Article

Research and Politics
July-September 2015: 1–11

DOI: 10.1177/2053168015590681
rap.sagepub.com
SAGE

Measurements of cognitive skill by survey mode: Marginal differences and scaling similarities

Andrew Gooch

Abstract

This paper addresses how measurements of cognitive skill differ based on survey mode, from a face-to-face interview to a self-completed survey, using the Wordsum vocabulary test found in the General Social Survey. The Wordsum acts as a proxy for general cognitive skill, and it has been used to predict a variety of political variables. Therefore, knowing differences in cognitive skill by mode are important for political science research because of the proliferation of self-completed Internet surveys. I leverage a large-scale mode experiment that randomizes a general population sample into a face-to-face or self-completed interview. Results show that historically easy questions are more likely to yield correct answers in the face-to-face treatment, but modest-to-difficult test questions have a higher rate of correct answers in the self-completed treatment (marginal distributions). A cognitive skill scale using item response theory, however, does not differ by mode because the ordering of ideal points does not change from a face-to-face interview to a self-completed survey. When applying the scale to a well-established model of party identification, I show no difference by mode, suggesting that a transition from face-to-face interviews to self-completed surveys may not alter conclusion drawn from models that use the Wordsum test.

Keywords

survey methodology, mode experiment, item response theory, Wordsum, party identification

Introduction

The importance of mode research for political science

Since the 1940s, most of the data used in political science research on elections, public opinion, and voting behavior was gathered by the American National Election Study (ANES) or General Social Survey (GSS) through in-person, face-to-face surveys. In-person interviewing, however, is costly, and becoming more so over time. In 2012, the cost per interview was approximated at US$2,100.00 (inclusive of both interviews but exclusive of staffing costs) to produce the ANES (Segura et al., 2010). An attractive alternative to face-to-face interviewing is online, self-completed surveys because of the proliferation of Internet usage, the increase in computer literacy, and the lower cost per interview. While self-completed surveys are becoming more popular in political science, much of the survey methodology literature does not directly address the mode differences between face-to-face interviewing and self-completed surveys. Regardless of whether a researcher prefers face-to-face or self-completed surveys, federal budgeting realities might force political scientists to pursue more online, self-completed surveys.

In this paper, I address how measures of cognitive skill differ by survey mode. To isolate the effects of mode, I use data from a large-scale experiment in which we randomly assigned respondents into a face-to-face or self-completed survey with identical questions (n=505 per mode).[1] The survey treatment assignments occurred *after* respondents agreed to participate in the experiment, which eliminates

Department of Political Science, University of California, Los Angeles, CA, USA.

Corresponding author:
Andrew Gooch, Department of Political Science,University of California, Los Angeles,4289 Bunche Hall, Los Angeles, CA 90095, USA.
Email: aagooch@ucla.edu

any of the confounders related to sampling and selection bias associated with the mode of survey administration. Results show a difference by mode for the marginal distribution of *individual* knowledge questions, where more respondents are answering moderate to difficult questions correctly in the self-completed mode. But once the knowledge test *as a whole* is scaled together using a two-parameter item response theory (IRT) model, little difference in cognitive skill exists by mode because the cognitive skill ideal points are order preserving regardless of mode. I then use the cognitive skill ideal points in a model of party identification to show that conclusions from the Wordsum test do not differ in a face-to-face interview relative to a self-completed survey. These results demonstrate that the presence of an interviewer can affect the marginal distribution of knowledge questions; respondents will be less likely to answer difficult questions correctly with an interviewer. But when the cognitive skill test is aggregated together, which is how most political scientists use knowledge tests, no difference by mode exists. Therefore, my results present an initial piece of evidence that a transition from a face-to-face interview to a self-completed survey might not alter conclusions from knowledge models even if the marginal distributions of each knowledge question will be different.

Measuring cognitive skill

I measure cognitive skill through the Gallup–Thorndike Verbal Intelligence Test (Thorndike, 1942; commonly called a "Wordsum" test, which became a part of the GSS in 1972 (Davies et al., 2007). Wordsum questions present respondents with a single vocabulary word and five answer choices: the answer choices are individual words as well. Respondents are then asked to select the answer choice that comes closest to the meaning of the prompted word.[2] As opposed to longer intelligence tests, the GSS Wordsum test is brief, only ten questions: six "easy" words (characterized by a high level of correct responses in the GSS) and four "hard" words (low level of correct responses). A wide range of fields such as political science, statistics, education and psychology use the Wordsum test to measure various dimensions of intelligence; for a comprehensive and thorough review of the Wordsum test in the social sciences, see Malhotra et al., 2007) found at the National Opinion Research Center and Cor et al., 2012).

Early research using the Wordsum test as a general measure of cognitive skill shows that the test is highly correlated (from 0.75 and above) with a more extensive, in-depth intelligence test (Miner, 1957). This connection between general intelligence and vocabulary is replicated in more recent studies as well (Alwin, 2010; Zhu and Weiss, 2005), demonstrating that a short vocabulary test can be an effective proxy for general intelligence and cognitive skill. In political science, the Wordsum test is used as a measure of cognitive skill to predict voter turnout, political knowledge, preferences on economic issues, and general ideology (Caplan and Miller, 2010; Erikson et al., 2002; Rempel, 1997; Verba et al., 1985).

I selected four of the ten Wordsum items that were used in this randomized experiment: the four words are Broaden, Space, Cloistered and Allusion (asked in that order). The first two words are historically considered easy and the latter two are consider difficult because a majority of respondents answer the first two correctly and a majority answer the second two incorrectly (Cor et al.,2012; Malhotra et al., 2007). The four items used in this study were randomly selected within each level of difficulty so that I tested two easy and two hard questions. How closely related is my abbreviated four-item scale with the full ten-item scale? To make this comparison, I first ran an IRT model on all 10 items from the 2010 GSS, and then I ran the same IRT model using only the four items taken from the 2010 GSS. The Pearson correlation on the two scales is 0.79. Although the scales are not perfectly correlated, the abbreviated scale correlates very strongly with the full scale.

