Appendix to: When Deliberation Produces Persuasion Rather Than Polarization: Measuring and Modeling Small Group Dynamics in a Field Experiment

Kevin M. Esterling
Professor
Department of Political Science
UC–Riverside
kevin.esterling@ucr.edu

Archon Fung
Professor and Dean
JFK School of Government
Harvard University
archon_fung@harvard.edu

Taeku Lee
Professor
Department of Political Science
UC – Berkeley
taekulee@berkeley.edu

April 26, 2019

A Appendix

TABLE OF CONTENTS

A.1 Event description

On Saturday, June 26, 2010, nearly 3,000 individuals spent most of the day discussing long term planning for the U.S. federal budget. The event was organized by the non-partisan, non-profit group America*Speaks*, and was called Our Budget, Our Economy (OBOE). The event was held in 19 communities across the United States and was organized specifically to provide citizen input into President Obama's National Commission on Fiscal Responsibility and Reform. The participants were aware that the commission would be briefed on the findings and recommendations that emerged from the event.

AmericaSpeaks made background reading material on the budget and fiscal policy available to potential participants via the website and in hard copy on the day of the event in a document, "Federal Budget 101: An Introduction to the Federal Budget and our Fiscal Challenges," http://usabudgetdiscussion.org/?page_id=17. These reading materials were drafted in consultation with a committee of 30 prominent, ideologically-diverse experts on fiscal policy, who covered the ideological range from very conservative to very liberal and everything in-between.

At the town halls, participants were seated at small discussion tables composed of 8-10 participants and one table facilitator. Participants were given randomized seating assignments, which helped to ensure that participants would encounter others with very different policy preferences and backgrounds.

Participants were charged with working through the technical reading materials and to complete a workbook with 42 policy options (spending cuts and tax increases) with the goal of reducing the deficit by $1.2 trillion in 2025. The options workbook estimated the revenues that would be realized by choosing each option, and outlined the pros and cons for each. The workbook was vetted by the diverse set of policy experts.

Our research team trained 24 field research assistants prior to the event and deployed them to each of the nineteen sites. These research assistants administered two written surveys. The first survey was distributed to participants in their packet of materials and constitutes our pre-event survey; the event organizers directed participants to fill out the survey before the event got underway. The research assistants were provided time at the conclusion of the event to distribute the post-event survey and both the research assistants as well as the event organizers encouraged participants to fill out the post-event survey as an important part of their participation. From the 19 sites, we received 2,576 pre-event surveys and 2,207 post-event surveys. These two rounds of surveys comprise our major source of quantitative data regarding the demographics, attitudes, and assessments of event participants.

A.2 Sample recruitment and characteristics

As we describe next, America*Speaks* did not use random sampling to recruit the participants to the event (as in Fishkin and Luskin, 2005; Luskin et al., 2002). Even if they had, the fact that the organizers had no power to require that those who were sampled actually participated would certainly destroy any randomization because of self selected participation. In this paper we only make statements regarding the in-sample counterfactual comparisons among participants who showed up to the event. In appendix section A.13 below we report the results of a replication study that used data from a different year, on a different policy topic, and that used different recruiting methods, and these results are largely similar to the findings of this paper. The replication provides evidence that the results are likely representative of a deliberative class of citizens who are attracted to this kind of public deliberative event.

Because they believe public deliberation is most constructive when differences of opinions are expressed, America*Speaks* went to great lengths to ensure that the participants were diverse and descriptively representative of their local communities. Their recruitment focused on local organized groups; virtually none of the participants were elite policy insiders. In the weeks leading up to the event, AmericaSpeaks set up a webpage (`http://usabudgetdiscussion.org`) where interested individuals could register to participate. America*Speaks* worked with hundreds of local groups in each of the nineteen localities to recruit a diverse and representative set of participants. They also hired grassroots organizers to recruit diverse participants unaffiliated with the collaborating groups. The registration form asked potential participants a variety of questions, including their age, income, race and party identification. The organizers used the registration database to monitor the representativeness of likely participants, and they targeted invitations to participants in order to preserve representativeness. At each site, if one demographic group appeared underrepresented in the registration database, they contacted local groups who could target and recruit the underrepresented groups most effectively.

For comparison, simultaneous to the event we conducted a random digit dialed (RDD) telephone survey conducted by the survey research firm Interviewing Services of America (ISA). ISA had no involvement with this study except for conducting the telephone survey, and in particular engaged in no communication with America*Speaks* and was not involved in any aspect of the planning for the deliberative events. For the RDD study we drew one sample of 1,929 respondents selected to be nationally representative and an oversample of 748 respondents from the six primary sites that America*Speaks* had selected for hosting large forums (Albuquerque, New Mexico; Chicago, Illinois; Columbia, South Carolina; Dallas, Texas; Philadelphia, Pennsylvania; Portland, Oregon). This sampling frame yields a final sample that includes between 234 and 285 completed interviews in each of these six main cities and a remaining sample of 1,119 respondents drawn from the rest of the United States.

