8

SCALABLE MULTIDIMENSIONAL RESPONSE MEASUREMENT USING A MOBILE PLATFORM

Philip Resnik, Amber E. Boydstun, Rebecca A. Glazier, and Matthew T. Pietryka

Today's astonishing level of technological connectivity presents new opportunities for measuring people's attitudes and obtaining a deeper understanding of opinion formation, ranging from online surveys (Evans & Mathur, 2005) to textual analysis of microblog postings (Lucas et al., 2015). As Maurer and Reinemann (2009) point out, however, when it comes to responses to communicative stimuli, not all measurement methods are equally effective: without continuously measuring responses over the course of a communication event, it is quite difficult to obtain valid data about the *causes* of opinions and changes to opinions in that event. Methods for real-time response (RTR) measurement help to make progress on the causal question by capturing reactions and making it possible to tie them back to moments within the stimulus, permitting inferences about which aspects of the communication were affecting which people, and in what way.

In this chapter, we describe an approach to real-time response measurement using a mobile platform that permits instantaneous responses from large numbers of participants using smartphones and other mobile devices. The technology is similar in spirit to dial testing and retains its primary advantage—responses are viewer initiated and virtually instantaneous, thereby allowing us to capture and analyze unmediated viewer reactions as opposed to digested opinions (Brubaker & Hanson, 2009; Fridkin, Kenney, Gershon, Shafer, & Woodall, 2007; Maurer & Reinemann, 2009; Tsfati, 2003). At the same time, our approach takes advantage of the scalability and flexibility of mobile technology to offer unique benefits in tracking real-time reactions to a live event.

In the next section, we discuss the considerations that went into our design and describe the core elements of the technology. Then we consider the crucial question of how accurately responses are mapped back to moments in the communicating stimulus, presenting empirical results on the temporal resolution of

reactions. We turn next to two illustrative applications of our approach. First, we describe a qualitative case study in one of the most common domains of application for RTR methods, measurement of responses to television advertisements—in this case a commercial engagement studying the effectiveness of ads broadcast during Super Bowl XLVII in February, 2013. Second, we describe a study focused on another common-use case: the measurement of citizen responses to candidate messaging during political debates. We illustrate our approach by using it to study the effects of candidate messaging on attitudes during the October 3, 2012, presidential debate between Barack Obama and Mitt Romney, including a recruiting technique taking advantage of our distributed platform, which made it possible to obtain a much larger and more representative sample than one would find in any single-location study. Finally, we summarize and discuss future directions.

Design

Motivating a Mobile Platform

The technological approach described here was inspired by an analysis of several major existing methods for opinion gathering during communicative events and the tradeoffs they create, with the idea of creating a "best of breed" combination to bring together the most important positive attributes of existing methods with as few of the negatives as possible.

Traditional questionnaire-based surveys permit careful fine-tuning of how questions are asked, make it easy to collect detailed individual-level data about demographics and issue preferences, and are equally easy to apply with either balanced samples or convenience samples. However, they do not allow in-the-moment, unmediated responses or the ability to relate reactions back to the stimuli that caused them (Maurer & Reinemann, 2009).

Conversely, dial tests, the most commonly used form of real-time response, measure instantaneous, user-initiated reactions. However, conventional hardware-based dials are expensive and limit the size of the sample that can be studied. Additionally, all dials (including more recent mobile variants) are restricted to tracking responses on a single dimension or scale (e.g., positive vs. negative impression, convincing vs. unconvincing, etc.). Maier and Fass (2009) suggest that this kind of single-dimensional measurement may be a poor match for analysis of reactions that arise from complex information processing, arguing that "researchers cannot reconstruct on what basis a subject has chosen a particular position on a given scale" and that "a specific dial position does not tell us very much about the underlying judgments" (p. 18). Considerations in their discussion include the fact that dials do not explicitly distinguish positive from negative impressions, instead recording a composite measure that can combine both influences; that it can be difficult to interpret reactions when two or more objects, such as candidates in a debate, are on screen simultaneously; and that after a new judgment

a dial remains in its position without returning to neutral, making it difficult to infer what is happening in the participant's mind when the position of the dial remains unchanged. Maier and Fass (2009) advocate instead for push button responses, which can address these concerns and provide more straightforwardly interpretable evidence about when cues in the stimulus have or have not engaged participants enough to evoke a specific response. Our method can be viewed as a large-scale instantiation of this approach.

As another opinion-gathering method, analysis of social media is providing a new and interesting alternative to traditional measurement methods; for example, during large shared-watching events, Twitter provides a large-scale, continuously running stream of opinions expressed in everyday language, and techniques have been developed to link the Twitter stream back to the event being watched (see Roy, 2005; Fleischman & Roy, 2008). Unfortunately, social media streams provide very little by way of reliable user demographics, and social media language is so unconstrained and messy that text analytics methods tend to have limited accuracy, even for coarse-grained distinctions like positive vs. neutral vs. negative sentiment.

These considerations led us to the design of a mobile platform, the core of which is a web application tailored for mobile devices such as smartphones, tablets, or laptop computers. Taking a mobile approach enables high scalability (in its current form, the platform can support on the order of 80,000 concurrent users during a shared viewing event, and the architecture supports further expansion). Implementing a "web app" in particular ensures high accessibility, since the same web app can run on most major platforms (iPhone, Android, etc.), and no download or installation is required. Users enter the app simply by tapping a link they receive via e-mail, text message, on a web page, or in social media, or by typing a URL into their device's browser. Furthermore, participants can be recruited using any sampling method, e.g., traditional probabilistic sampling, *ad hoc* convenience sampling, or snowball sampling. The ability to obtain appropriate samples is, of course, connected with the penetration and distribution of mobile devices in the population of interest, but the prospects for reaching individuals who use mobile technology are improving over time, particularly as compared with probabilistic sampling that relies on landlines. Below, we describe a sampling method especially well suited for recruiting from undergraduate populations.

A mobile app also enhances external validity, since not only are people able to respond in whatever their natural watching environment might be—rather than more artificial focus-group settings (see Ramanathan, McGill, Phillips, Schill, & Kirk, 2010)—but more and more people these days are already engaging in "second screen" experiences on their mobile devices as a natural part of how they consume media (Giglietto & Selva, 2014; Nielsen, 2014). At the same time, nothing prevents the app from being used in more formal experimental settings where, for example, outside distractions can be limited or controlled.

The User Experience

Having discussed design considerations associated with the *kind* of app we developed, we now turn to the design of the user experience itself. Once having entered the app, as described above, there are three parts to the user experience.

Entry Survey

The first component is a traditional questionnaire of any length, where the response types can include multiple choice (with radio buttons or pulldown menus for forced choice, or checkboxes to permit multiple answers), a horizontal slider (with labels at extreme left, midpoint, and extreme right, producing a value from 0 to 100), or free text response. By using the questionnaire on entry to the app, it is possible to collect detailed information such as demographics, issue preferences, and opinion baselines for studies of attitude change, as well as to provide participants with instructions.