Knowledge differences by survey mode

Does survey mode change a respondent's answer to a question? There are many reasons to expect response differences due to survey mode, whether it face-to-face, over the phone, or self-completed (Acree et al., 1999; Bishop et al., 1988; Chang and Krosnick, 2010; De Leeuw, 2002; Fowler et al., 1998; Gano-Phillips and Fincham, 1992; Kiesler and Sproull, 1986; Malhotra, 2009; Shulman and Boster, 2014; Sudman and Bradburn, 1974). In one other true mode investigation in which respondents (college sophomores) were assigned to a mode of interview, Chang and Krosnick (2010) found that lower cognitive skill respondents who completed the survey on a computer exhibit higher concurrent validity: "Oral presentation might pose the greatest challenges for respondents with limited cognitive skills, because of the added burden imposed by having to hold a question and response choices in working memory while searching long-term memory and generating a judgment" (Chang and Krosnick, 2010: 155). But in the Chang and Krosnick (2010) experiment, the researchers were not interested in how measurements of cognitive skill differ by mode, which is the focus of this study. Instead, they used SAT scores as a proxy for cognitive skill to show how intelligence can interact with survey mode (Chang and Krosnick, 2010).[3]

With knowledge tests in the self-completed treatment, I expect a lower rate of correct answers among the *easy* questions due to satisficing in the self-completed treatment (Krosnick, 1991; Malhotra, 2009). Satisficing occurs when individuals are presented with a task, and instead of maximizing their ability to complete the task, individuals will only exert a minimum amount of effort (Krosnick, 1991; Simon, 1957). Satisficing leads individuals to be "less than thorough" when interpreting survey questions (Krosnick, 1991; Krosnick et al., 1996). Recent work shows

that individuals can be less engaged with easy tasks in self-completed surveys, leading to a higher level of satisficing (Malhotra, 2009). As a result, I expect a higher level of satisficing in the self-completed treatment for the less-difficult Wordsum questions. But harder questions encourage more careful consideration in self-completed surveys, and so I do not expect satisficing to persist for the difficult Wordsum questions.

For more difficult questions, I expect a higher rate of correct answers in the self-completed treatment compared with the face-to-face treatment because of the presence of an interviewer. Tourangeau et al. (2000: 179) detail a mechanism from which the instability of responses arises in different survey settings. Which considerations a respondent retrieves and places weight on depends on the momentary accessibility of each consideration, and these considerations are influenced by many factors, some temporary (Tourangeau et al., 2000). I argue that the accessibility of such considerations can also be *mode* dependent. Tourangeau et al. consider this circumstance, arguing that judgments about considerations may also be affected by momentary changes to the environment, including the presence of an interviewer (Tourangeau et al., 2000: 180). Difficult questions asked by an interviewer, therefore, might reduce the level of correct answers because the interviewer inhibits the respondent from utilizing their retrieval and judgment abilities. When respondents sit down with an interviewer to complete a survey, they might feel added pressure to answer difficult questions correctly, which would not otherwise exist if respondents were alone behind a computer.

Data: randomized mode experiment with a block design

In order to isolate mode effects on measurements of cognitive skill, I leverage a large-scale randomized experiment conducted during the summer of 2011 that tested and evaluated self-completed surveys on a computer as a replacement to interviewer-assisted surveying.[4] The experiment took place at the CBS research facility within the MGM Grand Hotel in Las Vegas, Nevada, where CBS conducts daily focus groups on its programming. Face-to-face interviews were conducted by six professional interviewers in one of four simulated living rooms in the research facility. The self-completed computer surveys took place individually in small rooms, which resemble a small home office, and respondents could take as long as they needed.[5]

The randomized experiment used a blocking design on 3 key indicators, age, race and gender, which creates 18 distinct blocks. Blocking ensures that both the face-to-face and the self-completed modes are balanced on demographics that might confound estimated treatment effects (Green and Gerber, 2012). After respondents agreed to participate, they were brought into a waiting room where age, race and gender were estimated by graduate students and entered into an algorithm using R that made the treatment assignment.[6] After the treatment assignment (face-to-face or self-completed; a respondent is then matched with the next agreeable participant with identical demographics, who is then assigned to the *opposite* mode treatment. This blocking technique created a sample that is balanced on treatment assignment, age, race, and gender for a total sample size of 1010.

Results

Marginal distributions by mode

Figure 1 displays the percentage point difference by mode, taking self-completed responses minus the face-to-face responses, where positive values show more correct answers in the self-completed treatment and negative values indicate more correct answers in the face-to-face treatment. Each difference is accompanied by a 95% confidence interval using blocked standard errors (Green and Gerber, 2012).[7] Figure 1 shows that the easier words (Space and Broaden) have a higher rate of correct answers in the face-to-face treatment, while the two more difficult questions (Allusion and Cloistered) show more correct answers in the self-completed treatment relative to face-to-face. These results suggest that an interviewer might be more beneficial to respondents who might have trouble with historically easy words, where 90% of the population can answer correctly, but the interviewer might be a hindrance for more challenging words such as Allusion and Cloistered.

In addition, self-completed respondents are more likely to satisfice with easier questions. These results suggest that respondents give less thought to easy tasks with a self-completed survey, potentially because simply tasks "may cause respondents to become bored and not expend cognitive effort to carefully consider the item" (Malhotra, 2009: 182). On the other hand, difficult tasks using self-completed surveys do not encourage satisficing because they require more thought and effort to answer (Malhotra, 2009). My results support these conclusions: respondents might be less likely to answer easy questions correctly in the self-completed treatment relative to face-to-face. Moreover, my evidence shows that self-completed respondents did not feel the need to look up answers to factual questions on the Internet, which is a legitimate concern when testing knowledge levels online.[8]

Scaling cognitive skill by mode

Typically, survey knowledge items are not used by scholars on a question-by-question basis, but instead as a collection of questions that constitute cognitive skill or knowledge more broadly defined. To that end, I jointly scale both modes together using a two-parameter IRT model, and then compare the ideal points for differences by mode. IRT models show the relationship between some latent trait (in this case, cognitive skill) and the response given to each question (Albert, 1992; De Ayala, 2009; Clinton et al., 2004; DeMars, 2010; Embretson

