Electoral Studies 44 (2016) 1–24

Contents lists available at ScienceDirect

Electoral Studies

journal homepage: www.elsevier.com/locate/electstud

A measure of survey mode differences

Jonathan Homola [a], Natalie Jackson [b], Jeff Gill [c, *]

[a] Department of Political Science, Washington University in St. Louis, United States
[b] Huffington Post, United States
[c] Department of Political Science, Division of Biostatistics, Department of Surgery, Washington University in St. Louis, United States

ARTICLE INFO

Article history:
Received 1 March 2016
Received in revised form
28 June 2016
Accepted 30 June 2016
Available online 5 July 2016

ABSTRACT

We evaluate the effects of different survey modes on respondents' patterns of answers using an entropy measure of variability. While *measures of centrality* show little differences between face-to-face and Internet surveys, we find strong patterns of *distributional differences* between these modes where Internet responses tend towards more diffuse positions due to lack of personal contact during the process and the social forces provided by that format. We introduce an entropy measure of dispersion for survey responses and illustrate its utility with election data from 2012. Our results provide clear evidence that mode matters in modern survey research, and we make recommendations for interpreting results from different modes.

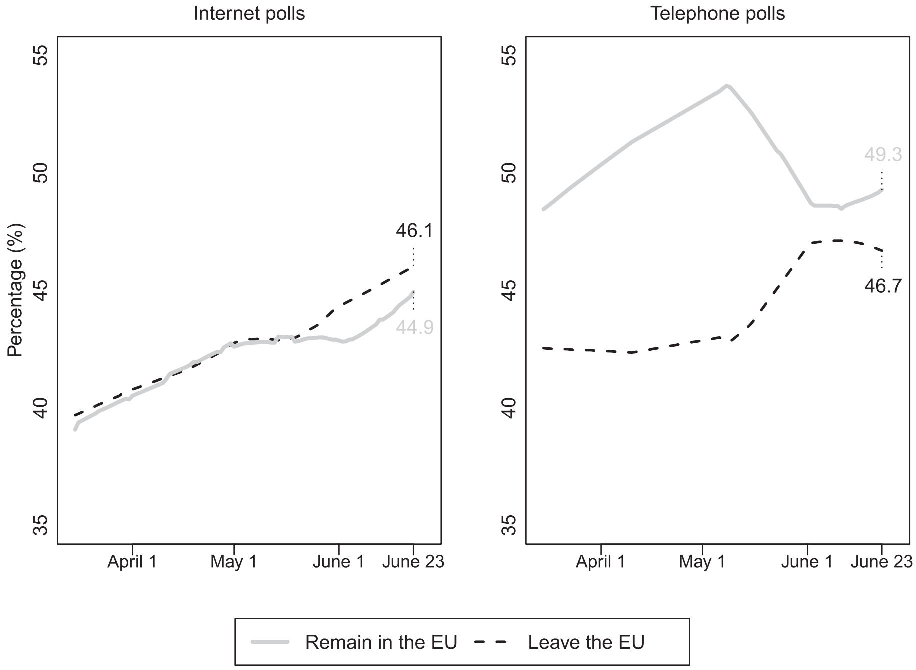
1. Introduction

Potential survey respondents are becoming more expensive and more difficult to contact with traditional telephone and face-to-face methods, forcing researchers to look toward new methods of finding and contacting sampled members of the population. Surveys conducted over the Internet have become a popular solution for obtaining survey responses faster and cheaper than via traditional telephone or face-to-face methods (Brick, 2011; Groves, 2011), to the point that expenditures on market research Internet survey usage increased from essentially zero to slightly less than 2.5 billion dollars from 1996 to 2011 in the U.S. alone (Inside Research, 2011). There is anecdotal and scholarly evidence that changing modes influences how respondents think about and react to survey instruments, causing measurement error that hurts comparability of surveys across modes and time (Ansolabehere and Schaffner, 2014; Atkeson et al., 2014; Ye et al., 2011; Bowyer and Rogowski forthcoming, Yeager et al., 2011). Such research has identified specific survey items for which responses do (and do not) seem to vary by mode (Christian, 2007; Christian et al., 2008, 2009; Dillman and Christian, 2005; Dillman et al., 2009), but provides little guidance on how measurement error due to a change in survey mode affects overall *distributions*, rather than individual categories, of responses for individual items and the instrument as a whole.

Given what is known about the differences between online surveys and other modes, we ask whether there is measurement error in the form of a critical distributional difference in aggregate survey responses. And if there is error, whether this difference has the potential to change our interpretation of the question responses. Here we develop an *entropy* (information) measure of variability to answer this exact question. Entropy is a commonly used measure in physics and communication studies to convey levels of information in a given system or message. It turns out that this measure is well-suited to describing measurement variability in categorical questions where measures of variance that assume continuous data, such as the standard deviation, are not appropriate. Therefore it can be used to find differences in how respondents react to *the exact same question* in surveys conducted using two different modes. We illustrate the power of this measure using data from the 2012 American National Election Study (ANES) where the survey was administered with over 500 questions asked identically via face-to-face and web modes across a total of 5914 respondents. Our results show that across a range of question types and circumstances mode effects are not reflected in *measures of centrality*, but manifest in greater *variability*. The entropy measure of dispersion tends to be higher with Internet responses than face-to-face interviews, indicating that sample respondents are more comfortable expressing diffuse views in a more private setting. Our new tool not only provides a direct quantification of measurement error between survey modes, it provides a means for future researchers to investigate specific reasons why respondents change

* Corresponding author.
E-mail addresses: homola@wustl.edu (J. Homola), nmjb09@gmail.com (N. Jackson), jgill@wustl.edu (J. Gill).

http://dx.doi.org/10.1016/j.electstud.2016.06.010
0261-3794/© 2016 Elsevier Ltd. All rights reserved.



Note: The two panels illustrate public opinion trends in advance of the United Kingdom European Union membership referendum that took place on June 23, 2016. The data come from the Huffington Post Pollster archive. The trend lines are based on loess regressions for the period from March 15 to June 23. The actual referendum resulted in an overall vote to leave the EU by 51.9% to 48.1%.

Fig. 1. "Brexit" poll averages by survey mode.

their values across different modes.

Why is this important? Survey research scholars and practitioners for the most part repeat their analyses many times over a career often with very similarly formatted instruments. For instance, the administrators of the American National Election Study are careful to maintain question wording over time for comparability of effects between elections. As the use of Internet polling increases, and it is not hard to envision some point in the future where its use will be exclusive, it is important to know how question mode changes survey results. Consider that the ANES has been asking the feeling thermometer (scaled 0–100) for "warmness" towards Democratic Party presidential candidates since 1968. A diligent researcher using the 2012 Internet mode version only (or a future researcher without the choice) might puzzle about the noticeably greater variability with more extreme evaluations apparent in both directions. Such a person is likely to try and explain this phenomenon by asking whether it is because: the candidate was African-American, the candidate was subject to four years of vitriolic criticism as a first term president, the candidate needed to deal with the biggest economic crisis in the United States since the Great Depression, or some other factor. Whereas we show that in the case of the 2012 ANES feeling thermometer for the Democratic presidential candidate (and other questions) the distributional difference was due to the survey mode only. This is not only important on its own in terms of understanding the 2012 election, it is critical for research designs that compare the same questions over different elections. Making such multiple-time comparisons without accounting for the mode effect naturally leads to the conclusion that the electorate has become more diffuse and more divided, which may be true, but if part of the reason is simply how the survey is conducted then such substantive conclusions will be exaggerated.

The influence of mode effects is not just a concern for academics. Media pollsters have demographic and attitudinal question formats that they prefer and maintain over time, and mode changes can affect what those questions show. For example, in polling on the recent United Kingdom European Union membership referendum, Internet polls and telephone polls indicated opposite trends up until the day of the actual referendum. As illustrated by Fig. 1, the Huffington Post Pollster averages showed a trend toward "remaining" in the EU in the telephone polls by a margin of nearly three percentage points. At the same time, Internet polls favored "leave" by just over one percent. Overall, the polls favored "remain," which meant that in the wake of the "leave" outcome much of the narrative was that the polls were wrong. However, the reality was more nuanced – polls were both right and wrong depending on which mode they used.

In this work we review the total survey error framework for describing and categorizing uncertainty in survey research, focus on measurement error due to mode of contact as a key part of this total, and provide a method for measuring differences between such modes. Our entropy measure of response variation reveals aggregate differences in how questions activate responses. We demonstrate the efficacy of this approach with an application to the 2012 American National Election study which used multiple survey modes with the same questions.[1]

2. Mode as part of the total survey error framework

The current academic literature does not completely identify how systematic measurement error affects Internet versus

[1] In addition, Appendix A presents a series of Monte Carlo simulations that further assess and highlight the efficacy of the approach and also help to measure its sensitivity to different types of observed distributions. We also include all of the R code for this experiment in Appendix B.

traditional surveying modes in terms of data quality, comparative dynamics, and statistical inference. We do know that asking the *same questions* with the *same answer options* might not yield exactly the same results between a web survey and a face-to-face or telephone survey. Previous literature has identified a large number of potential reasons for such a *mode effect* (Christian, 2007; Christian et al., 2008, 2009; Dillman and Christian, 2005; Dillman et al., 2009).

Mode effects, and variability between modes, are key aspects of total survey error: the conceptual framework that fully describes the statistical error that comes from the process of conducting a sample survey and calculating subsequent model-based estimates to describe a population (Biemer, 2010). The total survey error framework consists of five primary sources of error: sampling, coverage, non-response, measurement, and postsurvey (Weisberg, 2005; Groves and Lyberg, 2010). Survey mode has the potential to affect each of these sources, and is therefore a critical part of considering total survey error.

The structure of these constituents of total survey error is diagrammed in Fig. 2. Each source of error has many possible causes, which are not shown in the figure, but the components are serial in their occurrence. The arrows therefore only indicate the order of concern, not causal paths. Coverage error is preeminent in the process since it comes first and defines not only the subsequent sampling process, but also the personal and statistical components. The next three essentially occur in parallel as the researcher defines the sampling process and instrument structure. Finally postsurvey error contributes the last component. Note also the additive nature of these five sources of error; the resulting survey error is a *total* in a way that the name implies. See Groves et al. (2011) for a comprehensive description of these components and their constituent effects.

The literature has developed ways to deal with most types of error across different survey modes. Coverage and sampling differences between web surveys and traditional modes are typically attributed to the opt-in nature of most web surveys. Opt-in means respondents have volunteered to take surveys and often systematically differ from people contacted by telephone or in person who did not previously sign up to receive calls or visits (Couper, 2000; Groves and Lyberg, 2010; Hays et al., 2015; Baker et al., 2013; Chang and Krosnick, 2010). It becomes more difficult to measure sampling error when samples are not randomly drawn, but some online surveys still calculate and report sampling error in similar ways to probability samples (Rivers and Bailey, 2009). Cases that fall outside of the sampling frame due to the method of communication in recruitment contribute to the coverage error: the population accessible by telephone or in person is different from the population with access to the Internet, and especially from the Internet users that sign up to take surveys (Alvarez et al., 2003).

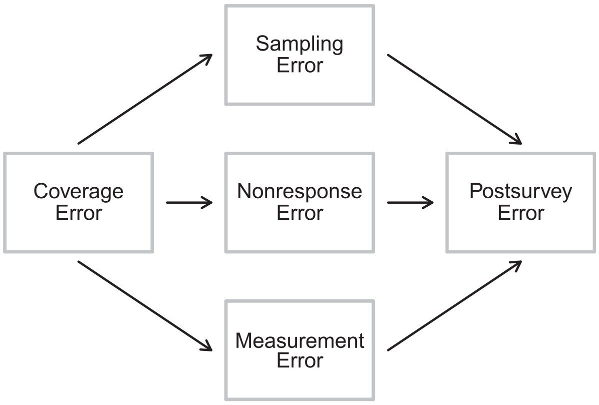


Fig. 2. Total Survey Error Diagrammed. Note: The two panels illustrate public opinion trends in advance of the United Kingdom European Union membership referendum that took place on June 23, 2016. The data come from the Huffington Post Pollster archive. The trend lines are based on loess regressions for the period from March 15 to June 23. The actual referendum resulted in an overall vote to leave the EU by 51.9%–48.1%.

Nonresponse and cooperation have been considered as well (Atkeson et al., 2014; Millar and Dillman, 2012). Slowinski (1988) presents a “hierarchy” of nonresponse error: *noncontact error* where the surveyor fails to contact a unit in the contact list, *unit nonresponse error* where the sampled unit fails to respond completely, and *item nonresponse error* from incomplete survey responses. Some studies have found similar cooperation rates between modes, but with demographic differences (e.g., Nagelhout et al., 2010); while other studies have found dramatic differences in response rates across modes (e.g., Couper, 2001). In the telephone or face-to-face modes, interviewers can sometimes persuade reluctant participants to take the survey, decreasing unit nonresponse (Chang and Krosnick, 2010; West et al., 2013). Interviewers can ask respondents follow-up questions, known as *probes*, and not immediately accept the answer when a respondent says they “don’t know” how to answer a question to reduce this error. Dillman et al. (2014) recently compare Internet, phone, mail, and mixed strategies to show tradeoffs and how design decisions should consider the strengths and weakness of each alternative (see also Dillman and Christian, 2005; McNabb, 2013; Weisberg, 2005).