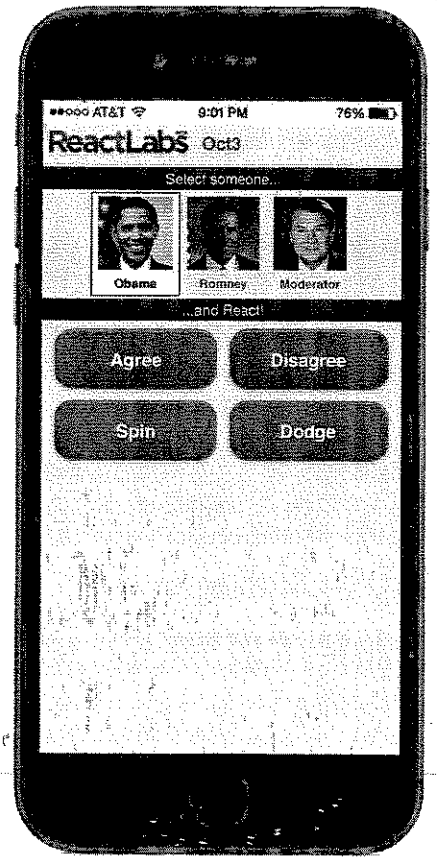

FIGURE 8.1 Real-Time Reactions Screen for the October 3, 2012 Presidential Debate.

Real-Time Reactions Screen

The second part of the user experience is the core of the app: a "real-time reactions" screen that permits user-initiated responses at any time. Figure 8.1 illustrates the real-time reactions screen as it was configured when our approach was applied during the October 3, 2012, presidential debate. At the top of the screen, a set of *reaction targets* is provided in the form of labeled thumbnail images. In Figure 8.1 these include the candidates as well as a target for the moderator.

Below the targets, a set of *reaction types* (or *reaction buttons*) permits multiple reactions. In this example, *Agree* and *Disagree* separately capture positive and negative impressions, and the choice of labels explicitly and continuously cues the user to specifically evaluate the *statements* candidates are making. Users specify reactions through a two-tap sequence, first tapping a target (which un-grays the reaction buttons to make them active) and then tapping a reaction button (which produces a very brief confirmation at the top of the screen, registers the reaction with its timestamp, and then grays the reaction buttons back out). Figure 8.1 shows the screen between the two taps, where a user has selected Obama but not yet tapped the reaction button.

Like the entry survey questions, the real-time reactions are flexibly configurable. For example, during a debate between Democratic candidates in the Maryland gubernatorial primary in May, 2014, the *Washington Post* was particularly interested in whether people watching the debate found candidates' statements truthful or believed they should be fact checked, in addition to whether the viewer supported or opposed what the candidate was saying. In a collaborative study for that event, the reaction buttons were *Truth*, *Fact Check!*, *I support*, and *I oppose*.

As Figure 8.1 illustrates, having multiple targets and responses makes it possible to collect finer-grained data than the single dimension offered by conventional dials. With regard to targets, this approach makes it possible to avoid uncertainty about who or what is being responded to; for example, during a debate the candidates might be engaged in a rapid back-and-forth exchange, or a viewer might simultaneously like what one candidate is saying while disliking the other candidate's facial expression in response to it. With regard to reactions, the availability of multiple reaction types makes it possible to fine-tune the data being gathered, for example by distinguishing whether a negative reaction is taking place because a viewer opposes the candidate's point of view or believes the candidate to be untruthful, without restricting the responses to one or the other. In a single-dimensional scenario, participants might be carefully instructed only to move the dial toward unfavorable when they hear a statement they oppose, not when they think the candidate is lying. However, even if such instructions are easily understood and followed consistently by participants, multidimensionality makes it easier to connect with a wider range of participants' responses, not only enriching the collected data but also creating a more natural experience. Indeed, as we discuss in the section on assessing advertising effectiveness, it is possible to

include elements of multidimensional response that exist not only for the purpose of data collection, but specifically to create a greater degree of engagement and naturalness.

Although easily overlooked, another configurable facet of the real-time reactions screen is the text above the reaction targets and reaction buttons (e.g., *Select someone ... and react!*). We have found that simple formulations of this kind permit users to immediately begin engaging in our two-tap mode of responses, even with minimal instruction. The configurability of these prompts also makes it easy to create multidimensional responses where the targets are *facets of a communication*, rather than distinct people speaking. For example, during a speech proposing a new policy, the app could be configured so that the first tap (*Select one of these ...*) selects among, say, *Idea*, *Feasibility*, and *Presentation*, and the second tap (*... and react!*) captures, say, *Excellent*, *Awful*, *Exciting*, and *Zzzzzz*. Multidimensional responses on this scheme would, therefore, be able to distinguish moments when participants found an idea exciting even though its feasibility was awful or the presentation was putting people to sleep.

Putting the pieces together, we have found that the real-time reaction experience provides a way to create significant engagement during a communicative event. During the 2012 presidential debates, for example, the app was used by WJLA-TV (the Washington, D.C. ABC affiliate) to engage viewers: the opportunity to participate was promoted on-air and digitally, and real-time charting of results (in the style of the CNN "People Meter") was available on the web and used in news coverage. Without any explicit incentive to participate, this invitation process for the first debate yielded 794 active participants, 233 of whom registered a reaction ten or more times during the debate (88 reacted 50 times or more). Moreover, users who are responding using our app tend to keep responding over the course of the entire event, providing measures of respondents' attitudes over a significant span of time.

The design of the real-time reactions screen addresses a number of the limitations of dial testing discussed by Maier and Fass (2009) by providing more specific information about participants' impressions and what elements in the stimulus are causing them. Of course, this advance is achieved at the cost of obtaining categorical rather than graded judgments. However, aggregation across a large sample provides an alternative way to quantify the strength of response to a cue, and recording discrete reactions—what Maier and Fass call the "reset mode" (p. 17)—which is a particularly good match for our interest in understanding when a cue has *engaged* participants enough to evoke a response. In the context of political communication, the choice to measure viewers' judgments discretely, focusing on the moments when they react on their own initiative, allows us to track not only the nature of the response, but also when they have passed a participant-specific, minimal threshold of effort to take action—even action as small as a click. If a candidate can get a viewer to click—analogous to other forms of minimal

political engagement (Shulman, 2009; White, 2010)—it may represent the first rung in a "ladder of engagement" (Karpf, 2010, p. 16) leading to more substantive mobilization.

"Snap" Surveys

A third functionality of the platform is the ability to easily create a set of one or more survey questions and push it out to participants at any time during the communicating event. Questions can be set up in advance, and *ad hoc* questions can also be generated on the fly by the researcher in response to specific statements or unexpected issues that emerge dynamically over the course of a live event. Generating and distributing an *ad hoc* question takes little more time than typing the question in, so snap surveys appear with minimal delay. The inventory of question types (radio buttons, checkboxes, etc.) is identical to the entry survey, and questions distract minimally from the user-initiated reactions experience on the real-time reaction screen. When a snap survey is initiated, a pop-up appears on the user's mobile device, a click brings the user to one or more questions, and a final "Submit" clears the snap survey and returns the user to the real-time reactions experience.

Snap surveys complement the real-time experience by making it possible for the researcher to generate specific prompts at any time, in addition to collecting user-initiated responses as described above. One important use for this capability is in looking at how responses change over time. For example, snap surveys asking the same question can be given to participants multiple times during the event, and then again after the event (e.g., "If you had to vote right now, which candidate would you choose?").

Alignment of Responses to Stimulus

A key issue for all RTR approaches is internal validity, which concerns "the question [of] whether RTR really measures what it is supposed to measure" (Maurer & Reinemann, 2009, p. 10). First among questions of internal validity is the reliable identification of what evokes participant response. Our user interface design creates greater confidence in internal validity by using the two-tap strategy to include the target of a reaction as an explicit component of the response, removing uncertainty about the target, exclusive of user error. However, just as for dial tests and other RTR approaches, alignment of reactions to the specific stimulus is unobserved and needs to be inferred.