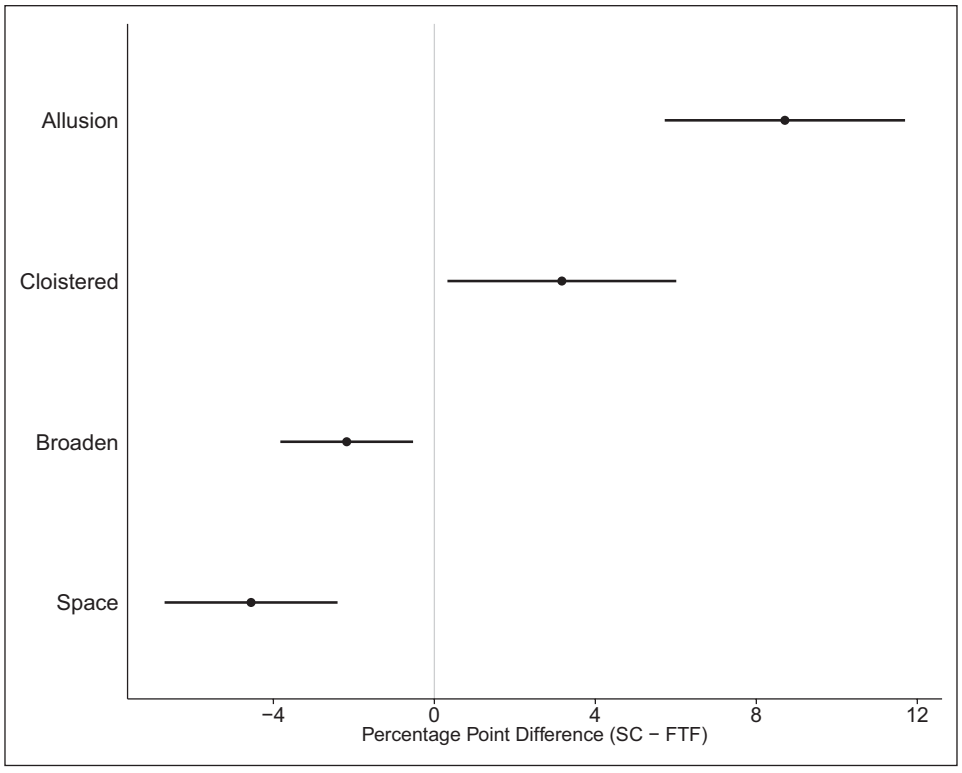

Figure 1. Percentage point difference in correct answers by mode.
Note: Percentage point (mean) differences in correct answers by mode with 95% confidence intervals. Positive values indicate more correct answers in the self-completed treatment, and negative values indicate more correct answers in the face-to-face treatment.

and Prenovost 1999; Hambleton and Swaminathan, 1999; Jackman, 2009; Lord, 1980).[9]

Figure 2 shows a density plot of the cognitive skill scale, separated by mode after the ideal points were estimated.[10] Visual inspection of both plots shows very little difference in the shape of each scale by mode. Face-to-face respondents tend to cluster just below average (zero; out-numbering self-completed respondents, but self-completed respondents out-number face-to-face respondents on the high and low ends of the scale. Ostensibly, the difference appears to be modest.

The current IRT literature does not provide a general test for comparing IRT scales, and as a result, I am employing a new method of comparing scales using the Markov chain iterations. I compared the mean ideal point estimates during each iteration of the chain for both modes of a jointly scaled IRT model. I can establish that the face-to-face and self-completed cognitive skill distributions do not differ from each other by comparing their mean estimates during each iteration (the posterior over the distributions). If both distributions are the same, we should observe substantial overlapping between the face-to-face and the self-completed ideal points during each iteration of the Markov chain. Figure 3 plots these mean differences as a density for each iteration with a vertical line at zero (indicating no difference in means). To calculate this difference, I subtracted the self-completed ideal point means for each respondent at each

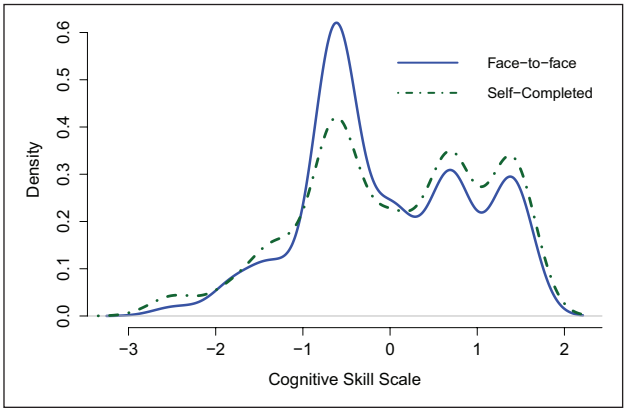

Figure 2. Jointly scaled cognitive skill compared by survey mode.
Note: Two parameter IRT model with mean zero and standard deviation of one, where both modes are scaled together.

iteration in the chain from the face-to-face ideal point estimates:

$$MeanDiff_i = \overline{FTF}_i - \overline{SC}_i$$

where $\overline{FTF}_i$ is the ideal point estimates for each iteration, i, in the face-to-face mode, and $\overline{SC}_i$ is the ideal point estimates for each iteration in the self-completed mode.[11]

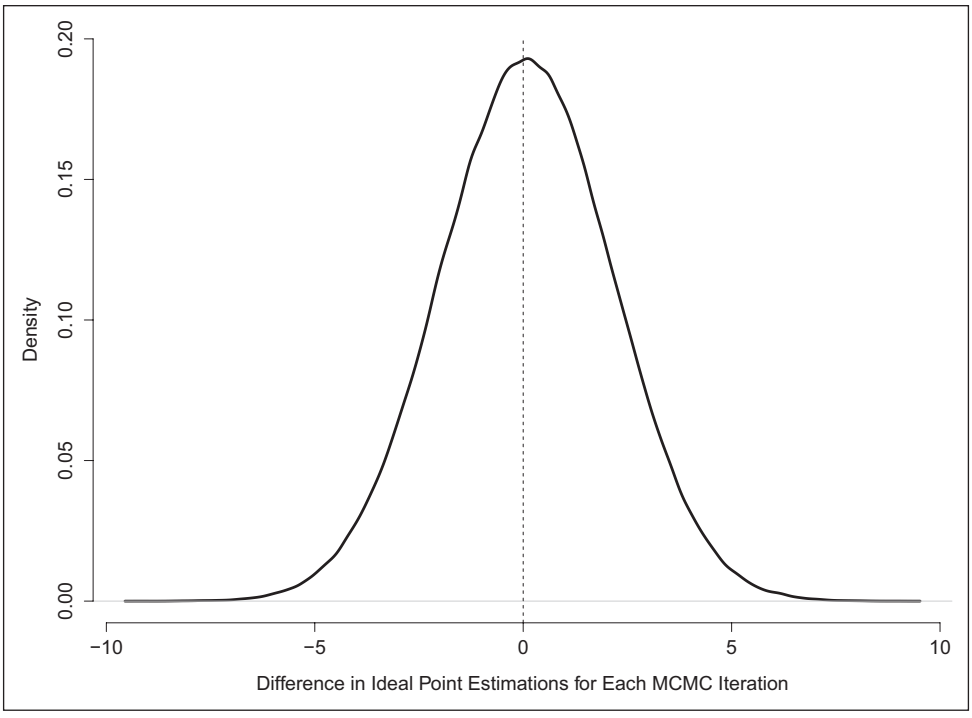

Figure 3. Comparing scales by mode: difference of ideal point estimations during each iteration.
Note: Jointly scaled IRT model separated by mode to see whether the scales differ by mode. Only 52% of differences exceed zero; 95% of differences must exceed zero for the distributions to be different at a 1.96 level.

