do Partisan Media Polarize Voters." University of Pennsylvania, Unpublished Paper.

Sharot, Trevor. 1991. "Attrition and Rotation in Panel Surveys." *The Statistician* 40:325–331.

Figure 1: Sample Demographics by Panel Wave



Event refers to the period between the second and third waves when the Supreme Court ACA decision was announced.