2.1. Measurement error across survey modes

What is missing in the literature is a way to quantify how much measurement error can be attributed to the difference in mode *that is not due to questionnaire design or question wording*. Even if sampling and coverage are held constant, nonresponse errors are similar, and postsurvey treatment is identical, there will still be a gap in measurement error between web and traditional survey modes because the web respondent is doing the survey privately, whereas traditional methods require an interviewer, adding a social component to the survey. The effect that interviewers can have directly on measurement error comes from several sources. During the survey interviewers force the respondents to pay attention to the survey, particularly when it is conducted face-to-face, and respondents complete the entire questionnaire all at once. Without an interviewer online respondents could be easily distracted by other tasks they need to do on the Internet or even search for information about questions on the survey, which damages the prospect of measuring Internet respondents’ unpolluted opinions. There is some evidence that social desirability bias is higher in interviewer-administered surveys than in Internet surveys, particularly on questions about controversial issues such as racism and discrimination (Davis et al., 2010). And, even though it is seldom discussed, there is always the possibility that interviewers can directly influence data by suggesting answers or even falsifying answers (Bredl et al., 2000). Both of these issues would increase measurement error for the telephone or face-to-face surveys.

Measurement error differences also stem from how the survey is presented in each mode. On the telephone or in person interviewers can read a list of options, leaving out “don’t know” or other non-answer responses, and still record a “don’t know” if the respondent insists on this as their answer. Online surveys cannot record anything that is not provided on-screen for the respondent, so “don’t know” or any other non-answer responses have to be either listed or eliminated completely. Listing “don’t know” results in higher proportions of respondents choosing that option (DeRouvray and Couper, 2002). There is also evidence that respondents choose answers on a scale differently online and over the phone (Christian, 2007; Christian et al., 2008, 2009; Dillman et al., 2009; Ye et al., 2011).

Survey researchers know that there are systematic differences in responses based on these sources of error, but the main way in which the error has been measured is to simply show the differences in responses question-by-question. These studies have sometimes shown substantial discrepancies between the modes (e.g., Yeager et al., 2011), and sometimes they show few differences across modes (e.g., Ansolabehere and Schaffner, 2014). However, there is currently no principled statistical means of comparing variance differences across alternative categorical survey modes. Understanding survey error in this context is important since most survey researchers rely upon data collected from others, and misunderstanding collection procedures, including survey mode, may lead to further downstream errors by such researchers that are avoidable. In the next section we introduce a useful dataset for this purpose along with a measure that avoids incorrectly applying standard variability measures to low-granularity ordinal survey measures. By comparing the value of this entropy measure for each mode we can get an idea of how much measurement error there is between the data collected using each mode.

3. Empirical data and entropy measure

In this section we introduce a measure of variability in discrete survey responses based on entropy (Shannon, 1948) and apply it to a recent academic survey. Since the bulk of survey responses are on discrete measures such as Likert scales and feeling thermometers, the standard statistical variance is inappropriate. The variance ($\text{Var}(X) = \frac{1}{n-1}\sum(X_i - \overline{X})^2$), and typical alternatives such as the median average deviation ($\text{MAD}(X)=\text{median}(|X_i-\text{median}(X)|)$) assume interval measured data, but are generally robust to ordinal data with roughly equally spaced distances. However, such statistics are *not* robust to ordinal data with *unequal* spacing because $\overline{X}$ and median(X) are not accurate measures of centrality in this context, and importantly, the summed distances of $X_i - \overline{X}$ and $X_i-\text{median}(X)$ do not add comparable units to create a coherent statistic.[2] One of the hallmarks of political science data, and survey data on politics in particular, is that these categories are *not* equally spaced (Johns, 2005; Evans et al., 2001; Green and Palmquist, 1994). For example, extremely liberal and extremely conservative responses to an ideological question are much further from the adjacent categories on the actual underlying metric distance than those next to the modal point from the central underlying point (Gill, 2005). Abascal and Rada (2014) also recently explain how standard variance measures are inappropriate with 0 to 10-point question formats specifically. The feeling thermometer questions in the American National Election Study (and elsewhere) are ordinal with 101 categories and would seem to be only a minor violation of the interval measure assumption, however a scan of the raw data for any of these questions for any year of the survey reveals few “non-round” numbers like 63, 91, 17, 28, etc., suggesting that the associated cognitive process is ordinal with much less granularity. Therefore a statistically appropriate measure of variability is needed to compare the dispersion of mode effects.

3.1. ANES 2012 survey data

To analyze the potential consequences of mode effects on the uncertainty that surrounds public opinion data, we examine the American National Election Studies 2012 Time Series Study. The ANES 2012 study is the 29th installment in a longstanding series of election studies that go back to 1948. The 2012 edition differs from its predecessors significantly, and lends itself exceptionally well to our analysis because it is the first ANES study that implements a dual-mode design by incorporating a traditional ANES face-to-face sample as well as a separate sample interviewed on the Internet. Both samples were independently drawn and data collection was conducted independently in the two modes as well.

There were slight differences in timing between the two sample collection approaches in 2012, but these timing differences are unlikely to be important in terms of electoral information. The pre-election face-to-face (F2F) interviews were conducted between September 8 and November 6 (election day), while the Internet (Inet) portion of the survey was fielded later between October 11 and November 6. The post-election interviews were conducted between November 7 and January 24, 2013 for face-to-face, and the Internet samples were collected between November 29 and January 24, 2013. The face-to-face part of the study employed an address-based sample with in-person recruitment and interviews. The Internet survey relied on a primarily address-based sample and mail recruitment, as well as some random digit dialing (RDD) based telephone sample and recruitment and Internet interviews for all respondents. All Internet study participants were members of the KnowledgePanel, a panel of regular survey participants administered by GfK. For more technical information on the study see the documentation in ANES (2012).

The data set for the ANES 2012 study is available online[3] and contains a total of 2240 variables for 5914 respondents, where 2054 respondents come from the face-to-face survey and the remaining 3860 respondents were interviewed online. To accomplish our goal of directly comparing the two different modes we need to process and subset the data to achieve direct comparability. We want to focus our analysis directly on Likert scales and feeling thermometers (plus similar percentage scales) reflecting political opinions/attitudes and behavioral intentions. Specifically,

- We removed all variables that have no variation and take on one of the missing data codes that the ANES coding scheme employs (between −1 and −9, see ANES, 2014: p.36–37) in the face-to-face (k=449) and subsequently Internet (k=452) part of the study.
- For the remaining 1339 variables, we recoded all entries that have a missing data code to NA, and consequently drop all variables that have more than two thirds of missing data in either the face-to-face (k=303) or Internet (k=39) portion of the survey.
- We removed all variables that provide background survey related information, such as the random ordering of answer options, or meta-data for the respondents, such as the names or gender of their current Congressional representatives (k=141). These deletions include additional variables that provide analytical information such as weights or strata characteristics.
- We then ignored all demographic variables (age, income, religion, etc.) as well as variables reflecting past (political) behavior, factual information about the respondents (whether they were contacted during the campaign, whether they are registered to vote or are gun owners, media use, etc.), and overall knowledge questions (k=342).[4]

[2] Relatedly, it is also inappropriate to use the standard Pearson’s Product Moment Correlation Coefficient for comparing two nominal or ordinal variables. Despite this fact, it is common to use the standard correlation when the tetrachoric (nominal) or polychoric (ordinal) forms are correct.

[3] http://www.electionstudies.org/studypages/anes_timeseries_2012/anes_timeseries_2012.htm.

[4] We exclude these variables to focus our main analysis on items that capture opinions and attitudes. However, the results reported below are robust to the inclusion of past behavior, factual information about the respondents, and knowledge questions. The robustness check results are reported in Appendix E.

Table 1
Comparing medians for face-to-face and internet modes.

	BSM	BSE	Difference	*t*-statistic	p-value
2 and 3 point items ($k = 232$)					
F2F	1.536	0.008	−0.038	−48.536	<0.001
Internet	1.574	0.009			
4 and 5 point items ($k = 155$)					
F2F	2.619	0.012	−0.155	−107.859	<0.001
Internet	2.774	0.012			
7 to 11 point items ($k = 78$)					
F2F	3.984	0.027	0.037	9.875	<0.001
Internet	3.947	0.023			
101 point items ($k = 49$)					
F2F	60.452	0.202	3.425	84.208	<0.001
Internet	57.026	0.196			

Note: the two-sided *t*-test presents the result of a test for significant differences in bootstrap means between Internet and face-to-face. This is not part of the entropy measure analysis.

This process ultimately leaves us with 465 Likert scale items and 49 0–100 scaled items (either feeling thermometers or questions that ask respondents to reply with a percentage) that capture respondents' political opinions and behavioral intentions, and do so in exactly the same way in both the face-to-face and Internet part of the study. Those K=514 questions contain a total of 99 items that are used to build composite indices, and 48 items that are already such indices, where values depend on a set of other variables in turn.[5] We collect these 465 Likert scale items into three groups such that 232 of them have only two or three answer categories (for example, “approve” and “disapprove”; or “favor”, “oppose”, and “neither favor nor oppose”), 155 have between four and five answer categories, and 78 have between seven and eleven categories. Modifying this categorization slightly did not have any noticeable effect on our eventual results (a robustness check). All of the R code for processing the data as just described is provided in Appendix D.

An obvious question is how the two modes differ in *measures of centrality* for the ANES 2012 study. Since these data are measured on fairly non-granular ordinal scales (except for those varying from 0 to 100), a regular mean is not an appropriate statistic. As noted above, we categorize the questions into four groups: 2 and 3 point items, 4 and 5 point items, 7 to 11 point items, 101 point items. For each question in each group we then obtain a median response to collect these medians within their groups. This allows us to bootstrap sample (e.g. with replacement) 5000 within each group to calculate a bootstrap mean (BSM) and bootstrap standard error (BSE) of the medians. These values are presented in Table 1 along with a test of differences. Bootstrapping the standard errors is necessary since medians do not have closed form expressions for the standard error. Note that while each difference is statistically reliable, the effect sizes are very small relative to the level of measurement for the corresponding questions: 0.038 on a scale of 1–3, 0.155 on a scale of 1–5, 0.037 on a scale of 1–11, and 3.425 on a scale of 0–100. So in the next section we will begin focusing instead on the differences in variability across survey modes.

3.2. Entropy statistic for response variability

For all K=514 remaining questions, we now calculate the *discrete* entropy measure given by Shannon (1948) and implemented by Gill (2005) in a political science context. Shannon's entropy has been applied in a wide variety of fields and is considered a measure of communicated information, although the definition and use of “information” differs widely (Ayres, 1994; Jaynes, 1957; Ruelle, 1991; Tribus, 1961, 1986).[6] More specifically, the entropy measure, H, for a given k-category discrete variable $f(X)=[p_1,p_2,\ldots,p_k]$ is calculated by:

$$H = -\sum p_i \ln(p_i), \quad \sum p_i = 1, \tag{1}$$

where the p_i are the empirically observed counts (normalized into proportions) for each possible answer category of the respective variable, and “ln” denotes the natural logarithm. Although this is a simple function, it has deep theoretical and practical implications. Shannon (1948, Appendix B) showed that this form is the only function that satisfies the following three critical properties:

- H is continuous in the discrete measure $\{p_1,p_2,\ldots,p_n\}$.
- H is at its maximum and is monotonically increasing with n if the p_i are uniformly distributed.
- If a set of alternatives can be reformulated as multiple, temporally consecutive sets of alternatives, then the first H equals the weighted sum of the following consecutive H values: $H(p_1,p_2,p_3) = H(p_1,1-p_1)+(1-p_1)H(p_2,p_3)$.

These properties show that the Shannon entropy function is mathematically ideal for relative information comparisons, which is well-documented in other literatures (Aczél and Daróczy, 1975; Van Campenhout and Cover, 1981; Cover and Joy, 1991; Bevensee, 1993; Ryu, 1993). In our case we will use this function to show different distributions of survey responses for the same question, thus highlighting variability in information supplied by respondents depending on the mode of sampling. Since the resulting quantity is interval measured over the entire real line, means and variances are now appropriate statistical summaries, thus allowing direct and intuitive relative comparisons.

The Shannon entropy form of uncertainty uses measures of uncertainty that are produced by the format in terms of the number of categorical responses and the full distribution of responses from the survey instrument. Furthermore, the entropy approach makes absolutely no assumptions about the distribution of the variability of uncertainty (Shannon, 1948; Jaynes, 1968, 1982). For our purposes we use entropy to describe the variability of a set of survey responses for a categorical question. On one extreme is the case where each of the categories had an equal number of responses. This is then a uniform distribution of responses, which gives the highest possible value for H in (1). On the other extreme, every one of the respondents picks the same category. This observed degenerate distribution gives the lowest possible value of H. Any values in between show relative distance on this scale and thus a continuous measure of discrete variables that can be compared in different settings for the same survey question.

It is important to note that the created value is a “measure”, and not a “statistic” since it does not possess an underlying distribution or associated standard error (Casella and Berger, 2002, Chapter 1). This means that there are no “tail-values” or thresholds for

[5] One classical example is the two-step party ID question, where a respondent is first asked whether they see themselves as Democrat, Republican or Independent, before a second question asks how strongly they feel about that view (for partisans), or whether they are leaning towards either party (for independents). Both answers together then allow for the construction of a 7-point party ID variable. Another example is first asking respondents whether they approve or disapprove of the way the president is handling the economy, and then asking whether their feeling is strong or not strong. In this case, both answers together then allow for the construction of a 4-point approval variable. Our analyses below are based on all K=514 items. However, we also ran every part of the analysis excluding the 99 items that are used to build composite indices and the results were virtually identical.

[6] Other uses in political science include John and Jennings (2010) who analyze the fragmentation of political attention in Queen's Speeches, and Greene (forthcoming) who studies the diversity of party policy rhetoric.

Table 2
Entropy descriptive statistics.