From this mean difference, I calculated the percentage of iterations that differ by mode, which can be found by summing the total number of mean differences that exceed 0, and then dividing the sum by the total number of iterations:

$$DiffTest = \frac{MeanDiff_i > 0}{I}$$

which creates a percentage of iterations that are different by mode. If both modes are different, then the mean difference should not exceed zero more than 5% of the time (mirroring a hypothesis test with a significance level of 1.96). As a point of comparison, I can use this difference test statistic to compare simulated distributions that we know are different to make sure my test works. Take two random normal distributions, for example, with sample sizes of 1000 each (simulating the Markov chain iterations; both with standard deviations of 1, but with differing means of 4 and 3 (simulating different posteriors over the distributions for the face-to-face mode and the self-completed mode). Using my $DiffTest$ difference statistic to compare the simulated normal distributions, I find that 1% of the mean differences are greater than zero, suggesting that the simulated distributions are different at a 99% confidence level with a 2.575 Z -score.

Returning to the survey mode data, I find that only 52% of the face-to-face means exceed the self-completed means, which is far from a standard 95% confidence level. In other words, my difference test statistic suggests that the face-to-face mode and self-completed mode are only different at a 48% level; therefore, I can conclude that the Wordsum IRT scales are not different by mode. These results show that the cognitive skill scales are order preserving within mode even if there are marginal differences found in the previous section. The top quarter of cognitive skill respondents in the face-to-face interview, for example, will still be the top quarter of cognitive skill respondents in the self-completed survey. For further evidence, please see Table 2 in which I predict the ideal points with education levels, a measure of knowledge that is unaffected by mode, I show no mode difference. In addition, I find that a simple additive scale does not differ by mode using a t -test of means ($p = 0.40$), demonstrating that my results are consistent across scaling procedures.

Application in political science

This section applies the cognitive skill scale by mode to a political science question, showing that a well-established finding does not change based on survey mode. A robust finding in American politics is that an increasing level of knowledge (usually operationalized as a fact-based test) is strongly associated with political constraint (Converse, 1964; Zaller, 1992). That is, higher levels of knowledge are associated with reporting attitudes that are identical to a respondent's preferred political party (Converse, 1964; Zaller, 1992). This type of knowledge effect, however, is uniquely ideological because high knowledge moderates are not any more likely to support a political party than low

Table 2. Predicting cognitive skill with education by mode.

	Jointly scaled			
Covariates	Pooled	FTF	SC	GSS 1972–2010
(intercept)	0.49 (0.09)	0.50 (0.08)	0.49 (0.09)	0.27 (0.01)
No High School Diploma	−0.96 (0.30)	−1.45 (0.31)	−0.96 (0.31)	−0.38 (0.01)
High School Diploma	−1.08 (0.15)	−0.93 (0.14)	−1.08 (0.16)	−0.31 (0.01)
Some and 2 Year College	−0.67 (0.12)	−0.71 (0.11)	−0.67 (0.12)	−0.24 (0.01)
Four-year College	−0.29 (0.12)	−0.51 (0.12)	−0.29 (0.12)	−0.09 (0.01)
Face-to-face Treatment	0.01 (0.13)			
Face-to-face * No High School	−0.49 (0.44)			
Face-to-face * High School	0.15 (0.22)			
Face-to-face * Some College	−0.04 (0.16)			
Face-to-face * Four Year College	−0.22 (0.17)			
R^2	0.12	0.11	0.12	0.07
Sample size	1010	505	505	54,925

Note: OLS regression results with standard errors in parenthesis. The education reference group is graduate degree. The dependent variable is the Wordsum questions scaled as a two-parameter IRT model. The first three columns come from our experimental data, and the last column uses data from the General Social Survey, 1972 to 2010. Insignificant interaction terms indicate no difference by mode using education, a proxy for cognitive skill that is unaffected by survey mode.

knowledge moderates. The effect of knowledge is borne out through ideology where high knowledge liberals (conservatives) are more likely to be strong Democrats (Republicans); and low knowledge liberals (conservatives) are more likely to be weak Democrats (Republicans) (Zaller, 1992).

This section shows that these well-established findings are confirmed regardless of survey mode. These findings represent an initial step toward showing that little difference exists in cognitive skill scales by survey mode. I model party identification using cognitive skill and ideology for both modes using an ordinary least squares (OLS) regression:

$$PartyID_{\text{im}} = \alpha_{\text{im}} + \beta I_{\text{im}} + \beta C_{\text{im}} + \beta (I * C)_{\text{im}}$$

where $PartyID$ is the party identification (seven-point) for respondent i in mode m, I is the ideology (three-point) for respondent i in mode m, C is the cognitive skill ideal point for respondent i in mode m, and $I*C$ is the interaction between cognitive skill and ideology. Party identification is also estimated as a pooled model using an interaction term for completing the survey in the face-to-face mode. These regression coefficients are in Table 3 with standard errors in parenthesis. The first column shows the interactive model, and none of the interactive variables are significant at a 95% level, which implies no difference by mode. Cognitive skill's influence on party identification is moderated by ideology, and this three-way interaction testing a difference by mode is modest with a large standard error.[12]

To help visualize the lack of difference by mode, Figure 4 plots these results. Each line in Figure 4 is the predicted level of party identification by cognitive skill for each type of ideology: conservatives, moderates and liberals. Although the plots are not identical, the general pattern shows no difference by mode. In addition to these results, similarities by mode for other political science models are found elsewhere too; for example, no difference by mode exists for models of retrospective voting and issue voting models (Fiorina, 1981; Gooch and Vavreck, 2015, manuscript A). But these results are only one knowledge scale from a single experiment, and more evidence using different conceptualization of cognitive skill is needed to generalize further.

Conclusion and implications for survey methodology in political science

If budgeting realities force political scientists to pursue a less costly mode of interview, such as a self-completed survey online, specific differences and similarities should be expected. The marginal differences by mode on the Wordsum test are driven by the level of question difficulty. Easy questions are answered correctly more often in the face-to-face treatment, and modest to difficult questions are answered correctly more often in the self-completed treatment.[13] But when the Wordsum items are considered together as a test, which is how most researchers use knowledge items, no mode differences exist. In addition, my party identification model replication by mode is one piece of evidence that inference drawn from statistical models will not change with a transition from face-to-face interviews to self-completed surveys. Future research needs to explore other aspects of cognitive skill, and how measurements of it may differ by mode.

Table 3. Predicting party ID with ideology and cognitive skill by mode.

	Pooled	Face-to-face	Self-completed
Intercept	3.53 (0.07)	3.66 (0.07)	3.53 (0.07)
Ideology	1.53 (0.10)	1.52 (0.10)	1.53 (0.10)
Cognitive Skill	0.22 (0.07)	0.05 (0.08)	0.22 (0.07)
Ideology * Cognitive Skill	0.27 (0.10)	0.33 (0.10)	0.27 (0.10)
Face-to-face	0.14 (0.10)		
Ideology * FTF	−0.01 (0.14)		
Cognitive Skill * FTF	−0.17 (0.11)		
Ideology * Cognitive Skill * FTF	0.06 (0.14)		
R^2	0.34	0.34	0.35
Sample size	1010	505	505

Note: OLS regression coefficients with standard errors in parenthesis. The dependent variable is a seven-point party identification scale. Party identification non-response and "other" responses are coded as independent, and ideology non-response is set as the mean.