In order to investigate this dimension of internal validity, prior to finalizing the design of the mobile interface we piloted a prototype with 359 respondents. Participants watched a segment of an unfamiliar debate (students in California and Arkansas watched a 2-3 minute excerpt of a 2008 Republican primary debate

among U.S. Senate candidates in New Jersey) and used a web-based version of the tool to register their reactions. Using the tool's timestamp information, these reactions were mapped to short segments of the debate transcript; the segments, of roughly equal length, averaged 14.6 words (95% confidence interval 13.7 to 15.5) lasting an average of 4.9 seconds (95% confidence interval 4.6 to 5.2), and each was typically smaller than a complete sentence. Immediately following the reaction session, each participant was shown an editable summary of how their reactions had been automatically mapped to segments in the debate using the timestamps. A user could move a reaction to the previous or next segment, delete it entirely, or leave it untouched to indicate that it was correct.

Results of this study established that 97% of the time, participants indicated that their responses had been synchronized correctly to the (sub-)sentence segment to which the user was responding. Studies comparing mouse-based and touchscreen reactions show that people using touchscreens are not significantly slower than—and are, if anything, faster—than people using a mouse, with respect to reaction times (Findlater, Froehlich, Fattal, Wobbrock, & Dastyar, 2013; Sears & Shneiderman, 1991). Therefore, we are confident that the mobile app captures people's reactions at a comparable temporal resolution to the web prototype.

Particularly because the mobile app enhances the ability to include geographically widely distributed participants during the same event, another question to consider is whether the alignment of reactions to stimulus might be affected by communication latency. For every reaction, the mobile app records the timestamp when the reaction was received on the system's server. (We experimented with using client device timestamps but found that too many people's device clocks are set inaccurately.) With measured averages of cellular network latency around 100 to 300 milliseconds as of 2012, and broadband much faster (Podjarny, 2012), network transmission time is of minimal concern. More interesting, especially for a nationally televised live event, is the assumption that all participants are watching the same thing at any given moment. This assumption turns out not to be strictly true. We analyzed the video for a sample of 20 local broadcasts of the October 3, 2012 debate from cities around the country, and found that, even for the same network broadcast, there were small variations in timing among local broadcasts, on the order of 2-3 seconds. For example, a participant watching the ABC broadcast on KVUE in Austin, responding to a statement by Obama at 9:00:06 pm U.S. Eastern time October 3rd, might have been reacting to the same thing a participant watching ABC/KMGH in Denver saw at 9:00:09 pm Eastern. Differences between standard and high definition cable broadcasts can introduce delays on a similar scale.

Our approach to this issue is to create a uniform mapping of the full dataset to a single time-stamped reference transcript, by manually synchronizing a small set of time points. For the October 3 debate, for example, we manually aligned the begin and end time points of 21 candidate statements at the start and end of the debate to their corresponding waves of response to that statement (aggregating over all reaction types). This alignment produced a single standard mapping between

transcript timestamps and reaction timestamps. The lengths of the speaking windows correlate with the lengths of the response windows with $r = 0.99$, lending confidence that the sets of reactions are mapping to the correct statements. Nonetheless, in order to account for uncertainty as to which specific sentence was being reacted to for any given reaction, owing either to broadcast differences or participants taking a few moments to react, we utilize five-second rolling windows in our analyses. We are also developing computational methods that use correlated responses among groups of participants to automatically infer a reference mapping of reaction-worthy moments and individuals' temporal offsets relative to that reference, with promising preliminary results (Julien & Resnik, 2013).

Assessing Advertising Effectiveness

Maurer and Reinemann (2009) observe that real-time response measurement originated in commercial research involving radio and TV, and also that many recent studies are done in the context of televised political debates. Here we briefly describe a commercial application of the app for data-driven qualitative analysis, and then in the next section we shift to a more formal study in the context of a U.S. presidential debate.

In a commercial context, few applications present a greater economic opportunity for real-time response technology than measurement related to advertising. Beyond a huge market for television ads alone, the field of advertising is rapidly evolving new settings in which ads are presented (e.g., very short spots at the beginning of YouTube videos). Moreover, particularly as digital advertising takes on an increasing role and connects with traditional advertising, there is increasing pressure in the industry to go beyond shallow measures, such as whether an ad was seen or clicked through, to richer and more informative measures of behavior and advertising effectiveness (Haile, 2014).

The application of our approach in commercial research is illustrated by an engagement with Frank N. Magid Associates, a large research-based strategic consulting firm. React Labs, a company commercializing the technology we describe here, worked with Magid Associates to collect data on advertising effectiveness for the 2013 Super Bowl for qualitative analysis by their research team. Magid Associates has historically been a well-known industry source for analysis of Super Bowl ads, providing widely-read commentary and assessments. However, their reports on the Super Bowl have conventionally been based on their experts' subjective impressions and their experience base, not on data-driven analysis.

The Magid Associates study illustrates several aspects of our approach described in the design section above. Most obvious are scalability and accessibility. Participants were solicited via a two-pronged strategy: a mailing to one of Magid's established panels offering entry in an Amazon gift card sweepstakes as an incentive to participate, and social media invitations to participate distributed on Facebook and Twitter. Without either significant cost for incentives (a total of $250 in gift

FIGURE 8.2 Real-Time Reactions Screen for the 2013 Super Bowl.

cards) or a concerted social media push, the study obtained a sample of about 400 participants distributed nationwide. Since React Labs is a web app, no download from an app store or installation was required—users simply tapped the URL that had been e-mailed or texted to them, on their smart device, and the app came up in their browser.

The multidimensional configuration of the app (see Figure 8.2) made it possible to design the survey with an emphasis on external validity. Beyond the fact that participants were responding in their natural game-watching settings, targets for reactions in the app included not only the commercials, but also the two teams and the halftime show. This design had the effect of keeping people engaged with the response experience throughout the entire game, as illustrated by significant bursts of response to touchdowns, other major plays, and disputed calls by the referees.

Finally, in addition to the discrete 4-point scale for like/dislike, two response types provided additional dimensions for analysis. One, labeled *I want it*, enabled the study to capture in-the-moment data not just on whether participants liked

an ad, but whether and where they were activated with regard to an interest in the product. This came out, for example, during a commercial for SodaStream (a consumer home carbonation product). Many other industry discussions overlooked this spot, focusing more on what the Magid analysis described as "the often silly theatricality" of other commercials (Frank N. Magid Associates, 2013, p. 2). However, their analysis suggests the SodaStream spot was effective at driving consumer activation, distinct from positive affect toward the ad: in addition to a concentrated activation of *Like lots* responses from participants as the product was clearly and simply demonstrated, an *I want it* peak identified a burst of favorable interest in the product. According to the Magid Associates qualitative analysis, "no other spot on Sunday night showed a similar density of activation scores ... Combined with powerful visuals and evocative sound effects, they hammered home a simple value proposition and drove consumer response" (Frank N. Magid Associates, 2013, p. 2).

The remaining response type, labeled *Blah,* provides a particularly interesting illustration of the value of multidimensional responses. It was included in this study as a way for participants to provide an *observable* response indicating lack of interest, not just because metrics from that response might prove valuable, but also to keep people engaged with the app experience even when they found what they were watching uninteresting. For example, ahead of the game Lincoln Motor Company had heavily teased and marketed their ad called "Steer the Script." Nonetheless, the results showed a fair proportion of *Blah* responses, particularly during the middle of the ad—which, distinct from *Dislike some* or *Dislike lots,* suggested that participants were not disliking this part of the ad so much as finding it boring or not compelling.

The app's "snap survey" functionality was also used during this engagement, eliciting questionnaire responses in a targeted way by popping questions up on their devices, and then returning participants immediately to the real-time reactions interface. This function made it possible to measure brand recall at a given interval after the conclusion of a particular commercial.