	Mean	Variance	Min	Max	*t*-statistic	p-value
2 and 3 point items ($k = 232$)						
F2F	0.761	0.045	0.176	1.097	1.334	0.183
Internet	0.787	0.044	0.168	1.098		
4 and 5 point items ($k = 155$)						
F2F	1.327	0.033	0.727	1.598	−0.445	0.657
Internet	1.317	0.040	0.518	1.608		
7 to 11 point items ($k = 78$)						
F2F	1.726	0.058	0.993	2.321	−0.259	0.796
Internet	1.716	0.057	1.167	2.334		
101 point items ($k = 49$)						
F2F	2.064	0.039	1.393	2.326	7.795	<0.001
Internet	2.470	0.093	1.230	2.792		

Note: 2-sided *t*-test presents the result of a test for significant differences in means between Internet and face-to-face.

standard hypothesis testing. Instead, its value is strictly as a comparative tool for *relative* distance much in the way that the Akaike Information Criterion (AIC) is used to compare fit between alternative models. However, once a set of entropy measures across questions have been calculated they are a numerical description of that survey instrument that is interval measured. Thus means, variances, etc. can be calculated *on that set of associated entropy scores* to make standard descriptive claims (see Table 2).

As an example of entropy usage, consider the ANES 2012 study question `iran_nukdip`, which asks respondents: “To try to prevent Iran from developing nuclear weapons, would you favor, oppose, or neither favor nor oppose direct diplomatic talks between the United States and Iran to try to resolve the situation?” Within the face-to-face sample, 1551 respondents answered “favor”, 146 answered “oppose”, and 171 respondents indicated that they would “neither favor nor oppose” such direct diplomatic talks between the United States and Iran. Based on these 1868 responses, we get the following vector of proportions: {0.830,0.078,0.092}. Consequently, the entropy score from the measure above is then:

$$H = -[0.830 \cdot \ln(0.830) + 0.078 \cdot \ln(0.078) + 0.092 \cdot \ln(0.092)] = 0.573. \tag{2}$$

For the same question in the Internet sample, the vector of proportions is {0.683,0.084,0.234} and the corresponding entropy measure is H=0.808. In other words, the Internet sample exhibits considerably greater entropy than the face-to-face sample on that specific question, and is therefore more varied and uncertain according to this measure. In Appendix A, we illustrate the efficacy of the entropy approach by implementing a series of Monte Carlo simulations, which also help to determine the sensitivity of our suggested measure to different types of observed distributions.

It is also important to note that entropy applied in this manner provides a measure of dispersion or spread only. So for instance, if we had one dichotomous response variable according to $\{p_1,p_2\}=\{0.1,0.9\}$ it would have an entropy value of $H=-[0.1 \cdot \ln(0.1)+0.9 \cdot \ln(0.9)]=$ 0.325. Now suppose we have another dichotomous response variable according to $\{p_1,p_2\}=\{0.9,0.1\}$. This is a very different response structure in substantive terms, yet it has the same entropy value: $H=-[0.9 \cdot \ln(0.9)+0.1 \cdot \ln(0.1)]=0.325$. This is because, although the structure of responses is flipped around, the categorical measure of variability is exactly the same. Observe that this is no more limiting than the use of variances for interval measured data: the variance of $\mathbf{x}=(6,5,1,5,3)$ is 4 as is the variance of $\mathbf{y}=(-6,-5,-1,-5,-3)$.

3.3. *Results from the application to the ANES 2012 study*

In the following, we present an analysis of all K=514 Likert scale items in the American National Election Studies 2012 Time Series Study that ask the exact same question in both survey modes and offer respondents the exact same answer categories. By comparing entropy measures for identical survey instruments across both survey modes, we examine the potential consequences of mode effects on the uncertainty that surrounds the data. The difference in the variability between modes is a quantification of measurement error.

3.3.1. *Aggregate entropy differences by categorization*

First we aggregate all questions according to our four types and analyze mode differences by entropy. Table 2 provides descriptive statistics for the entropy measures across all K=514 items included in this analysis and compares those entropy measures across the two survey modes and across the different types of response scales. The table includes the mean and variance for each category of entropy measures, as well as their minimum and maximum value and the result of a *t*-test comparing the differences in means for a given subgroup of items. For very short and very long scales, the Internet sample shows a greater mean entropy measure than the face-to-face survey, while for 4 to 11 point Likert scales, the mean entropy in the face-to-face sample is slightly greater. Apart from the feeling thermometers and percentage questions, the variance of the entropy measure is relatively similar for both survey modes. Moreover, while the minimum observed entropy value is greater in the face-to-face sample for some categories, and greater in the Internet sample for others, the maximum entropy value is always greater in the Internet sample. Perhaps most importantly, only for feeling thermometers and percentage questions do we observe a difference in entropy means between the two modes that is statistically reliable at conventional levels.

Fig. 3 combines a set of histograms that further inform us about the distribution of the entropy measure for different types of scales and the two different survey modes (with measures for the face-to-face sample in the left column, and measures for the Internet sample in the right column). As we would expect (and was indicted by the descriptive statistics), with increasing scale length, the entropy measure tends to increase in both survey modes. We can also see that whereas for short to medium length scales Internet and face-to-face entropy scores do not seem to differ significantly, the distributions look considerably different for the 101 point items, where the Internet sample exhibits noticeably heavier tails in the plotted distribution.

Fig. 4 presents similar information in a slightly different way. Here, we look at each of the K=514 survey items by category and subtract the face-to-face based entropy measure from the Internet based measure. Subsequently, we plotted histograms for these differences and overlaid these with the respective empirical density estimates. If both survey modes were to capture a given question with the same degree of uncertainty, we would expect the differences presented here to be (very close to) zero. For the first three graphs there are observable but modest differences. The last graph showing the density of differences for feeling thermometers and percentage questions provides a distinctly different pattern. While the right-hand side of the distribution still looks relatively normal, its mean and mode are now clearly positive, indicating that the entropy measure is larger in the Internet sample for almost all of these survey items.

All of this aggregate entropy analysis implies modest but noticeable difference between the two survey modes: 0–100 scaled items easily show statistically reliable mean differences and tail differences, and there are modal differences for all categories. This supports the general idea of using the entropy measure to detect overall differences in a set of questions across two modes, but it does not yet indicate how different, or how important, a difference

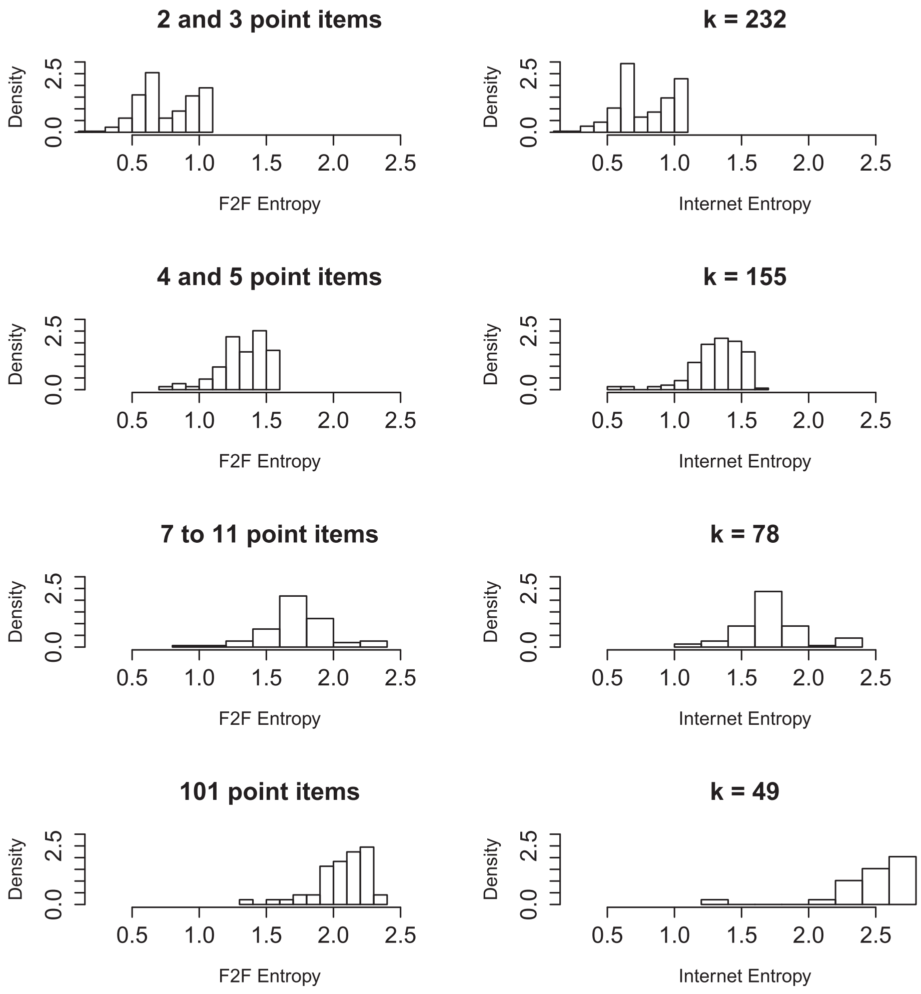


Fig. 3. Entropy histograms.

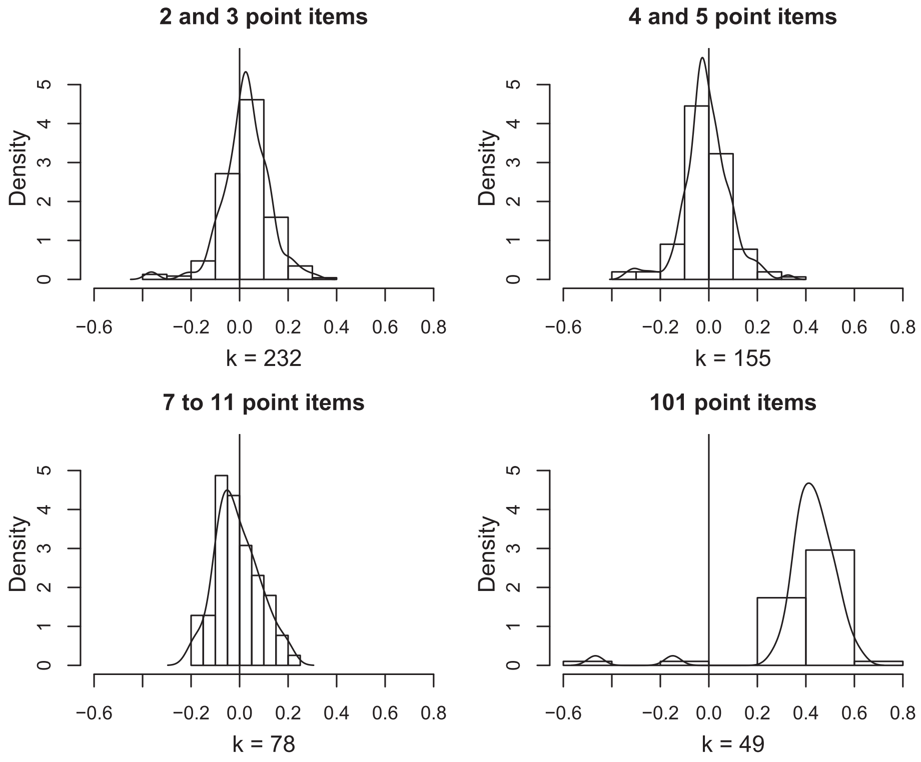


Fig. 4. Distribution of entropy differences (Internet - F2F).

Table 3
Greatest face-to-face entropies.

Variable name	Entropy Value	Variable description
2 and 3 point items ($k = 232$)		
cses_econ	1.0869	State of Economy
campfin_banads	1.0906	Ban Corporate/Union Ads
ineqinc_ineqreduc	1.0925	Gov't Reducing Income Inequality
econ_ecpast	1.0946	National Economy: Better/Worse
econ_unpast	1.0966	Unemployment: Better/Worse
4 and 5 point items ($k = 155$)		
resent_deserve	1.5742	Blacks: Gotten Less Than Deserved
cses_govtact	1.5768	Gov't Reducing Income Inequality
resent_try	1.5773	Blacks: Must Try Harder
ecperil_payhlthcst	1.5864	Able to Pay Health Care
egal_worryless	1.5976	Worry Less About Equality
7 to 11 point items ($k = 78$)		
cses_dptylike	2.1893	Democratic Party Like (0–10)
cses_rptylike	2.2786	Republican Party Like (0–10)
cses_rpclike	2.3022	Republican Pres Cand Like (0–10)
cses_dptyleft	2.3203	Left-Right Democratic Party (0–10)
cses_rptyleft	2.3208	Left-Right Republican Party (0–10)
101 point items ($k = 49$)		
ftpo_dvpc	2.2782	FT: Democratic Vice Presidential Candidate
ftgr_unions	2.2814	FT: Unions
ftgr_fedgov	2.2842	FT: Federal Government
ftpo_rpc	2.3098	FT: Republican Presidential Candidate
ftcasi_illegal	2.3260	FT: Illegal Immigrants

Note: The table presents the 5 items with the highest value for (entropy_f2f) in each category.

can be across a single question.

3.3.2. *Entropy differences by specific survey questions*

Table 3 shows the five items with the highest values for the entropy measure in the face-to-face sample, again separated by our four question type categories. In line with the previous analyses, we can see that as scale length increases, the maximum values for the entropy measure also tend to increase (although the differences between the items for 7 to 11 point scales and the 101 point scales is relatively small). This is a feature of the mathematical structure of (1). Substantively, the items for relatively short scales focus primarily on the economy and attitudes towards blacks, while the items for the longer scales focus more on party and candidate attitudes as well as illegal immigrants.

Table 4 presents the same information for the Internet portion of the survey. While the items with the highest entropy measure differ slightly for the shorter scales, they are almost identical to the items with the highest values in the face-to-face sample for the longer scales. Notice first that the 101 point items show considerably greater values for the entropy measure in the Internet sample. Second, when comparing Tables 3 and 4 side by side, the entropy value for a given entry is always greater in Table 4.