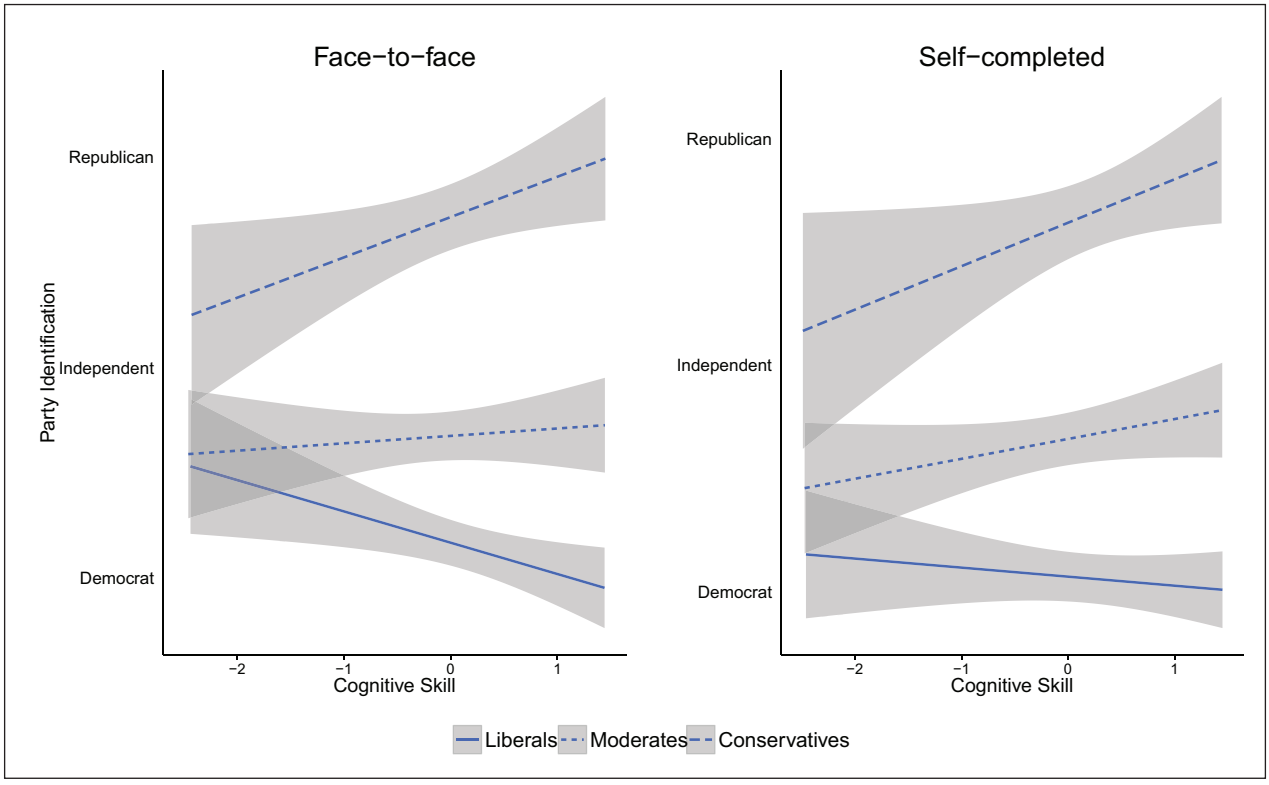

Figure 4. Face-to-face: predicting party ID with ideology and cognitive skill.
Note: Party identification predictions from a linear regression model. The left panel is face-to-face and the right panel is self-completed. Dependent variable is a seven-point party identification variable, where high values are associated with identifying as a Republican and low values are associated with Democratic identifiers. Independent variables are ideology, cognitive skill, and an interaction between ideology and cognitive skill. The top line in the plot is the predicted probability of party identification for conservatives, the middle line is independents, and the bottom line is liberals.

The lower rate of correct answers for the easy questions in the self-completed survey might be due to satisficing: individuals are less engaged with easy tasks in self-completed surveys (Malhotra, 2009). But difficult questions encourage more careful consideration in self-completed surveys, and so satisficing does not persist with difficult questions (Malhotra, 2009). Moreover, the higher rate of correct answers in the self-completed survey supports the findings in educational testing (Ben-Shakhar and Sinai, 1991; Casey et al., 1997; Cronbach, 1946; Shulman and Boster, 2014). And Tourangeau et al., (2000, 179) detail a mechanism from which the instability of responses arises in different survey settings. Which considerations a respondent retrieves and places weight on depends on the momentary accessibility of

each consideration, and these considerations are influenced by many factors, some temporary (Tourangeau et al., 2000). The accessibility of such considerations, demonstrated in this experiment, can also be *interviewer* dependent (Tourangeau et al., 2000: 180). Difficult questions asked by an interviewer, therefore, might reduce the level of correct answers because the interviewer inhibits the respondent from utilizing their retrieval and judgment abilities. When respondents sit down with an interviewer to complete a survey, they might feel added pressure to answer factual questions correctly, which would not otherwise exist if respondents were alone behind a computer.

Acknowledgements

I would like to thank Professor Lynn Vavreck for involving me in this project. I also thank the project's manager, Brian Law, who kept things running on time and effectively at the MGM Grand, and the graduate students who helped administer the experiment in Las Vegas: they are Felipe Nunes, Sylvia Friedel, Gilda Rodriguez, Adria Tinnin and Chris Tausanovitch. I also greatly appreciate helpful comments on this paper from Chris Tausanovitch, Lynn Vavreck, Jim DeNardo, Michael Chwe and John Zaller. I also appreciate the help of Doug Rivers and Jeff Lewis, who wrote parts of the backend program responsible for the randomization and blocking. Finally, I thank John Aldrich, Larry Bartels, Alan Gerber, Gary Jacobson, Simon Jackman, Vince Hutchings, Gary Segura, John Zaller and Brian Humes, who helped to design this experiment in the summer of 2010.

Funding

This research is supported by a grant from the National Science Foundation (award number SES-1023940 to Lynn Vavreck).