Although more a case study rather than a formal scientific study, the engagement with Frank N. Magid Associates illustrates the promise of a methodological approach in which a large, broad audience can be tapped for multidimensional responses in their natural viewing settings using an easily accessible mobile platform. Below, we illustrate more formal analysis using data collected in the first 2012 Presidential Debate.

Measuring Reactions During a Presidential Debate

Studying Debate Reactions

Debates serve a singular role in elections: they uniquely provide candidates unmediated access to a large and diverse audience (Trent & Friedenberg, 2008), including marginally attentive citizens (Pfau, 2003) and undecided voters (Geer,

1988) who use debates to learn about the candidates (Blais & Perrella, 2008 Holbrook, 1999; Lemert, 1993). In the U.S., debates are the most visible, widely watched events of a presidential campaign (Benoit, Hansen, & Verser, 2003 Schroeder 2008).

It can, however, be a challenge to measure the effect of specific candidate messages on individual attitudes. Most mainstream polls collect aggregate data only after a debate has finished (e.g., Holbrook, 1999; Shaw, 1999), making individual-level conclusions impossible. And most large-scale individual-level research on debates also relies on post-debate evaluations (e.g., Abramowitz, 1978; Geer, 1988 Hillygus & Jackman, 2003; Steeper, 1978). Whether surveys are cross-sectional (e.g., Lanoue, 1992; Sigelman & Sigelman, 1984) or panel designs (e.g., Kraus & Smith, 1977; Tsfati, 2003), the data cannot differentiate between the effects of the debate itself and other influences, such as media coverage of the debates (Brubaker & Hanson, 2009; Fridkin et al. 2007). Moreover, these studies cannot isolate *which* candidate messages are influencing viewers. Recent work indicates that researchers cannot trust survey respondents to self-report accurately even whether they watched the debate (Prior, 2012). Thus, while past research has contributed greatly to our understanding of debate effects (Bartels, 2009; Benoit et al., 2003; Geer, 1988; Holbrook, 1999), scholars have often been reduced to educated guesswork about which specific candidate cues produce these effects.

Recruiting Participants: Colleague Crowdsourcing

In order to rigorously test our approach, we needed a large and diverse set of participants. We recruited them through a process of colleague crowdsourcing. The term *crowdsourcing* was introduced in a *Wired* magazine article in 2006 (Howe, 2006a) and the idea rapidly caught on; as Howe (2006b, sidebar) defines it, "Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call" (see also Quinn & Bederson, 2011). The introduction of the crowdsourcing concept has had a dramatic effect in industry and in many fields of research, including political science, where, for example, scholars have turned to services such as Amazon Mechanical Turk to recruit individuals to complete surveys or code data (Grimmer & King, 2011), yielding larger and more granular datasets than previously possible.

Our approach to recruiting participants was to recruit instructors, who would themselves recruit students to our study. This approach created the need for a chain of incentives: we encouraged instructors to promote the app to their students and, if willing, offer students extra credit for participating in our study. In turn, we created an instructor package designed to help them incorporate watching the debates and using the app into their course materials. The instructor package included PowerPoint slides and lecture notes for a lecture on presidential debates, questions to generate class discussion, pointers to additional resources and

relevant literature, and alternative assignments. Additionally, participating instructors received the list of students who had earned the extra credit, and, within 12 hours after each debate, presentation-ready slides with preliminary results and analysis. (All of our materials are available at http://reactlabseducate.wordpress.com/.) The nationwide recruiting of instructors was done via e-mail contact with colleagues, invitations distributed on relevant mailing lists, and outreach to relevant political science blogs (e.g., *The Monkey Cage*, *Active Learning in Political Science*), which provided coverage of the app and our recruitment efforts. While neither broad-scale recruitment nor gaining access to student respondents through instructors are new ideas (for example, see Carlin & McKinney, 1994), to our knowledge our study is somewhat unique in the large scope of our data collection effort and the external validity provided by allowing respondents to participate in a natural environment.

Recruiting Outcomes and Properties of the Sample

Student participants represented all 50 states, the District of Columbia, and Puerto Rico, as well as participants in Canada, France, England, Ireland, the Netherlands, and Kuwait. In total, 263 instructors registered at least one course to participate in at least one debate, totaling 361 courses and more than 13,000 potential student respondents. Across the 3 presidential debates and the vice presidential debate, we received 8,006 participants (some of whom participated in more than one debate; nearly 5,000 unique individuals participated at least once). Participation peaked with the first presidential debate, with 3,340 participants. (Most students participated for some form of credit, with names and course IDs, collected in the post-debate survey, providing reasonable confidence in participant identities. In more general settings the app can support a login mechanism.)

In terms of demographics, the colleague crowdsourcing succeeded in recruiting a set of students with the diversity required to test hypotheses on the basis of individual-level, real-time observations. As Table 8.1 shows, the diversity of the students who participated is comparable to the national population means in terms of gender, income, race, party identification, and religion, though clearly not in age as our recruitment efforts were targeted at college undergraduates. Although this undergraduate sample still has a host of generalizability issues, the data represent a leap forward for sample quality, with a level of diversity that would be difficult to achieve in localized or regional samples of the same population. Moreover, the large multi-campus sample provided significantly more variation across a range of variables, allowing for unbiased estimates of heterogeneous treatment effects (Druckman & Kam, 2011). Table 8.2 shows the number of students who took part in the debate study by partisanship and race/ethnicity. The table shows that the large total number of respondents provided a significant number of responses in all cells—even for rare combinations such as African American conservatives.

TABLE 8.1 Study Demographics Compared to National Demographics

	App		*National*
	N	*%*	*%*
Gender[a]			
Women	3,789	48	51
Men	4,099	52	49
Family Income[a]			
<$25K	1,232	16	18
$25K–$49,999	1,236	16	24
$50K–$74,999	1,397	18	19
$75K–$99,999	1,140	14	14
>=$100K	2,868	36	26
Race[a]			
African American	694	9	13
Asian	679	9	5
Hispanic	1,054	13	17
Other	418	5	2
White/Caucasian	5,120	64	63
Party ID[b]			
Democratic (includes leaners)	4,215	54	50
Independent	1,235	16	11
Republican (includes leaners)	2,396	31	39
Religion[c]			
Christian	4,737	60	76
Jewish	381	5	1
Muslim	157	2	<1
Atheist or agnostic	2,069	26	15
Other	616	8	8
Age[d]			
18–24	6,830	85	13
25–29	448	6	9
30–39	366	5	17
40–49	183	2	18
>=50	179	2	43

Notes: App estimates include all 8,006 participants across the 4 debates, including those who participated in more than 1 debate. The numbers do not total 8,006 on any given demographic item due to non-response on that item.

[a] National estimates are from the U.S. Census.

[b] National estimates are from the Pew Research Center for the People, October 2012, accessed January 23, 2013, from the iPOLL Databank, The Roper Center for Public Opinion Research, University of Connecticut. Available at http://www.ropercenter.uconn.edu/data_access/ipoll/ipoll.html.

[c] National estimates are from the 2008 American Religious Identification Survey.

[d] National estimates are from the 2012 American Community Survey One-Year Estimates.

TABLE 8.2 Participant Frequencies by Ideology and Race/Ethnicity

	Asian	*African American*	*Hispanic*	*Caucasian*	*Other*	*Total*
Liberal	348	368	468	2,080	201	3,465
Moderate	249	262	429	1,390	152	2,482
Conservative	79	60	149	1,606	62	1,956
Total	676	690	1,046	5,076	415	7,903

Notes: Ideology and race were measured in the pre-debate survey. Ideology was measured with a 100-point sliding scale ranging from 0 (extremely liberal) to 100 (extremely conservative). In the table, participants scoring between 0 and 39 on this scale are classified as liberal, between 40 and 60 as moderate, and between 61 and 100 as conservative.