In Table 5 we combine the entropy measures for both samples, by subtracting the entropy value for a given question in the face-to-face sample from the value in the Internet sample. Table 5 then presents the three variables with the highest and lowest value for

Table 4
Greatest internet entropies.

Variable name	Entropy Value	Variable description
2 and 3 point items ($k = 232$)		
mip_prob2pty	1.0958	Best Party to Handle MIP #2
econ_unpast	1.0964	Unemployment: Better/Worse
iran_nuksite	1.0968	Bombing Iran's Nuclear Sites
cses_econ	1.0980	State of Economy
econ_ecpast	1.0983	National Economy: Better/Worse
4 and 5 point items ($k = 155$)		
ctrait_dpccare	1.5873	Dem Cand: Cares About People Like Me
ecblame_pres	1.5874	Blame President for Economy
ctrait_rpclead	1.5886	Rep Cand: Strong Leadership
ctrait_dpcmoral	1.5956	Dem Cand: Is Moral
ctrait_rpcmoral	1.6078	Rep Cand: Is Moral
7 to 11 point items ($k = 78$)		
cses_rptyleft	2.2585	Left-Right Republican Party (0–10)
cses_dpclike	2.2847	Democratic Pres Cand Like (0–10)
cses_rptylike	2.3200	Republican Party Like (0–10)
cses_dptylike	2.3311	Democratic Party Like (0–10)
cses_rpclike	2.3343	Republican Pres Cand Like (0–10)
101 point items ($k = 49$)		
ftpo_rvpc	2.7123	FT: Republican Vice Presidential Candidate
ftgr_unions	2.7272	FT: Unions
ftpo_pres	2.7480	FT: Democratic Presidential Candidate
ftpo_rpc	2.7495	FT: Republican Presidential Candidate
ftpo_dvpc	2.7919	FT: Democratic Vice Presidential Candidate

Note: The table presents the 5 items with the highest value for (entropy_inet) in each category.

Table 5
Greatest entropy differences.

Variable Name	Entropy Difference	Variable Description
	2 and 3 point items ($k = 232$)	
auth_consid	-0.3780	Important for Child: Considerate or Well-Behaved
finance_finpast	-0.3649	Better/Worse Off Than Year Ago
interest_wherevote	-0.3478	Know Where to Vote
tea_suppln	0.2685	Tea Party: Leaning Towards Support/Opposition
preswin_dutyst	0.2919	Voting as Duty: Feeling Strength
fedspend_schools	0.3220	Public Schools: More or Less Spending
	4 and 5 point items ($k = 155$)	
finance_finpast_x	-0.3403	Better/Worse Off Than Year Ago (5 point scale)
likelypct_howlikvt1	-0.3101	Likelihood of Voting
trustgov_trustgstd	-0.3003	Trust Gov't in Washington
cses_diffvote	0.2124	Vote Makes a Difference
gayrt_discstd_x	0.2377	Favor Laws Against Gays/Lesbian Job Discrim
egal_equal	0.3262	Provide Equal Opportunities
	7 to 11 point items ($k = 78$)	
wpres_gdbd_x	-0.1941	Good/Bad: Woman Pres
women_bond_x	-0.1859	Working Mother's Bond with Child
libcpre_dpc	-0.1845	Lib/Con Scale: Dem Pres Cand (1-7)
abort_sex_x	0.1890	Legal Abortion to Select Child Gender
budget_deficit_x	0.1962	Favor Reducing Budget Deficit
scourt_remove_x	0.2032	Possibility to Remove Sup Court Judges
	101 point items ($k = 48$)	
pctlikely_whatpct2	-0.4675	Percent Chance of Voting (Group 2)
likelypct_whatpct1	-0.1487	Percent Chance of Voting (Group 1)
ftpo_dpcsp	0.5432	FT: Spouse of Democratic Presidential Candidate
ftgr_military	0.5703	FT: Military
ftgr_working	0.5777	FT: Working Class People
ftpo_pres	0.6220	FT: Democratic Presidential Candidate

Note: The table presents the 3 items with the highest and the 3 items with lowest value for (entropy_inet − entropy_f2f) in each category. Negative values indicate more entropy for the face-to-face mode, and positive values indicate more entropy for the Internet mode.

that difference in each question category, where negative values indicate a higher entropy measure in the face-to-face sample, and positive values indicate higher measures in the Internet sample. These are listed in order that radiates outward from zero indicated by the dashed lines. The 101 point items had only two extreme items favoring face-to-face responses so the dashed line moves upward from the other cases.

For lower levels of measurement, 2 and 3 points items plus 4 and 5 points items, the extreme cases in Table 5 are bigger for the face-to-face survey mode, whereas this switches for the more granular measures. This suggests that when Internet respondents have more "room" to provide less centralized (perhaps more extreme) responses they tend to do so more often than face-to-face respondents, producing more variability for specific questions. Even though the less granular measurement categories tend to show greater negative differences (favoring more diffusion in face-to-face responses), these differences are smaller than the differences in the more granular measurements categories. For the top entropy measure cases this is clear:

- *2 and 3 point items*: −0.3780 versus 0.3220,
- *4 and 5 point items*: −0.3403 versus 0.3262,
- *7 to 11 point items*: −0.1941 versus 0.2032,
- *101 point items*: −0.4675 versus 0.6220.

Obviously the 101 point scale gives a greater range of entropy by mathematical construct, but it is important to look at the substantive political difference in these values.

Among 2 and 3 point items, auth_consid (Important for Child: Considerate or Well-Behaved) is the variable for which (entropy_f2f − entropy_inet) takes on the greatest negative value, whereas fedspend_schools (Public Schools: More or Less Spending) is the variable for which (entropy_inet - entropy_f2f) takes on the greatest positive value. The first question seems easy to respond to an interviewer with varying values given its relatively insensitive nature, regardless of other political leanings. The second question deals with a politically sensitive question that touches on race, class, and fiscal behavior. So it makes sense that there is more diffusion in the answers for the Internet respondents. For the ordinal question types excluding the 0–100 point items, the Internet entropy measures tend to be greater for controversial issues such as support for the Tea Party, laws against gay/lesbian job discrimination, or abortion.

It is noteworthy that among the 101 point items, only two variables have a higher entropy measure in the face-to-face sample than in the Internet sample. These questions, pctlikely_whatpct2 and likelypct_whatpct1, ask the respondents in the two modes what their percent chance of voting in the upcoming election is, and are notorious for producing higher values in the

Table 6
Logit model results for voting for Romney over Obama.

	Face-to-Face Mode				Internet Mode			
	Estimate	Std. Error	t value	Pr(>|t|)	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−4.43224	0.75929	−5.83733	0.00007	−3.86743	0.47010	−8.22673	0.00001
gender_respondent_x	0.07630	0.19080	0.39989	0.34836	−0.11625	0.14913	−0.77956	0.22588
dem_birthyr	0.00415	0.00528	0.78575	0.22351	0.00303	0.00390	0.77536	0.22517
dem_racecps_black	−1.96061	0.42177	−4.64858	0.00026	−1.45919	0.29549	−4.93821	0.00012
dem_hisp	−0.72328	0.30128	−2.40065	0.01588	−0.44175	0.21503	−2.05442	0.03113
pid_x	0.78939	0.05816	13.57346	0.00001	0.91645	0.04252	21.55245	0.00001
libcpre_self	0.49726	0.10285	4.83485	0.00033	0.40134	0.06481	6.19265	0.00002
interest_attention	−0.22951	0.08541	−2.68720	0.01035	−0.03281	0.06984	−0.46983	0.32370
candrel_dpc1	1.12917	0.20898	5.40327	0.00009	1.27015	0.17649	7.19657	0.00001
ftcasi_illegal	−0.00526	0.00387	−1.36046	0.10010	−0.01218	0.00280	−4.35654	0.00039
egal_worryless	−0.14788	0.06996	−2.11389	0.02752	−0.40761	0.06040	−6.74875	0.00001
	Null deviance: 2827 on 2053 df				Null deviance: 5340.3 on 3859 df			
	Residual deviance: 1335 on 2043 df				Residual deviance: 2271.5 on 3849 df			
	AIC: 1306.9				AIC: 2159.6			

person-to-person context.[7] It is also interesting that two of the highest entropy questions include the words "Democratic" and "Candidate" together but the words "Republican" and "Candidate" do not appear at all in the list. This suggests greater variability of responses and greater polarization around the political positions of Democratic candidates than Republican candidates. Importantly, this does not imply some electoral advantage for Republicans because the feeling thermometers are not tied directly to subsequent vote choice.

Notice here in this analysis that we do not specifically model entropy values as an outcome variable in a regression context. While this is tempting, it assumes that the entropy scores are produced from some *iid* process. Nothing could be further from the truth: in the professional context (academic or media) survey questions are customized by experts to gain insight into a variety of *specific* attitudes and knowledge. While some of these are obviously related, such as income, education, and ideology, the purpose in designing such surveys is not to impose a common underlying data generating process across questions that could be summarized with something like an entropy score and then modeled. For example, we looked at both the dichotomous "Know Where to Vote" variable and the feeling thermometer for "Working Class People." Even though entropy values can be calculated in both cases, it is not appropriate to treat these underlying data generation processes as *iid* and feed them into a regression specification.

3.3.3. Model differences by mode

In this section we highlight the extent of possible differences by mode by fitting the same logit model specification for voting for Romney over Obama using the ANES 2012 data. This modest sized specification mixes standard control variables for vote prediction along with three of the variables in Table 5 that were shown to have large entropy differences based on mode: `candrel_dpc1` (FT: Spouse of Democratic Presidential Candidate), `ftcasi_illegal` (FT: Illegal Immigrants), and `egal_worryless` (Provide Equal Opportunities). We also add these additional covariates: `gender_respondent_x` (0 for male and 1 for female), `dem_birthyr` (recoded to age), `dem_racecps_black` (self-identified as African-American), `dem_hisp` (self-identified as Hispanic/Latino/Latina), `pid_x` (7-point party ID: Democrat to Republican), `libcpre_self` (7 point ideology: liberal to conservative), and `interest_attention` (interest in politics: from 1 = always to 5 = never). Additional details about these variable are available in the ANES (2012) codebook. We used multiple imputation with the `R` package mice to handle missing data (Rubin, 2004). Fortunately, the ANES provides sampling weights separately by mode and these are incorporated into the two model estimation processes in the conventional manner. Both model results are presented in Table 6.

We see that all of the coefficients and standard errors differ between the models on the two sub-samples of the data. Some of these differences are important. The negative coefficient for self-reporting as African-American is statistically reliable in both models and reduces 26% in absolute value going from the face-to-face mode to the Internet mode. This implies some measure of black support for Romney that was reserved from telling the interviewer in person. Similarly, the negative coefficient for self-reporting as an Hispanic reduces 39% in absolute value going from the face-to-face mode to the Internet mode, and is also statistically reliable in both settings. Predictably the coefficient for party ID is positive and statistically reliable in both models, but does differ noticeably. There is a very similar story for ideology. The coefficient for interest in politics changes considerably between the two models: it is −0.22951 and statistically reliable in the face-to-face model but is about one order of magnitude smaller and not statistically reliable in the Internet model. Both coefficients are negative, implying greater interest leads to more support for Obama, but it could be that the face-to-face interviewees feel a social effect wherein they want to look interested to the interviewer. This is consistent with the literature (Davis et al., 2010; O'Muircheartaigh and Campanelli, 1998; Singer et al., 1983). The coefficient for the feeling thermometer for illegal immigrants is unsurprisingly negative, meaning higher values push respondents towards Obama over Romney, but the effect is about twice as strong for the Internet mode versus the face-to-face mode (both estimates are reliable). Seemingly respondents have a stronger link between their feelings on illegal immigrants and their vote choice in private. Lastly, the difference between the coefficient estimates is large for the importance of providing equal opportunities, and both are reliable at conventional levels. The Internet mode coefficient is a little over twice the size as the face-to-face mode coefficient, implying that people are more assertive about the relationship between this variable and vote choice without having to think about the social consequences.

Importantly, the model comparison provided above is an example not a proof of mode differences. It is entirely possible that one could use the same data and produce two relatively indistinguishable model results. But to know the differences, a researcher

[7] Another possible explanation could be the slightly different timing with which the two survey modes were fielded. Face-to-face interviewing started about one month earlier than the Internet interviews. Given the popularity of early voting, this means that a greater proportion of respondents in one mode than the other might have already voted. If this is the case, there would be more respondents that are 100% certain they would vote, which could have affected the result.

would have to enjoy the presence of both modes in the same dataset with the same questions, as we have here. The key point is that we have demonstrated non-trivial mode differences in model results that would certainly lead authors to different substantive conclusions about vote choice (in this case). In addition, it is completely unclear what the substantive results mean when mixing these two subsamples and running the same model specification. Even with reliable coefficients, the mixed standard error is not a logical expression of the contribution to total survey error.

Finally, on a technical note, the standard errors on the coefficients in Table 6 are *not* the same as the entropy values on the data. For example the entropy value for the variable `egal_worryless` in the face-to-face subsample is 1.5976, but the standard error for the associated coefficient in the associated model is 0.06996. This is not to imply that greater data variability does not affect the subsequent model coefficient variability, but these are two very different quantities.

4. Methodological advice

So far we have demonstrated that it is possible to use the exact same question wording across two modes and produce a different distribution of responses, even if measures of centrality are not very different. We also observed that Internet surveys tend to produce wider distributions with more extreme response categories, and that these differences can considerably affect the conclusions that we draw from such data. A deep body of literature on social desirability shows that interviewers have an effect on survey responses (Davis et al., 2010; O'Muircheartaigh and Campanelli, 1998; Singer et al., 1983), thus predicting differences in responses between face-to-face and Internet modes. There is also a sampling issue that could be important. Individuals with extreme opinions may opt-in to more impersonal nodes rather than submit to a face-to-face interview. For our purposes this is a secondary issue for researchers to consider: there is simply evidence that survey responses, when controlling for question wording as we do in our ANES example, tend to be more diffuse with the Internet mode than the face-to-face mode.