Notes

1. These data were collected with funding from the National Science Foundation (NSF). NSF reference number SES-1067949.
2. Please see the appendix for instructions and question wording.
3. See also Gooch and Vavreck (2015) for differences in political knowledge and non-response by survey mode.
4. Lynn Vavreck, NSF SES-1067949: Methodology, Measurement, and Statistics Grant, *Increasing Power and Decreasing Costs: A New Method for Drawing High-Quality National Probability Samples of U.S. Citizens, 2011–2014*.
5. Each computer connected to the Internet for respondents to complete the survey, and a new browser window was opened by a research assistant for each new respondent. Opening a new browser demonstrates to respondents that the survey is not confined within their computer terminal, and they are free to use the Internet if desired.
6. The algorithm was written by Douglas Rivers and Jeff Lewis. The algorithm contained three available age ranges (30 and below, 31–59 and 60 plus; three race categories (White, African American and Hispanic) and two gender categories.
7. Confidence intervals are used to characterize the error surrounding the mean differences by mode. They simply provide a visual test for the mean differences between respondents randomly assigned to a face-to-face or self-completed survey; I am not making inferences about the general population. When the confidence interval does not cross zero, the difference in means by mode are significant at a 95% level.
8. I downloaded the browser history of each self-completed respondent to confirm that no one cheated, and this can be found in the appendix.
9. Jointly scaled ideal points has the benefit of maintaining the same mean and standard deviation (zero and one, respectively) regardless of treatment assignment. If I scaled face-to-face and self-completed separately, comparing ideal points becomes untenable because they would not be on the same scale.
10. Table 1 shows IRT parameter estimates. Both modes had a chain length of 1 million with a burn-in of 10,000, thinned by 1000 using the R package Ideal. The mean effective sample size for each ideal point is 988.
11. The ideal points for each iteration are stored in a matrix, where each row is an iteration's mean estimates and each column is a respondent. I calculated the mean of each row, and then subtracted the face-to-face means in each row minus the self-completed means (Figure 3).
12. The cognitive skill coefficient in the un-pooled models in Table 3 is positive and significant in the self-completed treatment (0.22 with a standard error of 0.07; but the face-to-face treatment does not have a significant cognitive skill coefficient (0.05 with a standard error of 0.07). However, this difference is more apparent than real, and the interactive model shows that no difference actually exists from the two-way interaction of "Cognitive Skill * FTF" (–0.17 coefficient and a 0.11 standard error). At conventional levels of testing, the null hypothesis of no difference between modes is not rejected. The difference in coefficients and their standard errors shows that cognitive skill on party identification is more precisely estimated in the self-completed treatment and that the face-to-face estimate is noisier by comparison.
13. Elsewhere, research shows that self-completed surveys also have a higher level of correct answers among moderate to difficult political knowledge and non-political knowledge questions as well (Gooch and Vavreck, 2015: manuscript B).

References

Acree M, Ekstrand M, Coats TJ and Stall R (1999) Mode effects in surveys of gay men: a within-individual comparison of responses by mail and telephone. *The Journal of Sex Research* 36(1): 67–75.

Albert J (1992) Bayesian estimation of normal Ogive item response curves using Gibbs sampling. *Journal of Educational Statistics* 17: 251–269.

Alwin DF (2010) Family of origin and cohort differences in verbal ability. *American Sociological Review* 56(5):625–638.

Ben-Shakhar G and Sinai Y (1991) Gender differences in multiple-choice tests: the role of differential guessing tendencies. *Journal of Educational Measurement* 28(1): 23–35.

Bishop GF, Hippler H-J, Schwarz N and Strack F (1988) A comparison of response effects in self-administered and telephone surveys. In: Groves RM, Biemer PP, Lyberg LE, Massey JT, Nicholls WL and Waksberg J (eds), *Telephone Survey Methodology*. New York: Wiley.

Casey MB, Ronald LN and Elizabeth P(1997) Mediators of gender differences in mathematics college entrance test scores:

a comparison of spatial skills with internalized beliefs and anxieties. *Developmental Psychology* 33(4): 669.

Chang L and Krosnick JA (2010) Comparing oral interviewing with self-administered computerized questionnaires: an experiment. *Public Opinion Quarterly* 74(1): 154–167.

Caplan B and Miller SC (2010) Intelligence makes people think like economists: Evidence from the General Social Survey. *Intelligence* 38(6): 636–647.

Clinton J, Jackman S and Rivers D (2004) The statistical analysis of roll call data. *American Political Science Review* 98: 335–370.

Converse PE (1964) The nature of belief systems in mass publics. In Apter DE (ed.), *Ideology and Discontent*, pp. 206–261.

Cor MK, Haertel E, Krosnick JA and Malhotra N (2012) Improving ability measurement in surveys by following the principles of IRT: The Wordsum Vocabulary Test in the General Social Survey. *Social Science Research* 41(5): 1003–1016.

Cronbach LJ (1946) Response sets and test validity. *Educational and Psychological Measurement* 6: 475–494.

Davis JA, Smith TW and Marsden PV (2007) *General Social Survey, 1972–2006: Cumulative Codebook*. Chicago, IL: National Opinion Research Center.

De Ayala RJ (1997) *The Theory and Practice of Item Response Theory*. New York: Gulford Press.

DeBell M (2010) *How to Analyze ANES Survey Data*. Stanford University, American National Election Study.

De Leeuw ED (2002) *Data Quality in Mail, Telephone and Face to Face Surveys*. Cambridge: Cambridge University Press.

DeMars C (2010) *Item Response Theory: Understanding Statistics Measurement*. Oxford: Oxford University Press.

Embretson SE and Prenovost LK (1999) Item response theory in assessment research. In Kendall PC, Butcher JN and Holmbeck GN (eds), *Handbook of Research Methods in Clinical Psychology*. New York: Wiley.

Erikson RS, MacKuen MB and Stimson JA (2002) *The Macro Polity*. Cambridge: Cambridge University Press.

Fiorina MP (1981) *Retrospective Voting in American National Elections*. New Haven, CT: Yale University Press.

Fowler FJ, Roman AM and Zhu XD (1998) Mode effects in a survey of Medicare prostate surgery patients. *Public Opinion Quarterly* 62: 29–46.

Gano-Phillips S and Fincham FD (1992) Assessing marriage via telephone interviews and written questionnaires: a methodological note. *Journal of Marriage and Family* 54: 630–635.

Gerber AS and Donald PG (2012) *Field Experiments: Design, Analysis, and Interpretation*. New York: WW Norton.

Gooch AA and Vavreck L (2015) Face-to-face interviews vs. self-completed surveys: canonical Findings in American Politics.

Gooch AA and Vavreck L (2015) How face-to-face interviews and cognitive skill affect non-response: a randomized experiment assigning mode of interview.

Hambleton RK and Swaminathan H (1999) Item response theory: principles and applications. In Kendall PC, Butcher JN and Holmbeck GN (eds), *Handbook of Research Methods in Clinical Psychology*. New York: Wiley.

Hambleton RK, Swaminathan H and Rogers HJ (1991) *Fundamentals of Item Response Theory*. New York: Sage.