A.2.1 Descriptive characteristics

Here we consider the similarities and differences between OBOE participants, the random-digit dial (RDD) telephone sample, and Census estimates from the 2009 American Community Survey (ACS) in the six primary cities. OBOE and RDD data are weighted to be comparable to the Census American Community Survey (ACS) profiles in these cities. Weights are necessary because some cities (i.e., Chicago, Dallas-Fort Worth, Philadelphia) have substantially larger populations than other cities (i.e., Albuquerque, Columbia, Portland, Oregon). In addition, we also compare the OBOE participants to a survey conducted by Public Agenda of elite Beltway insiders, also on the topic of the budget and long term fiscal policy. The elite survey was conducted by Harris Interactive from February 10 to March 9, 2010. (The Harris sample had an N of 150.) Comparing OBOE participants to this latter sample is useful to see just how different the OBOE participants are from

Beltway insiders who are involved in policy making as a routine matter. Tables 1 and 2 provide the summaries.

First, consider the income distributions reported in Table 1. This table shows that the OBOE participants reasonably approximated the population of these six cities and were more representative than the sample drawn from random digit dialing. Specifically, we find that there is a roughly equivalent proportion of OBOE participants and RDD respondents in the lower income range (less than $50,000) as in the ACS Census data (41 percent in OBOE; 47 percent in RDD; and 44 percent in ACS). It appears that in both OBOE and RDD studies there were fewer participants in the higher income brackets (more than $100,000) than found in the ACS data (20 percent in OBOE; 19 percent in RDD; 24 percent in ACS). The OBOE participants were, as a result, markedly more socioeconomic diverse than policy elites, as shown in the Public Agenda survey who are all relatively wealthy.

Next consider age. Here too there are rough similarities between the OBOE participants, the RDD respondents, and the ACS Census data. The primary difference between the age distribution of OBOE participants and that of the Census of the six primary cities is that OBOE participants were likelier to be in the older age groups (56 percent were aged 55 or older, compared to 27 percent in the ACS in those age categories). OBOE participants were also somewhat less likely than ACS figures to be in the 25-44 year old age groups (27 percent of OBOE participants were in these categories, compared to 33 percent in the ACS; note, ACS data are for 15-24 years, and this reduces the comparability between Census Bureau age ranges and those in our surveys). The RDD telephone sample also substantially underrepresents the youngest adults (aged 18-24) compared to both OBOE and ACS data. In contrast to both the RDD and the OBOE samples, policy elites are typically in the middle age range.

Next consider race/ethnicity. The proportion of whites among OBOE participants in the six cities we examine (71 percent) and in the RDD sample (76 percent) is higher

Table 1: Characteristics of Participants (in percents)

	OBOE	RDD	ACS	Elites
INCOME				
Less than $50K	41	47	44	–
$50-100K	39	33	32	–
More than $100K	20	19	24	100
AGE				
18-24 years	9	6	15	3
25-34 years	9	10	19	14
35-44 years	9	12	19	18
45-54 years	17	20	20	25
55-64 years	28	23	14	34
65+ years	28	28	13	7
RACE/ETHNICITY				
White	71	76	59	86
African-American	17	11	16	8
Latino	5	6	18	0
Asian American	3	2	5	1
Other / Multiple	3	5	2	3
EDUCATION				
H.S. or less	9	26	40	–
Some college	19	28	21	9
College degree	32	24	27	38
Advanced degree	41	22	12	53

than that found in the ACS Census data (59 percent). By contrast, the proportion of African American OBOE participants (17 percent) matches the ACS (16 percent) and the proportion of Latinos is much lower (5 percent in OBOE and 18 percent in ACS). Compared to the RDD telephone sample, we find that OBOE participants are more likely to be African American and less likely to be white. This underrepresentation of Latinos is consistent with other deliberative town hall meetings and, we believe, likely related (at least in part) to language and the predominant use of English at the town halls (though translation services were provided for participants).

We find the biggest demographic differences are in education levels. OBOE participants were without question more educated than the general public. Fully 41 percent of OBOE participants reported having a post-baccalaureate degree, while only 12 percent of the underlying population in the six cities of focus held an advanced degree. Only 9 percent of OBOE participants had a high school degree or less, compared to 40 percent of the six-city Census. On this one measure, the characteristics of the RDD telephone sample sit in between the OBOE and ACS figures: RDD respondents were less educated on the whole than OBOE participants, but more educated than the general population in the six metro areas. Finally, compared to the Public Agenda sample, it is clear that Beltway policy elites are even more highly educated than participants in the OBOE event.

We next consider partisanship, ideology, and level of political interest reported in Table 2. Before discussing what we find, we note a few caveats to these comparisons. First, there are no data that are similar in their quality and generalizeability to Census data with respect to political markers. In this section, we use the 2006 Cooperative Congressional Election Study (CCES), which has the benefit of conducting a large enough number of interviews at the city level to allow us to say something reasonably reliable about political orientation in the six cities we focus on.