Those differences could be partly due to differences between the survey being conducted visually online and aurally in person. Previous research has shown this to be a critical distinction (Christian, 2007; Christian et al., 2008, 2009; Dillman and Christian, 2005; Dillman et al., 2009). However, in the case of the ANES, the in-person survey administration involves some visual elements as well, including showing respondents a visual "feeling thermometer" for the respective questions. Since we found the different feeling thermometers to exhibit some of the largest entropy differences, this suggests that these distributional differences are general measurement effects rather than purely consequences of survey design.

So when presented with multiple modes, which version should a diligent researcher use? The answer depends on the objectives. If the interest lies in understanding reflective responses in the absence of a social context, then the Internet mode may be the preferred choice. However, if the research questions have a social context, the face-to-face mode may be preferred to situate answers in a social environment to better reflect respondent behavior in the presence of others.

As the use of Internet-based instruments becomes more prevalent, the second question is whether researchers should change their interpretation of categorical responses based on the evidence of different distributions. We see little evidence of differences in centrality, which is reassuring and indicates that regression coefficients have essentially the same interpretation in most cases. However, in extreme cases there can be substantial differences in subsequent regression model quality as illustrated by the example in Table 6. An explanatory variable with large entropy provides less statistically reliable information (a greater standard error for the associated coefficient) than an explanatory variable with small entropy in terms of predicting levels of an outcome variable in a regression context. In the context where researchers compare the efficacy of some survey question over repeated studies where the mode changes to Internet, it might even be possible to incorrectly claim that this question no longer contains useful information in the regression sense.

As additional evidence for this claim consider a brief example using a 7-point ordinal question where the distribution of answers is highly concentrated in the middle for the face-to-face mode, proportional responses x_1={0.02,0.02,0.02,0.88,0.02,0.02,0.02}, and highly dispersed towards the endpoints for the Internet mode, proportional responses x_2={0.45,0.02,0.02,0.02,0.02,0.02,0.45}. To make this a simple point, consider a bivariate relationship only between these two explanatory variables individually on an n=200 random outcome variable correlated at ρ=0.5 with (x_1+x_2). Linearly regressing the outcome variable response on each of these variables separately produces regression coefficients and (standard errors) of $\beta_{x_1} = 0.468(0.105)$ for the centralized response and $\beta_{x_2} = 0.195(0.025)$ for the dispersed response. Notice that these are substantially different coefficient estimates, both of which are reliable at standard levels. This is illustrated in Fig. 5, where the jittered categorical variables are shown with the differing regression lines (dark line and points for the dispersed case and light line and points for the centralized response). Notice that the 95% confidence interval for the more centralized case covers the regression line for the dispersed case only in a small region in the middle of the data (the reverse setup shows the same effect). This example assumes equal spacing between the ordinal responses, and relaxing this assumption with a fixed effects (treatment contrast) or random effects (distribution) specification only exacerbates the observed effect by allowing greater differences.

In a regression context researchers simply need to be aware that differences in modal distributions can provide very different regression coefficients. From a modeling perspective this means that the same question wording can provide very different *substantive* information across different modes. Therefore results need to be placed in the context of the survey mode of collection so that readers are fully aware of potential differences. It also bifurcates results in the literature: it is not correct to directly compare specific coefficients in different models or published articles across survey modes because they reflect different statistical contexts. There also exist dispersion measures for categorical explanatory variables (Gilula and Haberman, 1995), and these can be used to measure the impact in regression specifications. However, such measures are model-specific and therefore limited to narrow comparisons.

Another issue that affects the use of the entropy measure described here is that of survey weighting in general. First, it is clear that standard survey weights provided by collectors of data in the traditional fashion are not going to make a more diffusely answered question from the Internet mode less diffuse than the *corresponding* weighted question from the face-to-face mode. Therefore the modeling considerations discussed above are immune from standard survey weight balancing. A more interesting and nuanced question is whether techniques like poststratification (Park et al., 2004) alter the effect of the H measure. To make the discussion as clear as possible suppose we are interested in a binary choice: (p_1,p_2). Now say that θ_ℓ= average response for each cross-classification of state and categorical demographics like sex (2 categories: male, female), race (4 categories: White, Hispanic, African American, other), age (3 categories), and education (3 categories). This setup produces 2×4×3×3=72 cross-classifications of interest, but with 50 states there are now ℓ=3600 cross-

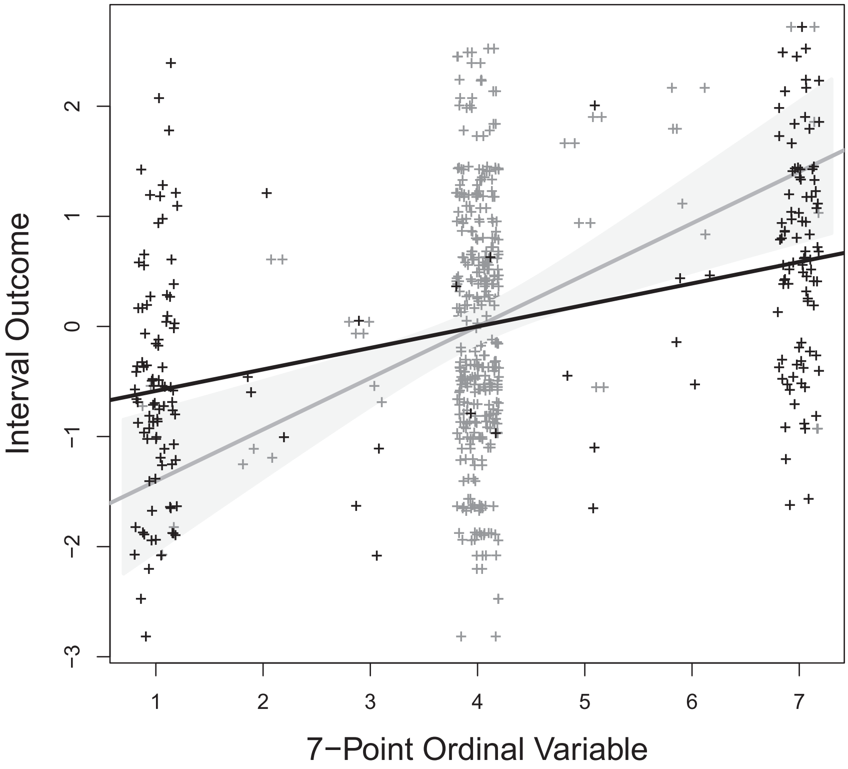


Fig. 5. Simulated regression results.

classifications of interest. For the ℓ th cross-classification we see the binary proportions for this survey question, $(p_{1\ell}, p_{2\ell})$ from our survey sample. Now obtain population information from the census or other sources of national demographics that give N_ℓ as the number of actual people in each of the ℓ categories. So the post-stratification proportions for each of the j=50 states is:

$$\widehat{p}_{1j} = \frac{\sum_{\ell \in j} N_\ell p_{1\ell}}{\sum_{\ell \in j} N_\ell} \quad \widehat{p}_{2j} = \frac{\sum_{\ell \in j} N_\ell p_{2\ell}}{\sum_{\ell \in j} N_\ell}, \tag{3}$$

which helps when categories in some states will be small or empty and we want to weight accordingly. This makes our entropy measure for this dichotomous question in the jth state:

$$H_j = -\sum_{i=1}^{2} \widehat{p}_{ij} \ln\left(\widehat{p}_{ij}\right). \tag{4}$$

So weighting by poststratification or other means is simply adjusting the categorical responses by the desired criteria, which are national demographics in this common case.

The entropy measure as we have described and used it here is restricted to analyzing single questions in isolation. It makes no inferences about surrounding questions and therefore does not detect sequential sources of survey error such as respondent straight-lining (picking the same category for multiple consecutive questions) or recency effects (a question influencing the response to the next question). These cross-question effects are known to be influenced by survey mode. However, it is possible to use the H measure to serially correlate question dispersion. Differences between modes may detect different levels of autocorrelation as a data exploration tool. Furthermore, since H is an interval measurement, common statistics can be used to summarize across a dataset from a given survey mode. Means, medians, variances, etc., are entirely appropriate as rough summaries of typical dispersion.

It is also important to think of the entropy measure contextually with regard to the sample or sub-sample under study. For example, support for welfare is typically skewed such that a large number of individuals are in favor of reducing spending, fewer support the status quo, and even fewer support an increase. However, any dataset that over-sampled African Americans, or subsetting on African Americans would likely produce a near uniform distribution, and thus have a higher entropy value. This is true without any real change in the quality of measurement, the question wording, the order of the question, and so on. Over-sampling or sub-sampling are not necessarily wrong strategies, but the effect on the dispersion of question answers must be considered in context of this decision.

5. Conclusion

In our long collective history of work in survey research the authors have never seen a more dynamic, interesting, and challenging period of technological change. The problems introduced by cell phones, diminishing landlines, lower inclination to participate in face-to-face interviews, and of course the introduction of Internet surveys all provide survey researchers with difficult decisions about the mode of data acquisition. Furthermore, the choice of survey mode affects all five types of survey error (Fig. 2), but the literature has not previously developed a way to quantify measurement error separate from the other types of error, especially as it arises as a consequence of the choice of mode. The entropy measure makes up for a deficiency in the standard set of analytical tools, by measuring variability in survey responses in a way that is free from the defects associated with the application of standard statistics. Our method is extremely easy to implement with or without the R code provided in Appendix C, and so we hope to improve the practice of general mode comparison as well as point out mode differences in a particular instance.

Despite some optimistic views (e.g. Gosling et al., 2004), there is strong evidence of measurement error between face-to-face and Internet surveys; it just turns out to be more subtle than expected. In the ANES 2012 study application, the error is in the distributional differences between the modes rather than centrality differences. We demonstrated that variability in mode effects exists with wider distributions associated with Internet instruments over traditional face-to-face approaches in a manner that could not be done before. This case is sufficiently large and general as to imply similar comparative ability with general academic and media surveys of political and social attitudes.

How does finding measurement error between face-to-face and Internet change subsequent analysis? We know from the analysis that it is wrong to pool face-to-face and Internet polling results for one item into the same regression (or other similar) models without careful explanation – otherwise the interpretation of variability would be assumed to be homogeneous across coefficients. In the mixed mode case, standard errors of specific coefficients could have fundamentally different meanings and statistical interpretations.

Our entropy measure of ordinal question variability is a summary measure for aggregated responses on individual questions that provides convenience. However, our measure cannot tell researchers that smaller entropy for a given question indicates some sort of partisan, general political, or strategic effect. Instead, it simply shows that the surveyed respondents are more or less cohesive and concentrated on a given issue. The entropy measure proposed here does not say *why* mode differences occur in the same way that a tire pressure gauge may indicate that the pressure is low but does not necessarily say that it is due to a nail or neglect. Our goal here is simply to provide a new tool to researchers of public opinion and provide evidence that it detects mode differences of a distributional nature. This is a useful contribution to the toolbox of survey researchers as it provides a statistically coherent measure of dispersion that is easy to interpret and compare.

As far as mode differences go, we assert no normative prescription assigned to higher or lower entropy. Survey-wide, the entropy measures are telling us that the distribution of responses is different between online surveys and face-to-face surveys, even when the sampling and measurement procedures are designed to make the surveys as comparable as possible. This is valuable information about measurement error even without saying which is "better" because a greater number of long-term surveys are increasingly making use of the less expensive web options to deploy their instruments. The entropy measure allows us to see that there are still differences between even the most carefully-constructed online and face-to-face surveys of which all researchers should be aware.

Appendix A. Monte Carlo Assessment of the Entropy Measure

In order to further assess and highlight the efficacy of our entropy measure of mode difference we also implement a series of Monte Carlo simulations to help measure the sensitivity to different types of observed distributions. More specifically, we first created 5000 random draws from a hypothetical 7-point categorical variable (coded 1 to 7) for 1000 respondents. We repeated this process five times assuming different underlying distributions and variances (a uniform distribution, a symmetric distribution with heavy tails, a symmetric distribution with a high centered peak, a left-skewed distribution, and a right-skewed distribution). This leaves us with five simulated data sets, each containing 5000 data points for every respondent. Based on these information, we can then calculate the entropy scores for all 5000 draws across the five data sets. Table 7 presents the five different distributions, their underlying vectors of normalized probabilities from which we simulated the data, and the mean observed entropy score $(\overline{H})$ for each distribution.

Table 7
Monte Carlo Simulation: 7-Point Scale.

Distribution	$p(x=1)$	$p(x=2)$	$p(x=3)$	$p(x=4)$	$p(x=5)$	$p(x=6)$	$p(x=7)$	$\overline{H}$
Uniform	1/7	1/7	1/7	1/7	1/7	1/7	1/7	1.943
Heavy Tails	0.25	0.15	0.05	0.10	0.05	0.15	0.25	1.789
Peaked	0.03	0.07	0.15	0.50	0.15	0.07	0.03	1.495
Left Skew	0.50	0.25	0.10	0.08	0.04	0.02	0.01	1.376
Right Skew	0.01	0.02	0.04	0.08	0.10	0.25	0.50	1.376

Note: $p(x=1)$ to $p(x=7)$ represent the vectors of normalized probabilities for the Monte Carlo simulations. $\overline{H}$ indicates the mean entropy score across all 5000 draws for a given distribution.

As expected, the uniform distribution has the highest mean entropy score. This makes intuitive sense, given that the uniform distribution clearly exhibits the most uncertainty (and provides the least information). Our distribution with "heavy tails" has a similarly high entropy score, but contains more information than the perfectly flat uniform distribution. The third distribution with a clear and centered peak ranks in the middle, while the two heavily skewed distributions have virtually the same mean entropy score and exhibit the least amount of uncertainty while providing the most information among our sample of five different distributions.