Analyzing the First Obama/Romney Presidential Debate, October 3, 2012

The adoption of this technology by college instructors around the country meant the student debate-watching experience changed for many students. Students are sometimes reluctant to engage politically, but the salience and scale of presidential debates present a key opportunity to encourage political engagement. Political engagement, in turn, is correlated with heightened political knowledge and civic skills, especially among those with lower initial levels of political interest (Beaumont, Colby, Ehrlich, & Torney-Purta, 2006), and can have a positive impact on students' future political engagement and voter turnout (Hillygus, 2005).

In addition to the potential democratic benefits our technology might offer, it provides scholars of presidential debates with data at a level of detail not seen before. What might these data be able to tell us? We illustrate by examining a research area that has long interested scholars of political communication: the development and control of political agendas. Agenda building (or agenda setting) is the process by which policy problems become topics of political discussion (Erbring, Goldenberg, & Miller, 1980; McCombs & Shaw, 1972). With finite agenda space, topics get attention at the necessary expense of other topics, meaning candidates can use the ability to define "what politics is about" (Schattschneider, 1960) as a powerful tool for building coalitions and gaining votes (e.g., Baumgartner & Jones, 2009; Jones & Baumgartner, 2005; Kingdon, 1995; McCombs & Shaw, 1972).

Prior research emphasizes a critical three-part question, what Iyengar and Valentino (2000) call "the classic shorthand of message learning theory—who says what to whom?" (p. 110). This mantra reminds us that we must attend to the entirety of a candidate's message: messenger, message, and audience. Within the debate literature, however, data limitations have precluded answering questions about how message sources and *specific* message cues influence viewers generally, or how responses might differ across viewers. As demonstrated below, our methodological approach allows us to illustrate how variation in each element—messenger, message, and audience—contributed to viewer responses in the first presidential debate of 2012.

Data and Methods

Participants used the app to complete a detailed pre-debate survey, including standard demographic and attitudinal questions and questions about issue priorities. In order to obtain real-time response data relevant to our questions, the real-time reactions screen of the app was configured as shown in Figure 8.1. Other than the instructions on the screen itself ("Select someone … and React!"), participants were given no explicit instruction on the interpretation of these labels. Although such instruction could easily be included in a pre-debate survey, our assumption was that the common, everyday understanding of these terms would be sufficiently clear to elicit coherent responses, and we did not wish to interfere with participants' naturally occurring thought processes or increase cognitive load by requiring that instructions be referred to or remembered.

Results

To identify candidate messages, we performed a manual content analysis of the debate transcript. The supplementary material for Boydstun, Glazier, Pietryka, & Resnik (2014), available online at http://poq.oxfordjournals.org/, contains our complete codebook. We divided the transcript into quasi sentences (i.e., separate clauses; see Boydstun, Glazier, & Pietryka 2013), which were manually time-stamped. Each quasi-sentence was coded for the candidate speaking and the *primary topic* (using the Policy Agendas Topics codebook, Baumgartner and Jones, 2006, available at www.policyagendas.org). Intercoder reliability was strong: based on a randomly sampled 75 quasi-sentences, coders registered 94.6% agreement (Cohen's kappa = 0.924).

Combining our time-stamped, coded transcript data and our dataset of participants' real-time reactions and individual-level variables enables us to investigate questions that have previously not been addressed through systematic data analysis. We structure our discussion around the concept of *net positive engagement*, which we define operationally as the average number of *Agree* clicks minus *Disagree* clicks per viewer targeted at a given candidate in the five seconds following that candidate's discussion of a given topic.

For our analyses of the effects of candidate cues on viewer engagement, we include any five-second rolling window in which the candidate discussed the topic. The unit of analysis is the participant-second. Since our study had 3,340 participants, and the debate lasted 5,443 seconds, our dataset contains a total of 18,179,620 observations (3,340 participants × 5,443 seconds), noting that absence of a reaction is also an observation in our statistical analysis. In order to prevent participants who logged into the app late and/or left early biasing downward our standard errors, we drop 5,205,421 participant-second observations where no clicks had yet registered or where no additional clicks would be registered for that participant, leaving us with just under thirteen million observations.

Issues and Net Positive Engagement

We focus here on the central discussions of the economy, healthcare, and foreign affairs from the first 2012 general debate. Our economy and foreign affairs categories each contain three Policy Agendas topics. From a citizen's point of view, macroeconomics, jobs, and banking discussions are all central to the most pressing question of the first 2012 general debate: the economy. Likewise, discussions of defense, foreign trade, and international affairs all shift viewers' focus from domestic to foreign affairs. We find that some messages were uniformly more resonant with viewers than others, even given variation in messenger and audience. Figure 8.3 displays net positive engagement (*Agree* clicks minus *Disagree* clicks) with each candidate by response topic. This figure shows that both candidates fared best among their base supporters and independents when discussing foreign affairs, although discussing foreign affairs also yielded the worst net results for Obama among Republicans.

One prescriptive interpretation of these results could be that to maximize net positive engagement with independents and their respective bases, both candidates should have emphasized foreign affairs. Yet, Obama's discussion of foreign affairs may have worked in Romney's favor, as foreign affairs was the only topic where Romney surpassed Obama in terms of net positive engagement among independents. Thus, from a heresthetics perspective (Riker, 1996), Romney was advantaged by shifting the agenda toward foreign affairs (though the difference is not statistically significant according to a Welch modified two-sample t-test), whereas Obama held the relative advantage on economic, health, and other topics (all three differences are statistically significant at $p < .05$, two-tail). Some of the prior research on debate responses aggregates reactions across the entire debate and therefore may not differentiate between audience responses to, say, foreign affairs vs. healthcare messages. Our data can show such fine distinctions, and our findings here reveal a tension between a candidate's pursuit of *absolute* net positive engagement and his desire to keep the agenda away from topics where the opponent has a *relative* advantage.

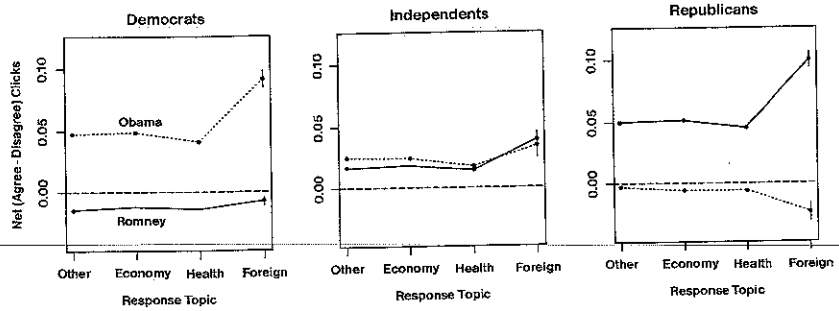

FIGURE 8.3 Net Positive Engagement with Each Candidate by Response Topic.

Audience Priority

Our fine-grained data also allow us to examine how audience characteristics beyond partisanship influence reactions. For example, viewers may respond differently to economic messages based on how strongly they prioritize the economy (Holbrook, Berent, Krosnick, Visser, & Boninger, 2005; Iyengar, Hahn, Krosnick, & Walker, 2008). Examining only candidate statements in response to moderator questions about the economy, we model viewers' responses as a function of candidate agenda building. The results are presented in Table 8.3. Focusing on economic questions in this way allows us to hold constant the content of the moderator's prompt and, thus, to better identify how viewers react to candidates' discussion of economic topics, relative to their use of other topics that are potentially less relevant to the question at hand.