Second, with respect to party identification and ideology, we are mindful of the fact that the categories that survey researchers use to label people politically representative are increasingly out of step with a growing number of Americans. Thus in our surveys to both OBOE participants and RDD telephone respondents, we included the option for someone to let us know that they did not think in terms of partisan labels like "Democrat," "Republican," or "Independent" or in terms of ideological labels like "liberal," "conservative," or even "moderate." Not surprisingly to us, a large proportion of individuals chose to tell us these labels are not meaningful to them. Importantly, the CCES asks about partisanship and ideology more conventionally, so these data are not fully comparable.

Third, the event organizers required that we place our party ID and ideology self-placement measures on the post-test, so the measures may reflect changes that occurred during the discussion. We report statistics regarding these measures here under the assumption that these measures are stable features of a respondent's political psychology; we note however that we purposefully do not use these self-placement measures in the statistical model and only rely on pretest measures to construct the ideological ideal point scale at the heart of the model so as to avoid these concerns with measurement validity.

We find that the rank order of Democratic identification being most common, Republican identification least common, and Independents in the middle is common to OBOE and CCES. At the same time, the overlap between OBOE participants and CCES respondents is much closer than either to the RDD respondents. These patterns are roughly similar with respect to ideology as well. A high proportion of people in America today choose not to think in terms of "liberal" or "conservative" labels. That said, OBOE participants were more likely to be liberal and somewhat less likely to be conservative than either RDD or CCES respondents.

The most dramatic difference between OBOE participants and the general population is in their very high degree of interest in politics and public affairs. Whereas only 41 percent of RDD respondents and 50 percent of CCES respondents report that they were "very" interested in politics, fully 81 percent of OBOE participants do so. This difference between OBOE participants and the general public is not surprising. There is little

Table 2: Characteristics of Participants (cont.)

	OBOE	RDD	CCES
PARTISANSHIP			
Democrat	39	28	47
Republican	15	24	21
Independent	23	22	27
Not applicable	24	26	5
IDEOLOGY			
Liberal	32	17	28
Conservative	21	31	24
Moderate	28	28	45
None of these	18	24	2
POLITICAL INTEREST			
Very interested	81	41	50
Somewhat interested	16	39	20
Slightly interested	3	14	5
Not at all interested	1	6	–*

*The CCES has a different set of response categories (only three categories), slightly different question wording, and a significantly higher proportion of respondents who indicated that they were ”not sure” or ”don't know.” The column percentages do not sum to 100 because the remainder (25 percent) are in this category.

reason for someone to volunteer to participate in an all-day event on the federal budget deficit unless one is very interested in the issue and the politics surrounding debates over the budget deficit. This point is most clearly made by comparing our data on OBOE participants to our survey of individuals who registered to participate in OBOE but did not make it to the event (results not shown). The distribution could not be more similar: 80 percent of these ”registered non-participants” report being ”very interested” in politics and a further 17 percent report being ”somewhat interested,” identical to what we find for OBOE participants.

A.2.2 Ideological common space comparison, OBOE and RDD

For a final comparison, figure 1 shows the densities for the ideal point distributions of the OBOE and RDD samples. We estimate these ideal points using the same ideal point estimator described in the main text. We are able to place the OBOE and RDD participants in a common space since we asked identical questions measuring policy preferences for both samples.

Figure 1 shows that, compared to the RDD sample, the OBOE event attracted more centrists relative to moderate-leaning ideologues (noting the higher kertosis of the OBOE distribution), a similar density of extreme conservatives, and a higher density of extreme liberals. Overall, however, Figure 1 shows that the OBOE sample mirrors the range of ideological differences that occur in the population. That is, the OBOE event was not simply an exercise in extreme liberals or conservatives echoing each others' views but instead, given the random assignment procedures was a truly cross-cutting event.

A.3 Sites and survey response summaries

In Table 3 we show the count of participants across the 19 sites in the study. As a part of the event planning, six sites were designated large sites (Chicago, Albuquerque, Portland, Philadelphia, Columbia and Dallas) and the rest were capped at 100 or fewer participants.