Table 8 presents the results from a similar exercise using 5000 random draws from hypothetical 101-point categorical variables (coded 0 to 100). This is done first with a uniform setup drawing directly from the 101 integers, then done by specifying different beta distributions (varying the shape parameters) and normalizing the draws up to 0–100. These specifications are designed to reflect the five generic distributional types presented in Table 7 above. While the first set of Monte Carlo simulations was implemented to illustrate the behavior of the entropy measure in a Likert scale setting, the goal of this second set of simulations is to analyze the entropy scores for different types of feeling thermometers or percentage questions as encountered in the ANES 2012 study.

The second results are very similar to the first and again confirm the intuition behind the entropy measure. The uniform distribution has the highest associated mean entropy score, followed by the distribution with "heavy tails" and the distribution with a relatively centered peak. The left and right skewed distributions have a virtually identical mean entropy measure and rank the lowest, indicating a relatively low degree of uncertainty and a high degree of information.

This simple Monte Carlo analysis illustrates the efficacy of the entropy approach by showing that the measure behaves the way it is intended to. As the amount of information that a given data set or distribution provides increases, the corresponding entropy measure decreases (indicating lower levels of uncertainty), whereas a highly uninformative distribution (such as the uniform distribution as our extreme example) is associated with a high entropy score.

What this means for our application is that evenly-distributed survey responses across the available categories will yield the highest entropy scores – so when we say entropy is higher on a certain type of survey, we are saying that response is more evenly distributed across the possible answers.

Table 8
Monte Carlo Simulation: Feeling Thermometer.

Type	Distribution	R code	$\overline{H}$
Uniform	$\mathscr{U}(0,100)$	c(0:100)	4.564
Heavy Tails	$\mathscr{Be}(0.7,0.7)$	rbeta(1000, 0.7, 0.7)	3.920
Peaked	$\mathscr{Be}(10,10)$	rbeta(1000, 10, 10)	3.708
Left Skew	$\mathscr{Be}(2,7)$	rbeta(1000, 2, 7)	3.656
Right Skew	$\mathscr{Be}(7,2)$	rbeta(1000, 7, 2)	3.656

Note: The Distribution column gives the mathematical notation for the distribution underlying the different simulations, while R code provides the code we used to actually generate the data in R. $\overline{H}$ indicates the mean entropy score across all 5000 draws for a given distribution.

Appendix B. R Code for the Monte Carlo Simulation.

```
set.seed(12435)
## Likert scale
entarray <- array(NA, dim=c(1000, 5000, 5))
for (i in 1:5000){
  entarray[,i,1] <- sample(1:7, size = 1000, replace = TRUE,
                     prob = rep(1/7, 7))
  entarray[,i,2] <- sample(1:7, size = 1000, replace = TRUE,
                     prob = c(.25, .15, .05, .1, .05, .15, .25))
  entarray[,i,3] <- sample(1:7, size = 1000, replace = TRUE,
                     prob = c(.03, .07, .15, .5, .15, .07, .03))
  entarray[,i,4] <- sample(1:7, size = 1000, replace = TRUE,
                     prob = c(.5, .25, .1, .08, .04, .02, .01))
  entarray[,i,5] <- sample(1:7, size = 1000, replace = TRUE,
                     prob = c(.01, .02, .04, .08, .1, .25, .5))
}

ent <- apply(entarray, 3, entropy)
round(apply(ent, 2, mean), 3)
[1] 1.943 1.789 1.495 1.376 1.376
summary(ent)
       V1              V2              V3              V4              V5
 Min.   :1.932   Min.   :1.728   Min.   :1.387   Min.   :1.257   Min.   :1.282
 1st Qu.:1.942   1st Qu.:1.778   1st Qu.:1.476   1st Qu.:1.357   1st Qu.:1.358
 Median :1.943   Median :1.789   Median :1.495   Median :1.376   Median :1.376
 Mean   :1.943   Mean   :1.789   Mean   :1.495   Mean   :1.376   Mean   :1.376
 3rd Qu.:1.944   3rd Qu.:1.800   3rd Qu.:1.514   3rd Qu.:1.395   3rd Qu.:1.395
 Max.   :1.946   Max.   :1.846   Max.   :1.605   Max.   :1.476   Max.   :1.476

## Thermometer
tails <- prop.table(table(round(rbeta(n=100, shape1=.7, shape2=.7)*100, 0)))
peak <- prop.table(table(round(rbeta(n=100, shape1=10, shape2=10)*100, 0)))
left <- prop.table(table(round(rbeta(n=100, shape1=2, shape2=7)*100, 0)))
right <- prop.table(table(round(rbeta(n=100, shape1=7, shape2=2)*100, 0)))

tails_num <- as.numeric(dimnames(tails)[[1]])
```

```
tails_val <- as.numeric(tails)
peak_num <- as.numeric(dimnames(peak)[[1]])
peak_val <- as.numeric(peak)
left_num <- as.numeric(dimnames(left)[[1]])
left_val <- as.numeric(left)
right_num <- as.numeric(dimnames(right)[[1]])
right_val <- as.numeric(right)

entarray_therm <- array(NA, dim=c(1000, 5000, 5))
for (i in 1:5000){
  entarray_therm[,i,1] <- sample(0:100, size = 1000, replace = TRUE,
                           prob = rep(1/101, 101))
  entarray_therm[,i,2] <- sample(tails_num, size = 1000, replace = TRUE,
                           prob = tails_val)
  entarray_therm[,i,3] <- sample(peak_num, size = 1000, replace = TRUE,
                           prob = peak_num)
  entarray_therm[,i,4] <- sample(left_num, size = 1000, replace = TRUE,
                           prob = left_val)
  entarray_therm[,i,5] <- sample(left_num, size = 1000, replace = TRUE,
                           prob = left_val)
}

ent_therm <- apply(entarray_therm, 3, entropy)
round(apply(ent_therm, 2, mean), 3)
[1] 4.564 3.920 3.708 3.656 3.656
summary(ent_therm)
       V1              V2              V3              V4              V5
 Min.   :4.538   Min.   :3.855   Min.   :3.669   Min.   :3.592   Min.   :3.596
 1st Qu.:4.559   1st Qu.:3.910   1st Qu.:3.703   1st Qu.:3.644   1st Qu.:3.644
 Median :4.564   Median :3.921   Median :3.709   Median :3.656   Median :3.656
 Mean   :4.564   Mean   :3.920   Mean   :3.708   Mean   :3.656   Mean   :3.656
 3rd Qu.:4.569   3rd Qu.:3.931   3rd Qu.:3.715   3rd Qu.:3.669   3rd Qu.:3.668
 Max.   :4.586   Max.   :3.975   Max.   :3.739   Max.   :3.719   Max.   :3.719
```

Appendix C. R Code for the Entropy Function.

```
entropy <- function(x){
  ## The helper function "entr" calculates the actual entropy scores
  ## by first computing the normalized probabilities (based on the
  ## empirically observed counts) for a given variable and then applying
  ## the calculation described in the text.
  entr <- function(z){
    probtable <- as.matrix(prop.table(table(z)))
    H <- -sum(apply(probtable, 2, function(y) y*log(y)))
    return(H)
  }
  ## If the input is only one variable (a vector), the function computes
  ## and returns a single numeric value for H as output.
  if(is.vector(x)){
    H <- entr(x)
  }
  ## If the input contains multiple variables (a matrix), the function
  ## computes an entropy score (H) for every variable (column).
  if(is.matrix(x)){
    H <- apply(x, 2, entr)
  }
  return(H)
}

## Vector toy example
vec <- c(1,1,2,3)
## Calculate entropy (H) "by hand"
-(.5*log(.5) + .25*log(.25) + .25*log(.25))
[1] 1.039721
## Compare to function
entropy(vec)
[1] 1.039721

## Matrix toy example (two variables)
mat <- matrix(c(1,1,2,3,1,2,3,4), ncol=2)
mat
     [,1] [,2]
[1,]    1    1
[2,]    1    2
[3,]    2    3
[4,]    3    4
## Calculate entropies (H) "by hand"
-(.5*log(.5) + .25*log(.25) + .25*log(.25))
[1] 1.039721
-(.25*log(.25) * 4)
[1] 1.386294
## Compare to function
entropy(mat)
[1] 1.039721 1.386294
```

Appendix D. R Code Processing the ANES 2012 Data for Analysis.

```
options(stringsAsFactors=F)
library("foreign")
anes <- read.dta("anes_timeseries_2012_stata12.dta",
                 convert.factors=F)

## mode 1 = f2f (k=2054), 2 = internet (k=3860)

f2f <- subset(anes, mode==1)
f2fdrop <- NULL
for (i in 1:2240){
  if (length(rownames(as.matrix(table(f2f[i]))))==1){
    if (rownames(as.matrix(table(f2f[i])))<0){
      f2fdrop <- c(f2fdrop, colnames(f2f)[i])
    }
  }
}

anes <- anes[, !(colnames(anes) %in% f2fdrop)]
## this drops 449 variables that are fully <0 for f2f

inet <- subset(anes, mode==2)
inetdrop <- NULL
for (i in 1:1791){
  if (length(rownames(as.matrix(table(inet[i]))))==1){
    if (rownames(as.matrix(table(inet[i])))<0){
      inetdrop <- c(inetdrop, colnames(inet)[i])
    }
  }
}

anes <- anes[, !(colnames(anes) %in% inetdrop)]
## this drops 452 additional variables that are fully <0 for inet
## leaving 1339 of originally 2240 variables that were asked in both modes

## recode all non-answers to NA
# -1 inapplicable
# -2 missing
# -3 restricted access
# -4 error
# -5 not asked, terminated
# -6 not asked, unit nonresponse
# -7 deleted, partial interview
# -8 don't know
# -9 refused
## Codebook p. 36-37
table(anes[,13])
for (i in 1:1339){
  anes[,i][anes[,i]<0] <- NA
  if (i%%250==0) {print (i)}
}
table(anes[,13])
table(anes[,13], useNA="always")

f2f <- subset(anes, mode==1)
f2fdrop <- NULL
for (i in 1:1339){
  if (sum(is.na(f2f[,i])) > 1370){
    f2fdrop <- c(f2fdrop, colnames(f2f)[i])
  }
}
anes <- anes[, !(colnames(anes) %in% f2fdrop)]
## this drops 303 additional variables that have more than 2/3 NAs for f2f

inet <- subset(anes, mode==2)
inetdrop <- NULL
for (i in 1:1036){
  if (sum(is.na(inet[,i])) > 2574){
    inetdrop <- c(inetdrop, colnames(inet)[i])
  }
}
anes <- anes[, !(colnames(anes) %in% inetdrop)]
## this drops 39 additional variables that have more than 2/3 NAs for inet
## leaving 997 of originally 2240 variables

anes <- anes[,1:856]
## this drops 141 variables that provide information on random ordering of
## answer options and general, factual information about the sample and
## respondents (who is current senator etc)
## leaving 856 of originally 2240 variables

# SUBSET THE DATA
f2f <- f2f[,c(3,531,10,282,277,317,318,316,280,156,100,11,220,781,820,760,
    14,17,18,19,21,24,33,34,43,46,47,50,53,56,59,62,63,64,65,68,69,74,75,78,81,82,85,87,108,
    109,113,118,120,125,131,150,174,190)]
inet <- inet[,c(4,531,10,282,277,317,318,316,280,156,100,11,220,781,820,760,
    14,17,18,19,21,24,33,34,43,46,47,50,53,56,59,62,63,64,65,68,69,74,75,78,81,82,85,87,108,
    109,113,118,120,125,131,150,174,190)]

# Outcome variable: vote for 0 Obama, 1 Romney
f2f$presvote2012_x <- as.numeric(f2f$presvote2012_x) - 1
f2f$presvote2012_x[f2f$presvote2012_x == 4] <- NA
inet$presvote2012_x <- as.numeric(inet$presvote2012_x) - 1
inet$presvote2012_x[inet$presvote2012_x == 4] <- NA

# 0 Male, 1 Female
f2f$gender_respondent_x <- as.numeric(f2f$gender_respondent_x) - 1
inet$gender_respondent_x <- as.numeric(inet$gender_respondent_x) - 1
```

```
# Education ordinal variable
f2f$dem_edu <- as.numeric(f2f$dem_edu)
inet$dem_edu <- as.numeric(inet$dem_edu)

# Age
f2f$dem_birthyr <- 2012 - as.numeric(f2f$dem_birthyr)
inet$dem_birthyr <- 2012 - as.numeric(inet$dem_birthyr)

# Self identify as black
f2f$dem_racecps_black <- as.numeric(f2f$dem_racecps_black)
inet$dem_racecps_black <- as.numeric(inet$dem_racecps_black)

# Self identify as hispanic/latino
f2f$dem_hisp <- 2 - as.numeric(f2f$dem_hisp)
inet$dem_hisp <- 2 - as.numeric(inet$dem_hisp)

# Party ID: Rep to Dem
f2f$pid_x <- as.numeric(f2f$pid_x)
inet$pid_x <- as.numeric(inet$pid_x)

# Ideology Lib to Con
f2f$libcpre_self <- as.numeric(f2f$libcpre_self)
inet$libcpre_self <- as.numeric(inet$libcpre_self)

# Interest in politics 1 always to 5 never
f2f$interest_attention <- as.numeric(f2f$interest_attention)
inet$interest_attention <- as.numeric(inet$interest_attention)

# FT: Illegal Immigrants
f2f$ftcasi_illegal <- as.numeric(f2f$ftcasi_illegal)
inet$ftcasi_illegal <- as.numeric(inet$ftcasi_illegal)

# Worry Less About Equality
f2f$egal_worryless <- as.numeric(f2f$egal_worryless)
inet$egal_worryless <- as.numeric(inet$egal_worryless)

# Survey Weights
f2f$weight_ftf <- as.numeric(f2f$weight_ftf)
inet$weight_web <- as.numeric(inet$weight_web)

# MULTIPLE IMPUTATION
library(mice); source("mice.output.R"); m <- 10
f2f.mice <- mice(f2f,m); inet.mice <- mice(inet,m)
f2f.imp.array <- array(NA,c(dim(f2f),m)); inet.imp.array <- array(NA,c(dim(inet),m))
for (i in 1:m)  f2f.imp.array[,,i] <- as.matrix(complete(f2f.mice,i))
for (i in 1:m)  inet.imp.array[,,i] <- as.matrix(complete(inet.mice,i))
model.names <- c("(Intercept)","gender_respondent_x","dem_birthyr","dem_racecps_black",
    "dem_hisp", "pid_x","libcpre_self","interest_attention","candrel_dpc1",
    "ftcasi_illegal", "egal_worryless")

# LOGIT MODEL FOR VOTING FOR ROMNEY (1) OVER OBAMA (0)
f2f.coef.mat <- matrix(NA,nrow=m,ncol=11); f2f.se.mat <- matrix(NA,nrow=m,ncol=11)
inet.coef.mat <- matrix(NA,nrow=m,ncol=11); inet.se.mat <- matrix(NA,nrow=m,ncol=11)
for (i in 1:m)  {
    current.f2f.df <- data.frame(f2f.imp.array[,,i]); names(current.f2f.df) <- names(f2f)
    current.inet.df <- data.frame(inet.imp.array[,,i]); names(current.inet.df) <- names(inet)
    prez.out.f2f <- glm(presvote2012_x ~ gender_respondent_x + dem_birthyr
        + dem_racecps_black + dem_hisp + pid_x + libcpre_self + interest_attention
        + candrel_dpc + ftcasi_illegal + egal_worryless, family=binomial(link=logit),
        data=current.f2f.df, weights=weight_ftf); xtable(summary(prez.out.f2f))
    prez.out.inet <- glm(presvote2012_x ~ gender_respondent_x + dem_birthyr
        + dem_racecps_black + dem_hisp + pid_x + libcpre_self + interest_attention
        + candrel_dpc + ftcasi_illegal + egal_worryless, family=binomial(link=logit),
        data=current.inet.df, weights=weight_web); xtable(summary(prez.out.inet))
    f2f.coef.mat[i,] <- summary(prez.out.f2f)$coef[,1]
    f2f.se.mat[i,] <- summary(prez.out.f2f)$coef[,2]
    inet.coef.mat[i,] <- summary(prez.out.inet)$coef[,1]
    inet.se.mat[i,] <- summary(prez.out.inet)$coef[,2]
}
f2f.out.table <- mice.output(t(f2f.coef.mat),t(f2f.se.mat),nrow(current.f2f.df))
inet.out.table <- mice.output(t(inet.coef.mat),t(inet.se.mat),nrow(current.inet.df))
dimnames(f2f.out.table)[[1]] <- model.names
dimnames(inet.out.table)[[1]] <- model.names
print(print(f2f.out.table); print(inet.out.table)
```