TABLE 8.3 Pooled Cross-Sectional Time Series Logistic Regressions of Audience Reactions on Candidate Agenda-Building Behaviors and Audience Characteristics

	Obama		*Romney*	
	Agree $n = 1739564$	*Disagree* $n = 1739564$	*Agree* $n = 1815663$	*Disagree* $n = 1815663$
Party ID				
Independent	—	—	—	—
Democrat	0.72 (0.077)	−2.03 (0.176)	−1.12 (0.092)	1.45 (0.107)
Republican	−1.35 (0.090)	1.73 (0.164)	1.00 (0.099)	−2.10 (0.138)
Economics Priority	0.00 (0.002)	0.01 (0.004)	0.00 (0.002)	−0.01 (0.002)
Economics Topic	−0.05 (0.012)	−0.01 (0.039)	−0.05 (0.014)	−0.13 (0.013)
Economics Priority × Economics Topic	0.04 (0.014)	0.04 (0.042)	0.05 (0.015)	0.09 (0.015)
Intercept	−4.07 (0.174)	−8.77 (0.417)	−4.99 (0.206)	−4.99 (0.232)
Var(Intercept)	0.78 (0.035)	1.86 (0.058)	1.08 (0.037)	1.27 (0.042)
σ_u	1.48 (0.026)	2.54 (0.074)	1.71 (0.032)	1.89 (0.039)
ρ	0.40 (0.008)	0.66 (0.013)	0.47 (0.009)	0.52 (0.010)
AIC	450125	102489.8	365746.8	252560.3

Note: Cell entries are model estimates (standard errors in parentheses).

ıe model in Table 8.3 is a pooled cross-sectional time series logit, in which the it of analysis is participant-seconds. The response variables (Agree, Disagree) ual one if a participant registered that response over the previous five-second an, zero otherwise, and are restricted to moments following statements by the cal candidate. Our first key explanatory variable is the priority the viewer ached to the economy. The pre-debate survey asked participants to prioritize e economy using a continuous slider ranging from "Not Important" to "Very portant," mapped to a value between 0 and 1. We also interact this viewer eco- omic priority value with the count of seconds that the focal candidate discussed ı economic topic in the preceding five-second span. As a candidate spends more ne on economic topics, and therefore less time discussing others, these count riables increase. Thus, the interaction term tests whether viewers' reactions to conomy-oriented messages were conditioned by the viewers' own economic rioritization.

Demonstrating the importance of individuals' issue priorities, the agreement odels show positive, statistically significant coefficients associated with the inter- ction between viewers' economic priority and the candidate's discussion of the conomy: people's tendency to click *Agree* in response to economic discussion ncreased with their economic prioritization. In contrast, for Obama's disagree- ent model, the coefficient associated with the interaction is small and statistically ndistinguishable from zero, suggesting that his economic discussion may have ffectively drawn issue publics (Converse, 1964) into the debate, producing agree- nent without necessitating disagreement. On the other hand, a significant inter- ction in Romney's disagreement model shows that the greater the priority viewers laced on the economy, the more likely they were to disagree with Romney's esponses about the economy. This finding may suggest viewers attuned to the economy were more likely to react negatively to Romney's comments in the context of the mixed economic climate (Vavreck, 2009) or his personal wealth (Adams, 2012). Further analysis is, of course, needed to adjudicate between these competing explanations. Our analysis illustrates our methodology's potential to yield detailed insight into specific audience reactions, such as how viewers' eco- nomic prioritization conditions receptiveness to economic discussion.

Spinning and Dodging

As a final illustration of the method, we briefly consider *Spin* and *Dodge* responses during the second presidential debate of 2012. The notion of "spin" is closely related to the concept of issue framing—emphasizing or de-emphasizing aspects of complex issues in order to connect with existing cognitive schemas (Scheufele & Tewksbury, 2007, p. 12)—but includes a perception of overt manipulation. "Dodging"—avoiding a question one would rather not answer by answering a different question—can also lead to negative perceptions when it is recognized, though it often goes undetected (Rogers & Norton, 2011).

Figure 8.4 illustrates the kind of data collected by our approach. The spil near 2:10 pertains to Obama's response to a question about Benghazi: "Who w it that denied enhanced security and why?" Partway through his answer, Oban criticized Romney: "While we were still dealing with our diplomats being threa ened, Governor Romney put out a press release, trying to make political poin and that's not how a commander in chief operates." Many participants appeared t interpret this response as dodging the question. Romney provoked a similar set (reactions when answering a question about gun control by shifting to a discussio of schools and the importance of two-parent families.

However, simultaneously collecting multiple kinds of response makes it pos sible to consider audience reactions in more depth. When the candidates wer asked how they differentiate themselves from President Bush and his policies, th *Spin* peak for Obama began to rise at the start of his response, "Well, first of al I think it's important to tell you that we did come in during some tough times .. we had been digging our way out of policies that were misplaced and focuse on the top doing very well and middle class folks not doing well." A *Dodge* spik was seen here, as well (unlabeled, around 1:50 in Figure 8.4), particularly as hi response shifted toward including criticism of Romney. The fact that his answe did include explicit comparison with the previous administration ("We've brough twice as many cases against unfair trading practices than the previous administra tion"), together with the *Spin* responses, suggests that some participant reaction were less about whether Obama's response was avoiding the question, and more about spinning the answer to cast doubt on Romney and place himself in a favor able light. As Boydstun, Glazier, and Pietryka (2013) observe, failing to respond directly to the question during a debate can have strategic advantages, but can also

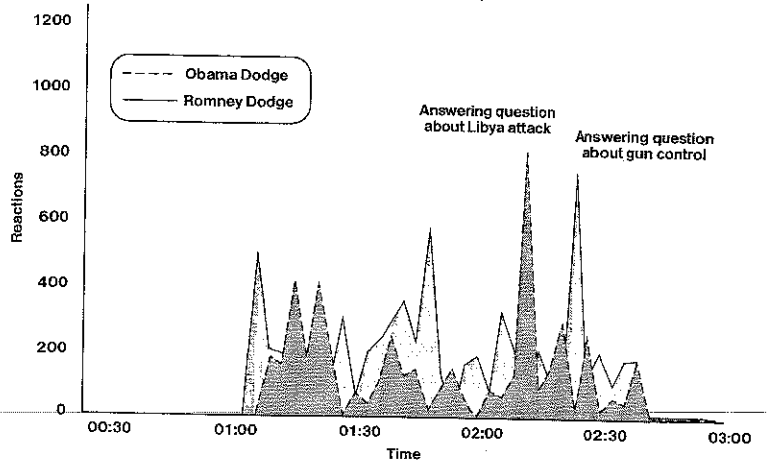

FIGURE 8.4 Counts of Dodge Reactions During the October 16, 2012 Presidential Debate.

incur a cost to the candidate. Collecting real-time, multidimensional data opens the door to finer-grained analysis of candidates' strategic choices and their effects.

Conclusion

We have described an approach to real-time response measurement based on a mobile app that enables collection of real-time data from a large, diverse population reacting at their own initiative in a natural environment. The platform captures the instantaneous and user-initiated aspects of dial testing, while at the same time incorporating a multidimensional push button design (Maier & Faas, 2009) to interpret what is being reacted to and why.