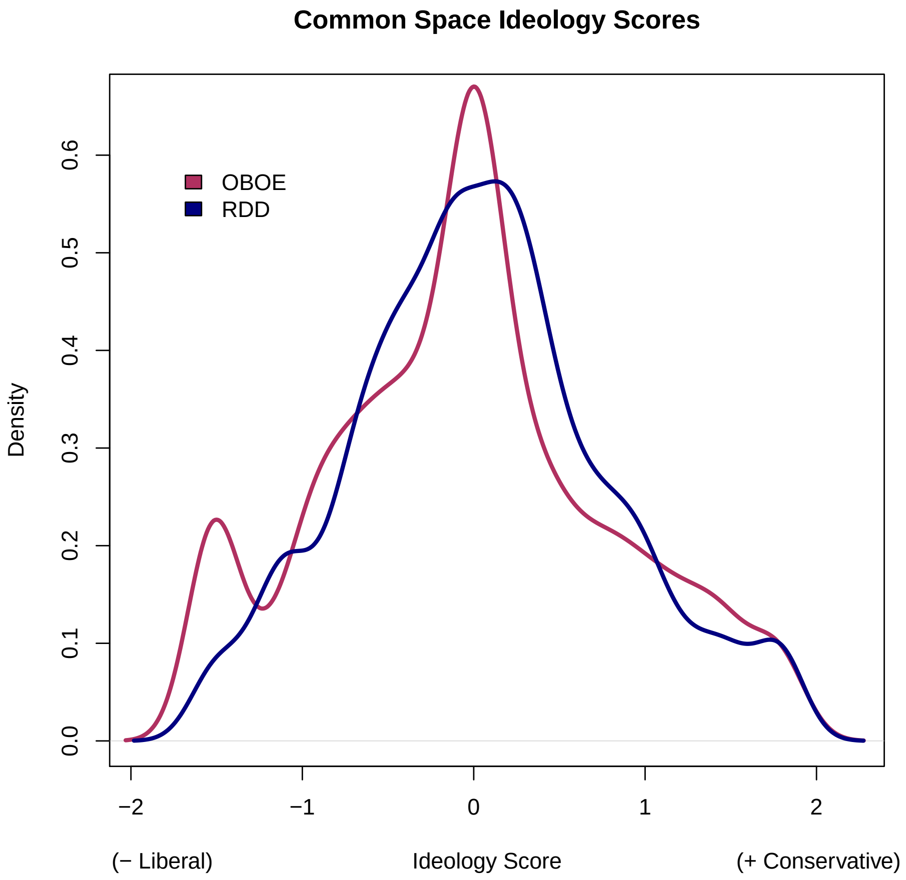

Figure 1: Ideological Common Space Comparison: RDD and OBOE Samples

The organizer's objective was to have 3,000 participants in all.

Table 4 reports the descriptive statistics for the data we use in the main statistical model in the paper. The wording for these questions are given in the data section of the paper. We use the first four items from the pretest (Tax the rich, cut programs, cut entitlements, and cut defense) to estimate participants' latent-space ideal points, as these load on a single left-right dimension. In order to study preference changes we condition on each pretest item in each outcome equation. For each pretest item we create a series of five dummy variables, where the first dummy variable is set to one if the respondent chose the first category on the item, and the other four dummies set to zero, and so on (omitting one category for identification).

Table 3: Event Sites and Number of Participants

Site City	Number of Participants
Pasadena/LA	93
Chicago	383
Des Moines	63
Overland Park	81
Louisville	90
Augusta, ME	60
Detroit	64
Jackson, MS	52
Missoula	59
Portsmouth, NH	110
Albuquerque	200
Grand Forks	21
Portland, OR	403
Philadelphia	303
Columbia, SC	343
Dallas	309
Richmond	75
Caspar	45
Palo Alto	87

Table 4: Summary statistics

Variable	Mean	Std. Dev.	N
Pretest			
Tax Rich	3.855	1.496	2523
Cut Programs	3.155	1.488	2452
Cut Entitlements	2.627	1.559	2499
Cut Defense	3.509	1.514	2500
Tax Rich and Middle Class	2.428	1.334	2482
Federal Sales Tax	2.499	1.389	2446
Post-test			
Tax Rich	3.921	1.476	2111
Cut Programs	3.207	1.568	2075
Cut Entitlements	2.863	1.583	2074
Cut Defense	3.932	1.401	2106
Tax Rich and Middle Class	2.5	1.358	2079
Federal Sales Tax	2.36	1.457	2090

Each item has a five point response scale, with 1 Strongly Disagree to 5 Strongly Agree. In the model, Tax Rich, Cut Defense, Tax Rich and Middle Class, and Federal Sales tax recoded so that 5 is Strongly Disagree, so that the conservative response is larger than (to the right of) the liberal response on each item (with the polarity of each item determined in a descriptive factor model; results not reported).

A.4 Randomization check

America*Speaks* chose to randomize participants to their seating assignments as a way to ensure that there was variation in the composition of participants at tables, and to ensure that people who knew each other (and may enter the event together) would not be seated together. Fortuitously, this randomization allows us to identify the causal effect as participant compositions vary across tables. That is, randomization allows us use the various tables as replicates and counterfactuals for each other.

In the statistical model we have two causal variables based on the ideal makeup of participants seated at each table: the average of the ideal points of the other participants seated at the respondent's table, and the standard deviation of these ideal points. Tables 5 and 6 show the results of balance tests, where in the first table the "treatment" here is a dichotomized variable that equals one if the members of the table other than the respondent are as a group above average for ideology (the respondent is seated at a conservative table) and zero otherwise (seated at a liberal table) and in the second table the "treatment" is one if the respondent is seated at a table with an above average standard deviation and zero if below average. The tables show that covariates measuring attributes of our sample are balanced for both mean and standard deviation, indicating the randomization worked well and that participants complied with their seating assignments.[1]

The omnibus test statistic in each table are estimated using the software of Hansen and Bowers (2008), which compares the joint distribution of the covariates across treatment arms using an omnibus test. Note that both for both treatment variables (the average table ideology and the standard deviation) the test cannot reject the null hypothesis that

[1] Age was systematically related to the standard deviation measure, but since this was only one out of 24 tests we can take that relationship as chance. In Table 6 we omit age and one can see that the remaining variables are balanced, both individually and jointly.