Jackman S (2009) *Bayesian Analysis for the Social Sciences*. Hoboken, NJ: Wiley.

Kiesler S and Sproull LS (1986) Response effects in the electronic survey. *Public Opinion Quarterly* 50: 402–413.

Krosnick JA (1991) Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5: 213–236.

Krosnick JA and Alwin DF (1987) An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly* 51(2): 201–219.

Krosnick JA, Narayan S and Smith WR (1996) Satisficing in surveys: initial evidence. *New Directions for Evaluation* 70(Summer): 29–44.

Lord FM (1980) *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale: Lawrence Erlbaum Press.

Malhotra N (2009) Order effects in complex and simple tasks. *Public Opinion Quarterly* 72(1): 180–198.

Malhotra N, Krosnick JA and Haertel E (2007) The Psychometric Properties of the GSS Wordsum Vocabulary Test. University of Chicago, National Opinion Research Center

Miner JB (1957) *Intelligence in the United States: A Survey*. New York: Springer Press.

Rempel M (1997) Contemporary ideological cleavages in the United States. In Clark TN and Rempel M (eds), *Citizen Politics on Post-Industrial Societies*. Boulder, CO: Westview Press.

Sears DO (1986) College sophomores in the laboratory: influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology* 51(3): 630–635.

Segura G, Hutchings V and Jackman S (2010). *ANES Budget Overview: Presentation at FURNES Meeting*, Rancho Palos Verde, CA.

Shulman HC and Boster FJ (2014) Effect of test-taking venue and response format on political knowledge tests. *Communication Methods and Measures* 8(3): 177–189.

Simon HA (1957) *Models of Man: Social and Rational.* New York: John Wiley and Sons Inc.

Smith TW (1981) *GSS Methodological Report, 19*. Chicago, IL: National Opinion Research Center.

Sudman S and Bradburn NM (1974) *Response Effects in the Electronic Survey*. Chicago, IL: Aldine Press.

Thorndike RL (1942) Two screening tests of verbal intelligence. *Journal of Applied Psychology* 26: 128–135.

Tourangeau R, Rips LJ and Rasinski K (2000) *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Verba S, Schlozman KL and Brady HE (1985) *Voice and Equality: Civic Voluntarism in American Politics*. Cambridge, MA: Havard University Press.

Zaller JR (1992) *The Nature and Origins of Mass Opinion*. Cambridge: Cambridge University Press.

Zhu J and Larry Weiss L (2005) The Wechsler scales. In Flanagan D and Harrison PL (eds), *Contemporary Intellectual Assessment: Theories, Tests, and Issues*. New York: The Guilford Press.

Appendix

IRT approximations and convergence diagnostics

Table 1. IRT parameter estimates for Wordsum scaling.

Wordsum	Pooled	
	Discrimination	Difficulty
Space	−0.379	−1.211
Broaden	−0.608	−1.890
Allusion	−0.739	0.231
Cloistered	−0.360	0.397
Sample size	1010	

Note: Results come from a two-parameter IRT model with a mean of zero and a standard deviation of one using the Ideal function in R (Jackman).

Both modes had a chain length of 1 million with a burn-in of 10,000, thinned by 1000 using the R package Ideal. The mean effective sample size is 988 across all ideal point approximations, found using the coda package, demonstrates that a hypothetical sample size for my chain is 988. This effective sample size demonstrating that the Markov chain Monte Carlo (MCMC) chain ran for a sufficiently long period.

Another way to evaluate MCMC convergence is through examining the traceplots of each respondent. Figure 5 shows a sample of nine respondent traceplots (there are 1010 traceplots, one for each respondent). The goal of a traceplot is to observe no patterns within the chain when converging to an ideal point. I gave a cursory examination to all of the plots, and none have a discernable pattern with each iteration.

Comparing the IRT model with an additive scale

Some researchers will opt to use an additive scale instead of an IRT model. Because of this, I compared an additive scale (contained between 0 and 1) with the IRT scale used throughout this paper. The correlation between these two scales is extremely high, 0.9758. This high correlation demonstrates that both scales measure the same construct of cognitive skill even though each was calculated differently. Moreover, using a *t*-test of means for each mode with the additive scale, I find a *p*-value of 0.40, demonstrating no difference by mode. This shows that my results are consistent across different scaling procedures.

Question wording

In the face-to-face treatment, we included a showcard with each word and corresponding answer choices to match the administration of the GSS. Below is the example provided for all respondents (or read out loud by the interviewer) before the Wordsum test is administered:

> We would like to know something about how people go about guessing words they do not know. On this card is a list of words. Some you may know; others you might be less familiar with. For each question, the first word is in capital letters, for example BEAST. Then there are five other words to choose from. Please select the word that comes closest to the meaning of the capitalized word. For example, if the word in capital letters is BEAST, you would select "animal" because it most resembles the meaning of the word BEAST.

1. Afraid
2. Words
3. Large
4. Animal
5. Separate

And the four Wordsum questions are:
SPACE

1. School
2. Noon
3. Captain
4. Room
5. Board

BROADEN

1. Efface
2. Make level
3. Elapse
4. Embroider
5. Widen

ALLUSION

1. Reference
2. Dream
3. Eulogy
4. Illusion
5. Aria

CLOISTERED

1. Miniature
2. Bunched
3. Arched
4. Malady
5. Secluded

Monitoring cheating in self-completed survey

Even though Google, Wikipedia and an endless amount of reference sites are readily available, only two respondents felt the need to cheat on two *political* knowledge questions but *not* the Wordsum used in this paper. All respondents were told that the survey was *not* contained in a program and was administered over the Internet. Figure 6 displays the browser history of one of the two cheaters: as you can see, the respondent used Google to find out about the Vice President and looked up the UK Prime Minister on Wikipedia. These results suggest that almost all respondents are not motivated to look up answers online.