Appendix E. Robustness Check

The main analysis reported in the text ignores variables reflecting past (political) behavior, factual information about the respondents (whether they were contacted during the campaign, whether they are registered to vote or are gun owners, media use, etc.), and overall knowledge questions in order to focus on items that capture opinions and attitudes. Here, we present additional analyses including past behavior, factual information about the respondents, and knowledge questions as a robustness check.

More specifically, the following tables and figures exactly replicate the tables and figures found in the main analysis. However, this time, the analyses they are based on includes a total of K=734 ANES question items (instead of K=514). This extended data set now contains k=411 2 and 3 point items, k=179 4 and 5 point items, k=95 7 to 27 point items, and still k=49 101 point items.

The findings confirm the results that we discuss in the main analysis. Table 9 shows that there are some differences in *measures of centrality* across the two modes, but that these differences are very small relative to the level of measurement for the corresponding questions. However, as the following tables and figures show, we still find substantial differences in variability across survey modes, where these differences are again the most pronounced for the 101 point items.

Table 9
Comparing Medians for Face-to-Face and Internet Modes

	BSM	BSE	Difference	*t*-statistic	p-value
2 and 3 point items ($k = 411$)					
F2F	1.092	0.011	−0.021	−25.212	<0.001
Internet	1.114	0.011			
4 and 5 point items ($k = 179$)					
F2F	2.537	0.013	−0.189	−129.987	<0.001
Internet	2.727	0.015			
7 to 27 point items ($k = 95$)					
F2F	3.928	0.027	0.041	11.499	<0.001
Internet	3.887	0.021			
101 point items ($k = 49$)					
F2F	60.452	0.202	3.425	84.208	<0.001
Internet	57.026	0.196			

Note: the two-sided *t*-test presents the result of a test for significant differences in bootstrap means between Internet and face-to-face. This is not part of the entropy measure analysis.

Table 10
Entropy Descriptive Statistics.

	Mean	Variance	Min	Max	*t*-statistic	p-value
2 and 3 point items ($k = 411$)						
F2F	0.601	0.082	0.000	1.097	0.716	0.474
Internet	0.615	0.087	0.004	1.098		
4 and 5 point items ($k = 179$)						
F2F	1.290	0.053	0.110	1.598	−1.148	0.252
Internet	1.259	0.078	0.023	1.608		
7 to 27 point items ($k = 95$)						
F2F	1.683	0.084	0.489	2.321	−0.443	0.658
Internet	1.663	0.094	0.347	2.334		
101 point items ($k = 49$)						
F2F	2.064	0.039	1.393	2.326	7.795	<0.001
Internet	2.470	0.093	1.230	2.792		

Note: 2-sided *t*-test presents the result of a test for significant differences in means between Internet and face-to-face.

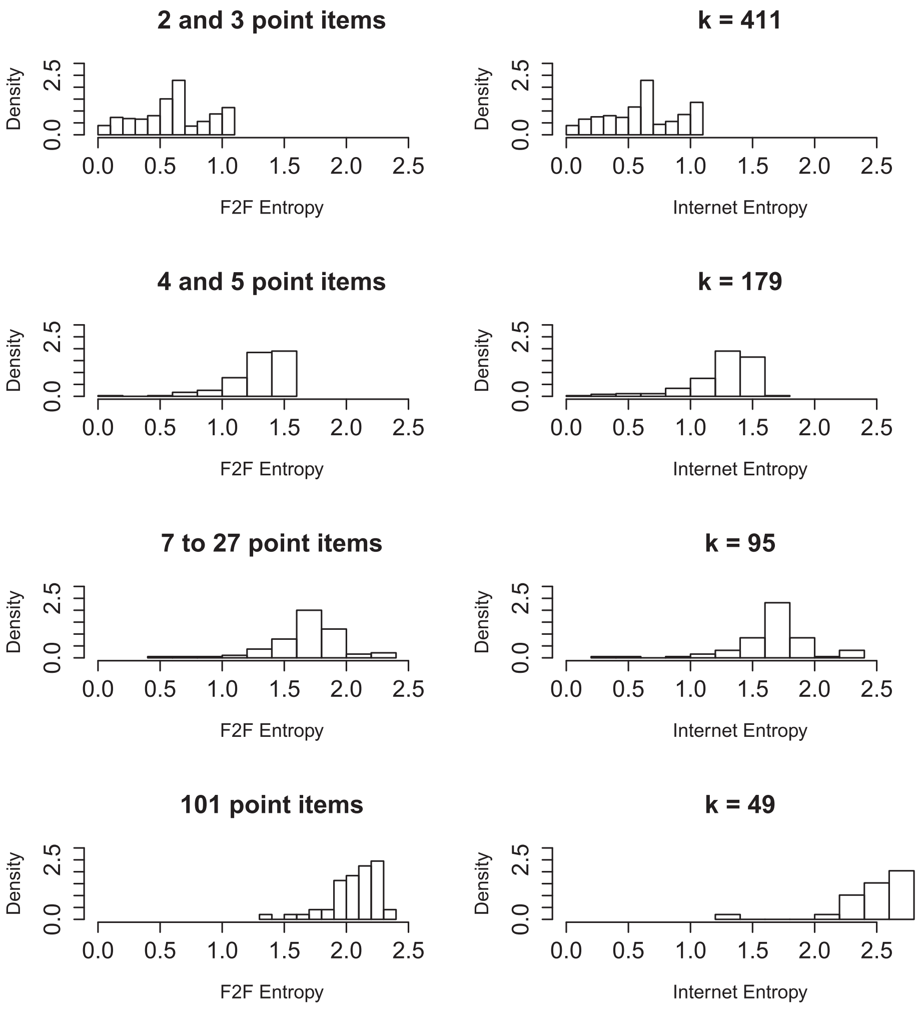


Fig. 6. Entropy Histograms.

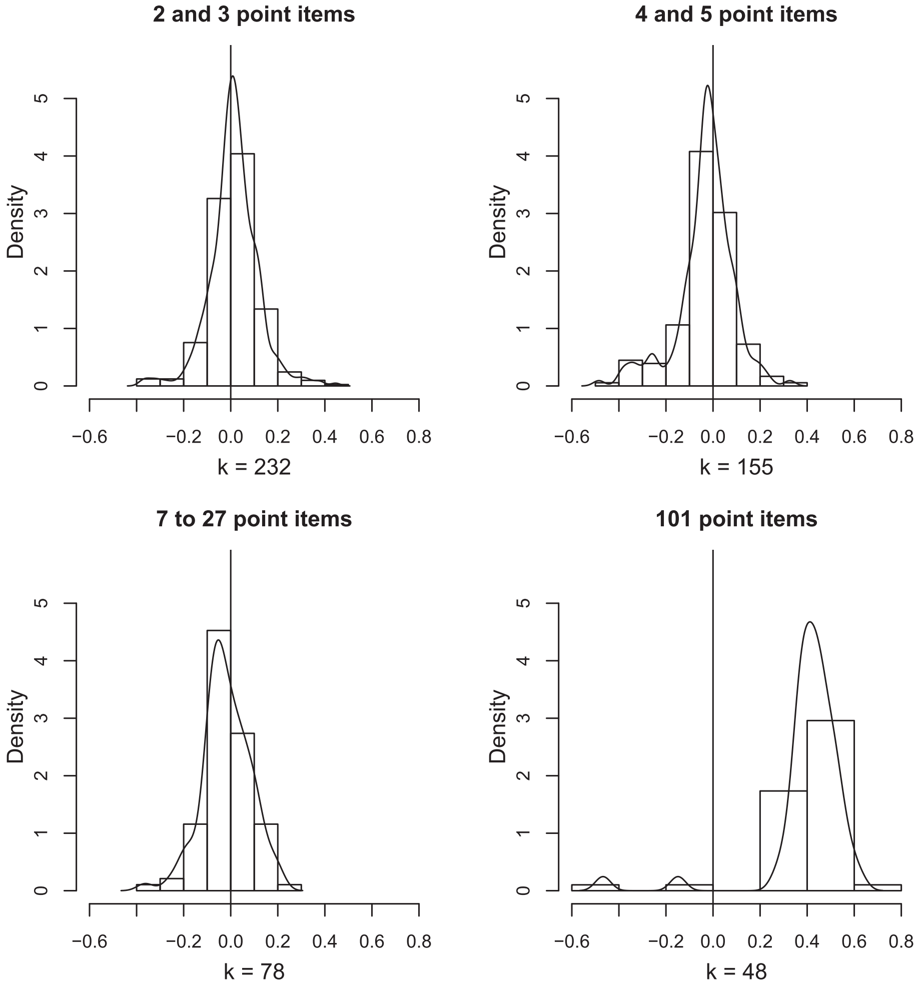


Fig. 7. Distribution of Entropy Differences (Internet - F2F).

Table 11
Greatest Face-to-Face Entropies.

Variable name	Entropy Value	Variable description
2 and 3 point items (*k* = 411)		
mediapo_tvamt	1.0883	Followed Campaigns on TV
campfin_banads	1.0906	Ban Corporate/Union Ads
ineqinc_ineqreduc	1.0925	Gov't Reducing Income Inequality
econ_ecpast	1.0946	National Economy: Better/Worse
econ_unpast	1.0966	Unemployment: Better/Worse
4 and 5 point items (*k* = 179)		
resent_deserve	1.5742	Blacks: Gotten Less Than Deserved
cses_govtact	1.5768	Gov't Reducing Income Inequality
resent_try	1.5773	Blacks: Must Try Harder
ecperil_payhlthcst	1.5864	Able to Pay Health Care
egal_worryless	1.5976	Worry Less About Equality
7 to 27 point items (*k* = 95)		
cses_dptylike	2.1893	Democratic Party Like (0–10)
cses_rptylike	2.2786	Republican Party Like (0–10)
cses_rpclike	2.3022	Republican Pres Cand Like (0–10)
cses_dptyleft	2.3203	Left-Right Democratic Party (0–10)
cses_rptyleft	2.3208	Left-Right Republican Party (0–10)
101 point items (*k* = 49)		
ftpo_dvpc	2.2782	FT: Democratic Vice Presidential Candidate
ftgr_unions	2.2814	FT: Unions
ftgr_fedgov	2.2842	FT: Federal Government
ftpo_rpc	2.3098	FT: Republican Presidential Candidate
ftcasi_illegal	2.3260	FT: Illegal Immigrants

Note: The table presents the 5 items with the highest value for (entropy_f2f) in each category.