Table 5: Balance test: Mean of Table Ideology

	Liberal Table	Conservative Table	Difference	Null SD	Std. Diff.	Z	
Age	66.22	62.37	-3.85	4.45	-0.03	-0.86	
Female	0.25	0.24	-0.01	0.01	-0.02	-0.80	
Income	3.83	3.83	-0.00	0.05	-0.00	-0.09	
Republican	0.14	0.15	0.02	0.01	0.05	1.60	
Democrat	0.33	0.31	-0.02	0.01	-0.04	-1.35	
Nonwhite	0.20	0.20	0.00	0.01	0.01	0.20	
Education	0.36	0.34	-0.02	0.01	-0.04	-1.28	
Age Missing	0.37	0.36	-0.01	0.01	-0.03	-1.07	
Gender Missing	0.51	0.51	0.00	0.01	0.01	0.32	
Income Missing	0.24	0.22	-0.02	0.01	-0.06	-1.84	.
PartyID Missing	0.16	0.17	0.01	0.01	0.03	0.96	
Race Missing	0.10	0.08	-0.01	0.01	-0.04	-1.17	
Education Missing	0.13	0.12	-0.02	0.01	-0.05	-1.55	

Omnibus balance test (Hansen and Bowers, 2008): $\chi^2 = 14.5(13df)p = 0.341$. Test is stratified by site.

Table 6: Balance test: Standard Deviation of Table Ideology

	Liberal Table	Conservative Table	Difference	Null SD	Std. Diff.	Z
Female	0.27	0.27	-0.00	0.01	-0.00	-0.23
Income	3.88	3.89	0.01	0.03	0.01	0.45
Republican	0.14	0.14	0.00	0.01	0.01	0.55
Democrat	0.34	0.35	0.01	0.01	0.02	0.93
Non-white	0.22	0.23	0.01	0.01	0.02	0.85
Education	0.36	0.37	0.01	0.01	0.02	1.13
Female Missing	0.45	0.45	-0.00	0.01	-0.01	-0.32
Income Missing	0.22	0.21	-0.01	0.01	-0.01	-0.61
Republican Missing	0.16	0.15	-0.01	0.01	-0.03	-1.35
Democrat Missing	0.16	0.15	-0.01	0.01	-0.03	-1.35
Non-white Missing	0.09	0.09	-0.00	0.01	-0.00	-0.13
Education Missing	0.13	0.12	-0.00	0.01	-0.01	-0.56

Omnibus balance test (Hansen and Bowers, 2008): $\chi^2 = 5.73(11df)p = 0.891$.Test is stratified by site. Assignment variable is centered.

the covariates are balanced.

A.5 Implementing the Farrar et al. Regression Model

The current standard method for testing for the causal effect of exposure to a small group discussion is given by β_2 in equation 1, which is found in Farrar et al. (2009). For comparison with our results, we estimate the Farrar model for the data from our application. To implement the model we model each outcome using ordered logit in Stata with standard errors clustered by discussion group. Since the Farrar model applies to each policy question separately, we estimate equation 1 six times, once for each outcome variable. Without adjusting for multiple comparisons, we find the estimate $\widehat{\beta_2}$ significant at conventional levels only for two items, *Tax Rich* $p = 0.00$ and *Cut Defense* $p = 0.00$, while the other four do not reach standard levels of significance: *Cut Programs*, $p = 0.10$ *Tax Both* $p = 0.37$, *Cut Entitlements* $p = 0.09$, and *Federal Sales Tax* $p = 0.72$. Applying the Bonferroni correction we find that the parameter estimate is not significant in the set of equations. In comparison with our model, the lack of precision in the estimates in the Farrar approach is due to modeling changes in the noisly-measured survey responses, rather than to a measured persuasion space.

As a further comparison, we re-estimated the Farrar model, but instead of modeling each item separately, we use the estimated factor scores for the set of four pretest items (*Tax Rich, Cut Defense, Cut Entitlements, Cut Programs*) that load on the latent (ideological) dimension, where we estimate one factor from the set of pretest responses and a separate factor for the set of post-test items. This model using factor scores for the pretest and post-test items is an approximation of the test for our main causal effect of interest, α_1 in equation 7b, although this is only an approximation since using the estimated factor scores as if they were measured scores yields incorrect standard error estimates. We use OLS regression with standard errors clustered by the discussion group. This approximation of 7b yields an estimate $\widehat{\alpha_1}$ that has a ratio of 3.6 to the standard error, which is approximately the same as the ratio we recover from the full model (4.5). While the results are very similar under this implementation of the Farrar model with factor scores as the outcome measures, we do not recommend using this method (see Treier and Jackman, 2013) since it is known that the standard errors when using estimated factor scores are incorrect. In this case, even though the t-ratio is similar in both cases, it is likely that the standard error estimate in the Farrar implementation is an unknown mixture of under-estimated uncertainty from ignoring measurement error and over-estimated uncertainty that comes from an incorrectly specified model.

A.6 A descriptive test of polarization

In addition to the full statistical model, it is worth examining a direct test of whether we observe the law of small group polarization in action among our tables (similar to Luskin et al., 2007). Recall that the law of small group polarization asserts that a group of all liberals will become even more extremely liberal, and a group of all conservatives will become more conservative, over the course of a small group interaction.