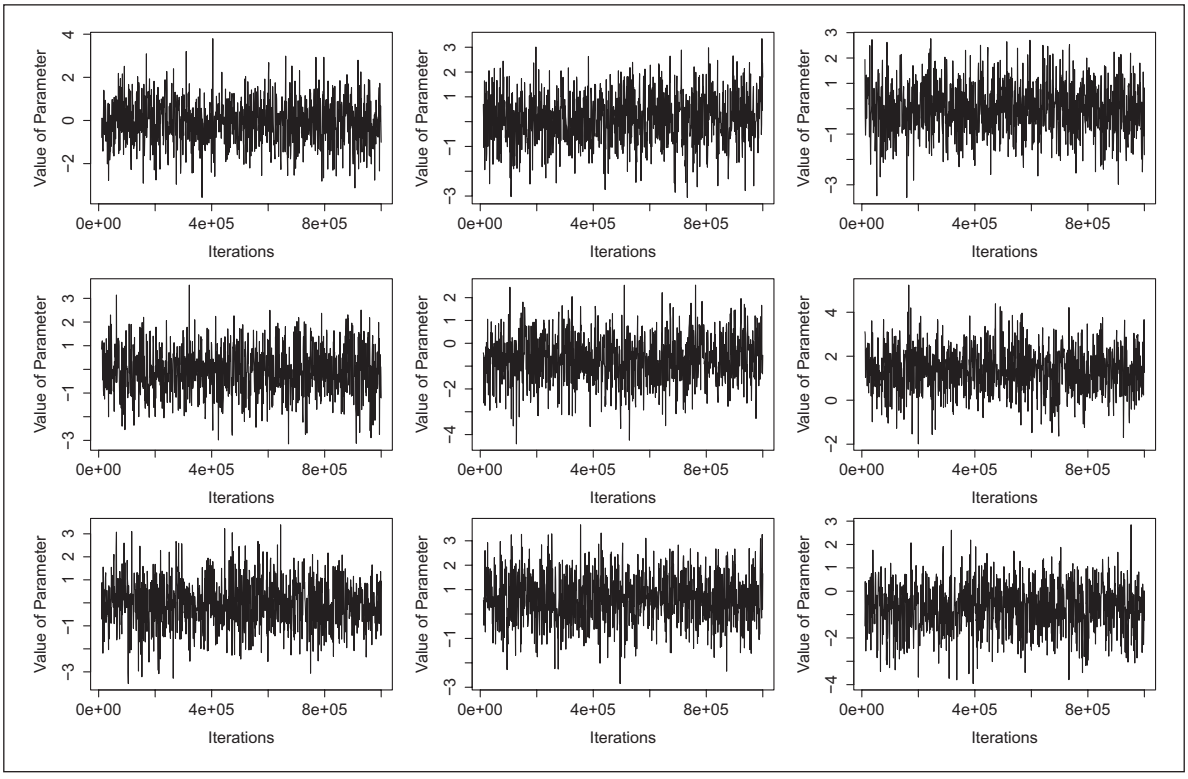

Figure 5. Sample of ideal point traceplots.

Name	Location	Visit Date
vwqPFsSWm0WL9C	https://isurvey-us.yougov.com/thanks/vwqPFsSWm0WL9C	7/28/2011 2:27 PM
YouGov Survey	https://survey-us.yougov.com/vwqPFsSWm0WL9C	7/28/2011 2:15 PM
vwqPFsSWm0WL9C	https://isurvey-us.yougov.com/vwqPFsSWm0WL9C	7/28/2011 2:15 PM
vXfQ7kCl07Wc72	https://isurvey-us.yougov.com/thanks/vXfQ7kCl07Wc72	7/28/2011 2:11 PM
YouGov Survey	https://survey-us.yougov.com/vXfQ7kCl07Wc72	7/28/2011 1:41 PM
vXfQ7kCl07Wc72	https://isurvey-us.yougov.com/vXfQ7kCl07Wc72	7/28/2011 1:41 PM
v9G5g1tzbcHmfg	https://isurvey-us.yougov.com/thanks/v9G5g1tzbcHmfg	7/28/2011 1:40 PM
YouGov Survey	https://survey-us.yougov.com/v9G5g1tzbcHmfg	7/28/2011 1:32 PM
v9G5g1tzbcHmfg	https://isurvey-us.yougov.com/v9G5g1tzbcHmfg	7/28/2011 1:32 PM
vjFWBnwJGmJzgp	https://isurvey-us.yougov.com/thanks/vjFWBnwJGmJzgp	7/28/2011 1:31 PM
vice president - Google Search	http://www.google.com/search?q=vice+president&ie=utf-8&oe=utf-...	7/28/2011 1:28 PM
Prime Minister of the United Kingdom - Wikipedia, the free en...	http://en.wikipedia.org/wiki/Prime_Minister_of_the_United_Kingdom	7/28/2011 1:27 PM
prime minister of uk - Google Search	http://www.google.com/search?q=prime+minister+of+uk&ie=utf-8&o...	7/28/2011 1:27 PM
YouGov Survey	https://survey-us.yougov.com/vcwm65SCdXYM2j	7/28/2011 1:27 PM
vcwm65SCdXYM2j	https://isurvey-us.yougov.com/vcwm65SCdXYM2j	7/28/2011 1:27 PM
YouGov Survey	https://survey-us.yougov.com/vjFWBnwJGmJzgp	7/28/2011 1:19 PM
vjFWBnwJGmJzgp	https://isurvey-us.yougov.com/vjFWBnwJGmJzgp	7/28/2011 1:19 PM
vxvTmbnXYJGDFD	https://isurvey-us.yougov.com/thanks/vxvTmbnXYJGDFD	7/28/2011 1:14 PM
YouGov Survey	https://survey-us.yougov.com/vxvTmbnXYJGDFD	7/28/2011 1:04 PM
vxvTmbnXYJGDFD	https://isurvey-us.yougov.com/vxvTmbnXYJGDFD	7/28/2011 1:04 PM
vtVwdX0TbN31wz	https://isurvey-us.yougov.com/thanks/vtVwdX0TbN31wz	7/28/2011 1:01 PM
YouGov Survey	https://survey-us.yougov.com/vtVwdX0TbN31wz	7/28/2011 12:52 PM
vtVwdX0TbN31wz	https://isurvey-us.yougov.com/vtVwdX0TbN31wz	7/28/2011 12:52 PM
vzFDZy4ZzNnJ8G	https://isurvey-us.yougov.com/thanks/vzFDZy4ZzNnJ8G	7/28/2011 12:47 PM
YouGov Survey	https://survey-us.yougov.com/vzFDZy4ZzNnJ8G	7/28/2011 12:30 PM
vzFDZy4ZzNnJ8G	https://isurvey-us.yougov.com/vzFDZy4ZzNnJ8G	7/28/2011 12:30 PM
vNngXCgbyDFntT	https://isurvey-us.yougov.com/thanks/vNngXCgbyDFntT	7/28/2011 12:29 PM
YouGov Survey	https://survey-us.yougov.com/vNngXCgbyDFntT	7/28/2011 12:11 PM
vNngXCgbyDFntT	https://isurvey-us.yougov.com/vNngXCgbyDFntT	7/28/2011 12:11 PM
vrwhC0dR5L1nlx	https://isurvey-us.yougov.com/thanks/screenout/vrwhC0dR5L1nlx	7/28/2011 11:46 AM

Figure 6. Most self-completed respondents do not cheat on knowledge tests.
Note: Browser history of a respondent that used Google to answer two political knowledge questions. Only 2 of out 507 respondents cheated, and they were removed from the experiment. Both instances occurred on political questions, and not the Wordsum questions used in this paper. The vast majority of respondents do not look up answers even though they are aware of the Internet's availability.