Many extensions to our approach are possible. One restriction of the app as we have described it is that it is limited to sets of participants watching the same event at the same time, since timestamps are used to align reactions to the stimulus. We are currently developing a version of the app that overcomes this limitation and permits the app to be used anywhere, any time, as long as the audio of the stimulus is available. Similar to the popular Shazam app (Wang, 2006), which "listens" to a snippet of music and then identifies what song it is via lookup in a song database, this adapted version of the app uses audio input to keep track of what part of the communicative stimulus the participant is watching at any given moment. This advance removes the limitation to shared-watching events and makes it possible to collect real-time reactions on an individual-by-individual basis.

A second extension we plan to explore is the connection of real-time responses with the social media stream generated during an event (e.g., on Twitter). Our data and Twitter data are in many respects complementary, since the Twitter stream contains vast quantities of difficult-to-interpret, unprompted language data, and our stream is relatively smaller but contains highly interpretable data points from individuals with known attributes (identifiable via sampling, in the entry survey, or both). Well-known computational methods (Wang et al., 2013) should make it possible to align time series based on Twitter responses, which take place with greater and higher-variance delays relative to the stimulus, with real-time reactions from our app, which are more tightly coupled to the stimulus. Text analysis techniques (for example, Blei, Ng, & Jordan, 2003) are then well positioned to identify the concepts and themes in the Twitter stream associated with peaks of response in our reactions data. This strategy illustrates the more general potential of a highly scalable approach to real time responses to bring together theoretical considerations, technical methods, and a rich range of data sources in order to produce not only measurements but causal insights.

Acknowledgments

Portions of this work were supported in part by the University of Arkansas at Little Rock Presidential Studies Center; the University of California, Davis Department

of Political Science; and React Labs LLC. Parts of this chapter are adapted with permission from Boydstun, Feezell, Glazier, Jurka, and Pietryka (2014) and Boydstun, Glazier, Pietryka, and Resnik (2014). The authors gratefully acknowledge the editors of this volume and several anonymous reviewers from *PS: Political Science and Politics* and *Public Opinion Quarterly* for their contributions and commentary on this work, and are also grateful to Jeffrey Korn, Drew Stephens, Chad Yan, Andy Garron, Craig Dye, Julie Schroeder, and the Maryland Technology Development Corporation for discussions and support for these efforts. The technological platform described in this chapter is being commercialized by React Labs LLC; further information is available by writing to the first author or info@reactlabs.com.

References

Abramowitz, A. I. (1978). The impact of a presidential debate on voter rationality. *American Journal of Political Science, 22*(3), 680–690.

Adams, G. (2012, September 20). Romney's wealth in spotlight again after tax probe: New evidence of Republican candidate's low payments follows poor TV ratings. *The Independent*. Retrieved from http://latestupdatednews.org/news/world/americas/mitt-romneys-wealth-in-spotlight-again-after-tax-probe-8101500.html.

Bartels, L. (2009). Priming and persuasion in presidential campaigns. In H. E. Brady & R. Johnston (Eds.), *Capturing campaign effects* (pp. 78–114). Ann Arbor, MI: University of Michigan Press.

Baumgartner, F. R., & Jones, B. D. (2006). Policy agendas project topic codebook (updated by E. Scott Adler and John Wilkerson). Retrieved from http://www.policyagendas.org/page/topic-codebook.

Baumgartner, F. R., & Jones, B. D. (2009). *Agendas and instability in American politics* (2nd ed.). Chicago, IL: University of Chicago Press.

Beaumont, E., Colby, A., Ehrlich, T., & Torney-Purta, J. (2006). Promoting political competence and engagement in college students: An empirical study. *Journal of Political Science Education*, 2(3), 249–270.

Benoit, W. L., Hansen, G. J., & Verser, R. M. (2003). A meta-analysis of the effects of viewing U.S. presidential debates. *Communication Monographs, 70*(4), 335–350.

Blais, A., & Perrella, A. M. L. (2008). Systemic effects of televised candidates' debates. *International Journal of Press/Politics, 13*(4), 451–464.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.

Boydstun, A. E., Feezell, J. T., Glazier, R. A., Jurka, T. P., & Pietryka, M. T. (2014). Colleague crowdsourcing: A method for fostering national student engagement and large-N data collection. *PS: Political Science & Politics, 47*(4), 829–834. doi:10.1017/S1049096514001127.

Boydstun, A. E., Glazier, R. A., & Pietryka, M. T. (2013). Playing to the crowd: Agenda control in presidential debates. *Political Communication, 30*(2), 254–277. doi:10.1080/10584609.2012.737423.

Boydstun, A. E., Glazier, R. A., Pietryka, M. T., & Resnik, P. (2014). Real-time reactions to a 2012 presidential debate: A method for understanding which messages matter. *Public Opinion Quarterly, 78*(S1), 330–343. doi:10.1093/poq/nfu007.

Brubaker, J., & Hanson, G. (2009). The effect of Fox News and CNN's postdebate commentator analysis on viewers' perceptions of presidential candidate performance. *Southern Communication Journal, 74*(4), 339–351. http://doi:10.1080/10417940902721763.

Carlin, D. B., & McKinney, M. S. (Eds.). (1994). *The 1992 presidential debates in focus.* New York, NY: Praeger.

Converse, P. (1964). The nature of belief systems in mass politics. In D. E. Apter (Ed.), *Ideology and discontent* (pp. 206–261). New York, NY: Free Press.

Druckman, J. N., & Kam, C. D. (2011). Students as experimental participants: A defense of the "narrow data base." In J. N. Druckman, Green, D. P., Kuklinski, J. H. & Lupia, A. (Eds.), *Handbook of experimental political science* (pp. 41–57). New York, NY: Cambridge University Press.

Erbring, L., Goldenberg, E. N., & Miller, A. H. (1980). Front-page news and real-world cues: A new look at agenda-setting by the media. *American Journal of Political Science, 24*(1), 16–49. Retrieved from http://www.jstor.org/stable/2110923.

Evans, J. R., & Mathur, A. (2005). The value of online surveys. *Internet Research,* 15(2), 195–219.

Findlater, L., Froehlich, J. E., Fattal, K., Wobbrock, J. O., & Dastyar, T. (2013). Age-related differences in performance with touchscreens compared to traditional mouse input. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 343–346. ACM. doi:10.1145/2470654.2470703.

Fleischman, M. & Roy, D. (2008). Grounded language modeling for automatic speech recognition of sports video. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* 121–129.

Frank N. Magid Associates. (2013). *Frank N. Magid Associates Super Bowl study focuses on ads that work.* Unpublished.

Fridkin, K. L., Kenney, P. J., Gershon, S. A., Shafer, K., & Woodall, G. S. (2007). Capturing the power of a campaign event: The 2004 presidential debate in Tempe. *The Journal of Politics, 69*(3), 770–785. Retrieved from http://www.jstor.org/stable/4622579.

Geer, J. G. (1988). The effects of presidential debates on the electorate's preferences for candidates. *American Politics Research, 16*(4), 486–501.

Giglietto, F., & Selva, D. (2014). Second screen and participation: A content analysis on a full season dataset of tweets. *Journal of Communication, 64,* 260–277. doi:10.1111/jcom.12085.

Grimmer, J., & King, G. (2011). General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences, 108*(7), 2643–2650.

Haile, T. (2014, March 9). What you think you know about the web is wrong. *Time.* Retrieved from http://time.com/12933/what-you-think-you-know-about-the-web-is-wrong/

Hillygus, D. S. (2005). The missing link: Exploring the relationship between higher education and political engagement. *Political Behavior, 27*(1), 25–47. doi:10.1007/s11109-005-3075-8.