A.6.1 Polarization in ideological groupings

To conduct a direct test of this, for now we will ignore the issue of test-retest error, and simply examine differences in preferences pre- and post-discussion. To characterize the ideological composition of the tables, we construct a point estimate for the ideological ideal point for participants by extracting the first principle component using responses to the first four policy questions (Q1 to Q4). These items both have a clear left-right ideological direction on their face, and they all load heavily and uniquely on a single factor. We then identify the set of "homogeneous" tables where everyone seated at the table was on the same side of the centered ideological space using the pre-discussion ideal points.[2] Under

[2] Recall that participants are randomly assigned to tables so participants at these tables should be representative of all participants.

this procedure, we create a variable *Liberal at homogeneous liberal table* that equals one if everyone seated at the table was left of center and zero otherwise, and another variable *Conservative at homogeneous conservative table* that equals one if everyone seated at the table was right of center and zero otherwise.

Of the 339 discussion tables in our study, a total of 24 homogeneous tables emerged from the randomization, with 16 homogeneously liberal and 8 homogeneously conservative. Under the random assignment, table composition is a binomial process, and the probability of a homogeneous table decreases as the size of the table increases. As a result, most of these 24 tables are relatively small, with a median number of participants equal to seven. Further these tables were not distributed uniformly across the sites, but instead most of the all liberal tables were in liberal dominated cities such as Philadelphia and Detroit, and the all conservative tables were mostly in conservative cities such as Casper, Dallas and Columbia.[3] There is no reason to believe, however, that table size or site location would be related to any effect of the law of small group polarization.

We construct a set of difference variables corresponding to Q1 to Q4 by subtracting the pretest response from the post-test response.[4] If the law of small group polarization were in effect at this event, we would expect to see liberals to move toward stronger endorsement of Q1 and Q4 and toward stronger rejection of Q2 and Q3, and vice versa for conservatives, if they are seated at a homogeneous table. We also create an ideology difference variable by subtracting the respondent's pretest ideological ideal point from her ideological ideal point estimated from the same items on the post-test.

We test for polarization at these tables by regressing each of the five difference vari-

[3] We account for site differences using fixed effects in the model below.

[4] We do not use the responses to Q5 and Q6 in this analysis in that these items do not have a clear ideological direction; they both propose new taxes, which liberals might prefer and conservatives might oppose, but these taxes fall on liberal constituencies. Further these items do not load on the ideal point factor.

ables on the two indicators for homogeneous liberal and homogeneous conservative tables. In these regressions we also include a number of control variables to hold constant participants' own attributes. We include scales for both internal[5] and external[6] efficacy, as well as indicators for race (African American, Hispanic, Asian rather than white), education (some graduate education rather than less education), and pretreatment ideology (liberal, conservative rather than moderate). We also include site fixed effects and a random effect for each table.[7] We also estimate reduced regressions leaving out the control variables.

Of the 10 tests of polarization within these models (five outcomes each for liberals and conservatives), in not one equation is the difference statistically different when comparing participants at homogeneous tables and those at non-homogeneous tables. And this null finding is not only a matter of statistical power in that the standard error of these effect estimates are on the order of only 0.1 to 0.2. That is, our data show no sign of small group polarization at this event even with this simple, descriptive analysis.

We do note the possibility of a ceiling effect that may underestimate, but would not eliminate, any effect of small group polarization, in that many of the respondents chose the extreme response on the five point scale that matches their ideological predispositions on the pretest on at least one of the items.[8] Note however that virtually no respondent

[5] "I consider myself well-qualified to participate in politics." "I think I am as well-informed about politics and government as most people."

[6] "Elected officials in Washington, DC don't care about what people like me think." "People like me don't have any say about what the government does." "We can trust the government in Washington to do what is right."

[7] We estimate a random-effects GLS regression in Stata with table-level random intercepts and an $N = 1839$ complete cases.

[8] For taxing the rich, 76 percent of liberals strongly agreed and 32 percent of conservatives strongly disagreed; for cutting social programs 37 percent of liberals strongly disagreed and 54 percent of conservatives strongly agreed; for cutting entitlements 69

chose the extreme category for the full set of items,[9] and as we demonstrate in appendix section A.2, the distribution of participants' ideological ideal points in this town hall closely mirror the distribution of ideal points in a national population sample; the main difference is that this event over-represents ideological moderates. Thus, to the extent there are ceiling effects, these effects would also occur naturally in the population and likely would be larger than what we observe here.

While the first cut analysis is inconsistent with the findings in much of the literature on polarization in small group discussions, we are able to examine this question as well as others more systematically in a full econometric model. The main test of this model moves beyond the literal statement of the law of small group polarization, which only focuses on homogeneous discussion groups. As we show above, homogeneity in discussion groups alone does not drive preference change either toward extremism or moderation.