Table 12
Greatest Internet Entropies

Variable name	Entropy Value	Variable description
2 and 3 point items ($k = 411$)		
mip_prob2pty	1.0958	Best Party to Handle MIP #2
econ_unpast	1.0964	Unemployment: Better/Worse
iran_nuksite	1.0968	Bombing Iran's Nuclear Sites
cses_econ	1.0980	State of Economy
econ_ecpast	1.0983	National Economy: Better/Worse
4 and 5 point items ($k = 179$)		
ctrait_dpccare	1.5873	Dem Cand: Cares About People Like Me
ecblame_pres	1.5874	Blame President for Economy
ctrait_rpclead	1.5886	Rep Cand: Strong Leadership
ctrait_dpcmoral	1.5956	Dem Cand: Is Moral
ctrait_rpcmoral	1.6078	Rep Cand: Is Moral
7 to 27 point items ($k = 95$)		
cses_rptyleft	2.2585	Left-Right Republican Party (0–10)
cses_dpclike	2.2847	Democratic Pres Cand Like (0–10)
cses_rptylike	2.3200	Republican Party Like (0–10)
cses_dptylike	2.3311	Democratic Party Like (0–10)
cses_rpclike	2.3343	Republican Pres Cand Like (0–10)
101 point items ($k = 49$)		
ftpo_rvpc	2.7123	FT: Republican Vice Presidential Candidate
ftgr_unions	2.7272	FT: Unions
ftpo_pres	2.7480	FT: Democratic Presidential Candidate
ftpo_rpc	2.7495	FT: Republican Presidential Candidate
ftpo_dvpc	2.7919	FT: Democratic Vice Presidential Candidate

Note: The table presents the 5 items with the highest value for (entropy_inet) in each category.

Table 13
Greatest Entropy Differences.

Variable Name	Entropy Difference	Variable Description
	2 and 3 point items ($k = 411$)	
auth_consid	-0.3780	Important for Child: Considerate or Well-Behaved
finance_finpast	-0.3649	Better/Worse Off Than Year Ago
interest_wherevote	-0.3478	Know Where to Vote
medsrc_inetnews_dkrf	0.3650	Regularly Reading Online Newspaper
ofcrec_cj_correct	0.3858	Knowledge Correct: Supreme Ct Chief Justice
medsrc_websites_dkrf	0.4456	Regularly Visiting Website
	4 and 5 point items ($k = 179$)	
wordsum_setl	-0.4850	Wordsum Module L
wordsum_setb	-0.3892	Wordsum Module B
medsrc_tvprog_none	-0.3863	Regularly Watching TV Program
cses_diffvote	0.2124	Vote Makes a Difference
gayrt_discstd_x	0.2377	Favor Laws Against Gays/Lesbian Job Discrim
egal_equal	0.3262	Provide Equal Opportunities
	7 to 27 point items ($k = 95$)	
preknow_prestimes	-0.3633	Presidential Term Limit
preknow_senterm	-0.2621	Seante Term Length
medsrc_socmedia	-0.2276	Weekly Social Media Use
abort_sex_x	0.1890	Legal Abortion to Select Child Gender
budget_deficit_x	0.1962	Favor Reducing Budget Deficit
scourt_remove_x	0.2032	Possibility to Remove Sup Court Judges
	101 point items ($k = 49$)	
pctlikely_whatpct2	-0.4675	Percent Chance of Voting (Group 2)
likelypct_whatpct1	-0.1487	Percent Chance of Voting (Group 1)
ftpo_dpcsp	0.5432	FT: Spouse of Democratic Presidential Candidate
ftgr_military	0.5703	FT: Military
ftgr_working	0.5777	FT: Working Class People
ftpo_pres	0.6220	FT: Democratic Presidential Candidate

Note: The table presents the 3 items with the highest and the 3 items with lowest value for (entropy_inet − entropy_f2f) in each category. Negative values indicate more entropy for the face-to-face mode, and positive values indicate more entropy for the Internet mode.

References

Abascal, E., Rada, V.D.D., 2014. Analysis of 0 to 10-point response scales using factorial methods: a new perspective. Int. J. Soc. Res. Methodol. 17 (5), 569–584.

Aczél, J., Daróczy, Z., 1975. On Measures of Information and Their Characterizations. Academic Press, New York.

Alvarez, R Michael, Robert, P Sherman, VanBeselaere, Carla, 2003. Subject acquisition for web-based surveys. Polit. Anal. 11, 23–43.

American National Election Studies, 2012. The ANES 2012 Time Series Study [dataset]. the University of Michigan and Stanford University, Ann Arbor, MI and Palo Alto, CA.

American National Election Study, 2014. User's Guide and Codebook for the ANES 2012 Time Series Study. the University of Michigan and Stanford University, Ann Arbor, MI and Palo Alto, CA.

Ansolabehere, Stephen, Schaffner, Brian F., 2014. Does survey mode still matter? findings from a 2010 multi-mode comparison. Polit. Anal. 22, 285–303.

Atkeson, Lonna Rae, Adams, Alex N., Michael Alvarez, R., 2014. Nonresponse and mode effects in self-and interviewer-administered surveys. Polit. Anal. 22, 304–320.

Ayres, David, 1994. Information, Entropy, and Progress. American Institute of Physics Press, New York.

Baker, Reg, Michael Brick, J., Nancy, A. Bates, Battaglia, Mike, Mick, P. Couper, Jill, A. Dever, Krista, J. Gile, Tourangeau, Roger, 2013. Summary report of the AAPOR task force on non-probability sampling. J. Surv. Statistics Methodol. 1, 90–143.

Bevensee, Robert M., 1993. Maximum Entropy Solutions to Scientific Problems. Prentice Hall, Englewood Cliffs, NJ.

Biemer, Paul P., 2010. Overview of Design Issues: Total Survey Error. Handbook of Survey Research, pp. 27–57.

Bowyer, Benjamin T. and Jon C. Rogowski. Forthcoming. Mode matters: evaluating response comparability in a mixed-mode survey. *Political Sci. Res. Methods*, available on CJO2015. doi:10.1017/psrm.2015.28.

Bredl, Sebastian, Winker, Peter, Kötschau, Kerstin, 2000. A statistical approach to detect interviewer falsification of survey data. Surv. Methodol. 38, 1–10.

Brick, J. Michael, 2011. The future of survey sampling. Public Opin. Q. 75, 872–888.

Casella, G., Berger, R.L., 2002. Statistical Inference, second ed. Duxbury Advanced Series, Belmont, CA.

Chang, L., Krosnick, J., 2010. Comparing oral interviewing with self-administered computerized QuestionnairesAn experiment. Public Opin. Q. 74 (1), 154–167.

Christian, Leah Melani, 2007. How Mixed-mode Surveys Are Transforming Social Research: the Influence of Survey Mode on Measurement in Web and Telephone Surveys. Doctoral dissertation. Washington State University.

Christian, Leah Melani, Dillman, Don A., Smyth, Jolene D., 2008. The effects of mode and format on answers to scalar questions in telephone and web surveys. Adv. Teleph. Surv. Methodol. 12, 250–275.

Christian, Leah Melani, Parsons, Nicholas L., Dillman, Don A., 2009. Designing scalar questions for web surveys. Sociol. Methods & Res. 37, 393–425.

Couper, Mick P., 2000. Web surveys: a review of issues and approaches. Public Opin. Q. 64, 464–494.

Couper, M.P., 2001. The promises and perils of web surveys. The challenges of the internet. In: A. Westlake for the Association for Survey Computing, vol. 35, p. 56.

Cover, Thomas M., Joy, A. Thomas, 1991. Elements of Information Theory. John Wiley and Sons, New York.

Davis, Rachel E., Couper, Mick P., Janz, Nancy K., Caldwell, Cleopatra H., Resnicow, Ken, 2010. Interviewer effects in public health surveys. Health Educ. Res. 25, 14–26.

DeRouvray, Cristel, Couper, Mick P., 2002. Designing a strategy for reducing 'No Opinion' responses in web-based surveys. Soc. Sci. Comput. Rev. 20, 3–9.

Dillman, D.A., Christian, L.M., 2005. Survey mode as a source of instability in responses across surveys. Field Methods 17, 30–52.

Dillman, D.A., Phelps, Glenn, Tortora, Robert, Swift, Karen, Kohrell, Julie, Berck, Jodi, Messer, Benjamin L., 2009. Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the internet. Soc. Sci. Res. 38, 1–18.

Dillman, D.A., Smyth, J.D., Christian, L.M., 2014. Internet, Phone, Mail, and Mixed-mode Surveys: the Tailored Design Method. John Wiley & Sons.

Evans, J.H., Bryson, B., DiMaggio, P., 2001. Opinion polarization: important contributions, necessary limitations. Am. J. Sociol. 106 (4), 944–959.

Gill, Jeff, 2005. An entropy measure of uncertainty in vote choice. Elect. Stud. 24, 371–392.

Gilula, Z., Haberman, S.J., 1995. Dispersion of categorical variables and penalty functions: derivation, estimation, and comparability. J. Am. Stat. Assoc. 90 (432), 1447–1452.

Gosling, S.D., Vazire, S., Srivastava, S., John, O.P., 2004. Should we trust web-based studies? a comparative analysis of six preconceptions about internet questionnaires. Am. Psychol. 59 (2), 93–104.

Green, D.P., Palmquist, B., 1994. How stable is party identification? Polit. Behav. 16 (4), 437–466.

Greene, Zachary. Forthcoming. Competing on the Issues: how experience in government and economic conditions influence the scope of parties' policy message. *Party Politics* FirstView doi:10.1177/1354068814567026.

Groves, Robert M., 2011. Three eras of survey research. Public Opin. Q. 75, 861–871.

Groves, Robert M., Lyberg, Lars, 2010. Total survey error: past present, and future. Public Opin. Q. 74, 849–879.

Groves, Robert M., Fowler Jr., F.J., Couper, M.P., Lepkowski, J.M., Singer, E., Tourangeau, R., 2011. Survey Methodology. John Wiley & Sons.

Hays, Ron D., Liu, Honghu, Kapteyn, Arie, 2015. Use of internet panels to conduct surveys. Behav. Res. Methods 1–6.

Jaynes, Edwin T., 1957. Information theory and statistical mechanics. Phys. Rev. 106, 620–630.

Jaynes, Edwin T., 1968. Prior probabilities. IEEE Transm. Syst. Sci. Cybern. SSC 4, 227–241.

Jaynes, Edwin T., 1982. On the rationale of maximum-entropy methods. Proc. IEEE 70 (9), 939–952.

John, Peter, Jennings, Will, 2010. Punctuations and turning points in British politics: the policy agenda of the Queen's Speech, 1940-2005. Br. J. Political Sci. 40, 561–586.

Johns, Robert, 2005. One size Doesn't fit all: selecting response scales for attitude items. J. Elections, Public Opin. Parties 15 (2), 237–264.

McNabb, David E., 2013. Nonsampling Error in Social Surveys. Sage Publications, Thousand Oaks, CA.

Millar, M., Dillman, D., 2012. Do mail and internet surveys produce different item nonresponse rates? an experiment using random mode assignment. Surv. Pract. 5 (2).

Nagelhout, Gera E Nagelhout, Marc, C Willemsen, Mary, E Thompson, Geoffrey, T Fong, van den Putte, Bas, de Vries, Hein, 2010. Is web interviewing a good alternative to telephone interviewing? Findings from the international tobacco control (ITC) Netherlands Survey. BMC Public Health 10, 351.

O'Muircheartaigh, C., Campanelli, P., 1998. The relative impact of interviewer effects and sample design effects on survey precision. J. R. Stat. Soc. Ser. A 161 (1), 63–77.

Park, D.K., Gelman, A., Bafumi, J., 2004. Bayesian multilevel estimation with post-stratification: state- level estimates from national polls. Polit. Anal. 12 (4), 375–385.

Rivers, Douglas, Bailey, Delia, 2009. Inference from matched samples in the 2008 US national elections. Proc. Jt. Stat. Meet. 627–639.

Rubin, Donald B., 2004. Multiple Imputation for Nonresponse in Surveys, second ed. Wiley, New York.

Ruelle, David, 1991. Chance and Chaos. Princeton University Press, Princeton.

Ryu, Hang K., 1993. Maximum entropy estimation of density and regression functions. J. Econ. 56, 397–440.

Shannon, C., 1948. A mathematical theory of communication. Bell Syst. Technol. J. 27 (379–423), 623–656.

Singer, E., Frankel, M.R., Glassman, M.B., 1983. The effect of interviewer characteristics and expectations on response. Public Opin. Q. 47 (1), 68–83.

Slowinski, S.M., 1988. Control and measurement of nonresponse error in establishment surveys. SRMS Proc. 321–325.

Tribus, Myron, 1961. Information theory as the basis for thermostatics and thermodynamics. J. Appl. Mech. 28, 1–8.

Tribus, Myron, 1986. Information and thermodynamics: bridging the two cultures. J. Non-Equilibrium Thermodyn. 11, 247–260.

Van Campenhout, Jan M., Cover, Thomas M., 1981. Maximum entropy and conditional probability. IEEE Trans. Inf. Theory 27 (4), 483–489.

Weisberg, Herbert F., 2005. The Total Survey Error Approach: a Guide to the New Science of Survey Research. University of Chicago Press, Chicago.

West, Brady T., Kreuter, Frauke, Jaenichen, Ursula, 2013. Interviewer effects in face-to-face surveys: a function of sampling, measurement error, or nonresponse? J. Off. Statistics 29, 277–297.

Ye, Cong, Fulton, Jenna, Tourangeau, Roger, 2011. More positive or more extreme? A meta- analysis of mode differences in response choice. Public Opin. Q. 75, 349–365.

Yeager, David S., Krosnick, Jon A., Chang, LinChiat, Javitz, Harold S., Levendusky, Matthew S., Simpser, Alberto, Wang, Rui, 2011. Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. Public Opin. Q. 75, 709–747.