Hillygus, D. S., & Jackman, S. (2003). Voter decision making in election 2000: Campaign effects, partisan activation, and the Clinton legacy. *American Journal of Political Science, 47*(4), 583–596.

Holbrook, A. L., Berent, M. K., Krosnick, J. A., Visser, P. S., & Boninger, D. S. (2005). Attitude importance and the accumulation of attitude-relevant knowledge in memory. *Journal of Personality and Social Psychology, 88*(5), 749–769.

Holbrook, T. M. (1999). Political learning from presidential debates. *Political Behavior, 21*(1), 67–89.

Howe, J. (2006a). The rise of crowdsourcing. *Wired Magazine, 14*(06), 1–5. doi:10.1086/599595.

Howe, J. (2006b). Crowdsourcing: A definition. [Blog post, June 2, 2006]. Retrieved from http://crowdsourcing.typepad.com/.

Iyengar, S., Hahn, K. S., Krosnick, J. A., & Walker, J. (2008). Selective exposure to campaign communication: The role of anticipated agreement and issue public membership. *The Journal of Politics, 70*(1), 186–200.

Iyengar, S., & Valentino, N. A. (2000). Who says what? Source credibility as a mediator for campaign advertising. In A. Lupia, M. D. McCubbins, & S. L. Popkin (Eds.), *Elements of reason: Cognition, choice, and the bounds of rationality*. Cambridge, UK: Cambridge University Press.

Jones, B. D., & Baumgartner, F. R. (2005). *The politics of attention: How government prioritizes problems*. Chicago, IL: University Of Chicago Press.

Julien, I., & Resnik, P. (2014). Aligning real-time opinion poll responses with an expectation-maximization algorithm. Unpublished manuscript.

Karpf, D. (2010). Online political mobilization from the advocacy group's perspective: Looking beyond clicktivism. *Policy & Internet, 2*(4), 7–41.

Kingdon, J. W. (1995). *Agendas, alternatives, and public policies* (2nd ed.). New York, NY: HarperCollins.

Kraus, S., & Smith, R. G. (1977). Issues and images. In S. Kraus (Ed.), *The great debates: Kennedy vs. Nixon, 1960*. Bloomington, IN: Indiana University Press.

Lanoue, D. J. (1992). One that made a difference: Cognitive consistency, political knowledge, and the 1980 presidential debate. *Public Opinion Quarterly, 56*(2), 168–184. doi:10.1086/269309.

Lemert, J. B. (1993). Do televised presidential debates help inform voters? *Journal of Broadcasting & Electronic Media, 37*(1), 83–94.

Lucas, C., Nielsen, R., Roberts, M., Stewart, B., Storer, A., & Tingley, D. (2015). Computer assisted text analysis for comparative politics. *Political Analysis, 23*(2), 254–277.

Maier, J. & Fass, T. (2009). Measuring spontaneous reactions to media messages the traditional way: Uncovering political information processing with push button devices. In J. Maier, M. Maier, M. Maurer, C. Reinemann, & V. Meyer (Eds.), *Real-time response measurement in the social sciences: Methodological perspectives and applications* (pp. 15–26). Frankfurt am Main, Germany: Peter Lang.

Maurer, M. & Reinemann, R. (2009). RTR measurement in the social sciences: Applications, benefits, and some open questions. In J. Maier, M. Maier, M. Maurer, C. Reinemann, & V. Meyer (Eds.), *Real-time response measurement in the social sciences: Methodological perspectives and applications* (pp. 1–14). Frankfurt am Main, Germany: Peter Lang.

McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly, 36*(2), 176–187.

Nielsen. (2014). Living social: How second screens are helping TV make fans. Retrieved from http://www.nielsen.com/us/en/insights/news/2014/living-social-how-second-screens-are-helping-tv-make-fans.html.

Pfau, M. (2003). *The changing nature of presidential debate influence in the new age of mass media communication*. Paper presented at the 9th Annual Conference on Presidential Rhetoric, College Station, TX: Texas A&M University.

Podjarny, G. (2012, June 26). Quantifying the mobile difference. [Slides]. Retrieved from http://www.slideshare.net/guypod/the-mobile-difference-in-numbers/22.

Prior, M. (2012). Who watches presidential debates? Measurement problems in campaign effects research. *Public Opinion Quarterly, 76*(2), 350–363. doi:10.1093/poq/nfs019.

Quinn, A. J., & Bederson, B. B. (2011). Human computation: A survey and taxonomy of a growing field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1403–1412. doi:10.1145/1978942.1979148.

Ramanathan, S., McGill, A., Phillips, J., Schill, D. & Kirk, R. (2010). Are political opinions contagious? An investigation on the effects of seating position and prior attitudes on moment-to-moment evaluations during the presidential debates. In M. C. Campbell, J. Inman, & R. Pieters, *Advances in Consumer Research, 37*, 242–245. Duluth, MN: Association for Consumer Research.

Riker, W. H. (1996). *The strategy of rhetoric: Campaigning for the American Constitution*. New Haven, CT: Yale University Press.

Rogers, T., & Norton, M. I. (2011). The artful dodger: Answering the wrong question the right way. *Journal of Experimental Psychology. Applied, 17*(2), 139–147. doi:10.1037/a0023439.

Roy, D. (2005). Grounding words in perception and action: Computational insights. *Trends in Cognitive Sciences, 9*(8), 389–396.

Schattschneider, E. E. (1960). *The semi-sovereign people: A realist's view of democracy in America*. New York, NY: Holt, Rinehart, and Winston.

Scheufele, D. A. & Tewksbury, D. (2007). Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of Communication 57*, 9–20.

Schroeder, A. (2008). *Presidential debates: Fifty years of high-risk TV*. New York, NY: Columbia University Press.

Sears, A., & Shneiderman, B. (1991). High precision touchscreens: Design strategies and comparisons with a mouse. *International Journal of Man-Machine Studies, 34*(4), 593–613.

Shaw, D. R. (1999). A study of presidential campaign event effects from 1952 to 1992. *The Journal of Politics, 61*(2), 387–422.

Shulman, S. W. (2009). The case against mass e-mails: Perverse incentives and low quality public participation in US federal rulemaking. *Policy & Internet, 1*(1), 23–53.

Sigelman, L., & Sigelman, C. K. (1984). Judgments of the Carter-Reagan debate: The eyes of the beholders. *Public Opinion Quarterly, 48*(3), 624–628.

Steeper, F. (1978). Public responses to Gerald Ford's statement on Eastern Europe in the second debate. In G. F. Bishop, R. G. Meadow, & M. Jackson-Beeck (Eds.), *The Presidential debates: Media, electoral, and policy perspectives* (pp. 81–101). New York, NY: Praeger.

Trent, J. S., & Friedenberg, R. V. (2008). *Political campaign communication: Principles and practices*. Lanham, MD: Rowman & Littlefield.

Tsfati, Y. (2003). Debating the debate. *The International Journal of Press/Politics, 8*(3), 70–86. doi:10.1177/1081180x03008003005.

Vavreck, L. (2009). *The message matters: The economy and presidential campaigns*. Princeton, NJ: Princeton University Press.

Wang, A. (2006). The Shazam music recognition service. *Communications of the ACM, 49*(8): 44–48. doi:10.1145/1145287.1145312.

Wang, X., Abdullah, M., Ding, H., Trajcevski, G., Scheuermann, P. & Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery, 26*(2): 275–309. doi:10.1007/s10618-012-0250-5.

White, M. (2010, August 12). Clicktivism is ruining leftist activism. Retrieved from http://www.theguardian.com/commentisfree/2010/aug/12/clicktivism-ruining-leftist-activism.