A.6.2 Polarization in policy agreement groupings

We present a supplemental assessment of the direction of change in respondents' preferences from the pretest to the post-test (that is, preference change rather than persuasion) among tables where participants began the day largely in agreement on specific policy items. To do this supplemental assessment, we identified the set of homogeneous tables for each policy item. To identify homogenous tables in this policy-relevant sense, we selected tables where there were no participants who responded "strongly agreed" or "agreed" on the pretest to a given policy preference item, and the set of tables where no partici-

percent of liberals strongly disagreed and 43 percent of conservatives strongly agreed; and for cutting defense 57 percent of liberals strongly agreed and 37 percent of conservatives strongly disagreed. Thus the best items for this test are cutting programs for liberals and taxing rich, cutting entitlements and cutting defense for conservatives.

[9] Only 8.4 percent of liberals chose the lowest category for each pretest preference item, and no conservatives conservative chose the highest category for each.

pants "strongly disagreed" or "disagreed" on the pretest with a given item. That is, we identified tables where everyone offered either a neutral or a liberal response to a policy preference item, and the tables where everyone offered either a neutral or conservative response.[10] Because of randomization to groups of size 10, we had no tables that contained only participants who only "agreed" or only "disagreed" with the preference item on the pretest (which would reflect moderate liberal or moderate conservative responses to the items, depending on the item), so we must include those who "strongly agreed" or "strongly disagreed" with the item. Including these respondents should bias this test in the direction of even more polarization.

While one would expect some test-retest error, under polarization one should expect to see a tendency for post-discussion responses to be biased in favor of the consensus view at the table. We find, however, no evidence to this effect. We evaluated the percentage of respondents who changed their response in the expected direction relative to all respondents who changed their responses, and tested whether the resulting percentage statistically differed from 50 percent. In this analysis we had a total of eight tests, where there were enough tables of either all liberals or all conservatives on an item to conduct a meaningful test. Among the eight tests, we found four that did not differ from 50 percent; in two tests participants displayed a polarized pattern of greater than 50 percent; and in two tests participants displayed a moderating pattern of less than fifty percent.[11] These

[10] Specifically, a table was retained if everyone either strongly agreed, agreed, or neither agreed nor disagreed; or if everyone either strongly disagree, disagreed, or neither agreed nor disagreed on a given item. We conducted this analysis separately for each item. For this descriptive analysis, we disregard missing observations.

[11] The items where the preference changes were equally in both directions were: liberals on cutting programs (3 tables, 30 participants, 19 changing responses); conservatives on cutting programs (11 tables, 71 participants, 39 changing responses); liberals on increasing the federal sales tax (two tables, 10 participants, 7 changing responses); and conservatives

supplemental results, like the results we present in the main text, are not consistent with any "law" of group polarization.

A.7 Statistical model

As we describe in the text, the statistical model is designed to identify and measure the systematic component of preference change that is due to interpersonal communication. That is, there are many reasons, including test-retest error, for why a respondent would report a different opinion on a post-test compared to a pretest. To identify the systematic interpersonal effect of persuasion, the statistical model relies on spatial methods to capture dependence in preferences among participants seated at the same table, and the model is based on the spatial regression approach described in Congdon (2003, chapter 7). These methods estimate a random effect based on the design structure of participants nested within tables.

The statistical model we use is shown below. In this model, because we estimate the six outcome equations simultaneously, we can nest a portion of the random effect ω_{ik} within the policy preference items and so can estimate the amount of this random effect that is due to changes common to all items, captured in $\Delta\theta_i$. Because this portion of the random effect measures changes on the latent dimension that explains the full set of preferences (Poole and Rosenthal, 1997), we label this component latent-space persuasion.

on increasing the federal sales tax (17 tables, 138 participants, 65 changing responses). Items that showed a polarized pattern were: liberals on taxing the rich (49 tables, 385 participants, 119 changing responses); and liberals on cutting defense spending (23 tables, 174 participants, 61 changing responses). And the items that showed a moderating pattern were: conservatives on taxing the middle class as well as the wealthy (21 tables, 163 participants, 86 changing responses); and liberals on cutting entitlements (18 tables, 117 participants, 58 changing responses).

The residual of this random effect, $\Delta\zeta_{ik}$, which is specific to each item, we label topic-specific persuasion. These are the two systematic components of persuasion that we can model directly.

Likelihood:

$$\left.\begin{array}{rcl}
O_{post,ik} & \sim & \text{OrderedLogit}(\boldsymbol{\beta_{1k}}\mathbf{O_{pre,ik}} + \beta_{2k}\theta_i^0 + \boldsymbol{\beta_{3k}}\mathbf{Site_i} + \omega_{ik}),\\
\omega_{ik} & = & \Delta\theta_i + \Delta\zeta_{ik}\\
RaiseTaxes_i & \sim & \text{OrderedLogit}(\theta_i^0)\\
CutPrograms_i & \sim & \text{OrderedLogit}(\lambda_2\theta_i^0)\\
CutEntitlements_i & \sim & \text{OrderedLogit}(\lambda_3\theta_i^0)\\
CutDefense_i & \sim & \text{OrderedLogit}(\lambda_4\theta_i^0)\\
\theta_{ij}^0 & \in & \left\{\theta_j^0 : j \text{ is seated at } i\text{'s table}, j \neq i\right\}\\
H_i & = & \text{mean}(\theta_{ij}^0) = \sum_j(\theta_{ij}^0)/(N_i^-)\\
S_i & = & \text{mean}([\theta_{ij}^0]^2) - \text{mean}(\theta_{ij}^0)^2\\
\Delta\theta_i & \sim & \phi(\Delta\theta_i^*, 1)\\
\Delta\theta_i^* & = & \alpha_1 H_i + (\delta_1 \cdot Liberal_i + \delta_2 + \delta_3 \cdot Conservative_i) \cdot H_i^2\\
& & +(\gamma_1 \cdot Liberal_i + \gamma_2 + \gamma_3 \cdot Conservative_i) \cdot S_i\\
& & +\delta_4 \cdot Liberal_i + \delta_5 \cdot Conservative_i\\
\Delta\zeta_{ijk} & \in & \left\{\Delta\zeta_{jk} : j \text{ is seated at } i\text{'s table}, j \neq i\right\},\\
\Delta\zeta_{ik} & \sim & \phi(\Delta\zeta_{ik}^*, 1)\\
\Delta\zeta_{ik}^* & = & (\rho_{1k} \cdot Liberal_i + \rho_{2k} + \rho_{3k} \cdot Conservative_i) \cdot \sum_j(\Delta\zeta_{ijk})/(N_i^-)\\
N_i^- & = & \#\left\{\text{participants sitting at } i\text{'s table, not including } i\right\}
\end{array}\right\}
\begin{array}{c} 1 \leq k \leq K\\ 1 \leq i \leq N\end{array}$$

i indexes N participants
j indexes i's N_i^- discussion partners
k indexes K policies

Priors:

The prior distributions for $\alpha_.$, $\delta_.$, and $\gamma_.$ are each Uniform(-0.25, 1) due to a constraint in the model, where the sum of each parameter type is bounded by the min/max eigenvalue of the normalized adjacency matrix formed by the table assignments for each observation. The priors for $\rho_.$ are distributed Uniform(-1, 1) to ensure bounds for the correlations. The factor coefficients in the θ_i^0 scale are distributed Uniform(0, 100) in order to ensure the correct direction labeling in the factor model. All other priors are unrestricted and flat.

The θ_i^0 factor is estimated from the pretest responses to the *Tax rich*, *Cut programs*, *Cut entitlements*, and *Cut defense* items, where the factor is estimated dynamically within the model (summarized in the likelihood above for simplicity of presentation). All of the policy preference items are recoded so that high numbers indicate a conservative response, as indicated in a factor model (results not shown). We define the factor coefficient the

the equation for each pre-treatment response on items Q1 to Q4 as $\{1, \lambda_2, \lambda_3, \lambda_4\}$. Since all $\lambda_.$ are estimate as positive, that means that movement to the right along the latent dimensions $\{\theta, H\}$ are in a conservative direction. We estimate $\rho_.$ separately for each of the six policy preference items. To constrain $\Delta\theta_i$ to the underlying ideological space, and to ensure identification, we constrain $\{\alpha_., \delta_., \gamma_.\}$ to be equal across all six items.

In the model the estimated covariances between the pre- and post-treatment response on each item is given by $\{\beta_{2k}, \beta_{2k}\lambda_2, \beta_{2k}\lambda_3, \beta_{2k}\lambda_4\}$, for Q1 to Q4, respectively. In effect, θ_i^0 is the portion of the total error component of the model, $\beta_{2k}\theta_i^0 + \omega_{ik} + \epsilon_{ik}$ (where ϵ_{ik} is the non-systematic error component) that accounts for and partials out dependence between O_{ik}^0 and O_{ik}^1.[12] The remaining error, including ω_{ik}, is conditionally independent of O_{ik}^0 for a given value of θ_i^0.

We can take $\omega_{ik} = \Delta\theta_i + \Delta\zeta_i$ as a valid measure of interpersonal persuasion provided that ω_{ik} is uncorrelated with the included predictors in equation 6a (Skrondal and Rabe-Hesketh, 2004, p. 50). This assumption is met on its face with $Site_i$ since this covariate is fixed and it is implausible that respondents would travel to a different city in response to anything endogenous to our study. This assumption also is met for the O_{ik}^0 covariate since the model includes a covariance parameter that captures any dependence between the full error terms of O_{ik}^0 and O_{ik}^1, including ω_{ik} and its components. Since we use the pretreatment policy preference items (Q1 to Q4) to measure the respondent's ideological ideal point, including the common latent variable θ_i^0 in the equations for both pre- and post-treatment responses captures all dependence between the pre- and post-treatment response (Skrondal and Rabe-Hesketh 2004, 107-8).

[12] The statement that θ_i partials out dependence does not hold for questions Q5 and Q6. We instead justify the validity of ω_{ik} for these two equations under the more common but stronger assumption that pretreatment values for O_{ik} are fixed and not endogenous to the design.

A.8 Estimation

We run the model until the posterior distribution of the structural estimates are stationary, and then sample from the joint posterior distribution to create marginal distributions of each parameter of interest. The pretest variables have missing data rates ranging from 9 percent to 28 percent, and the post-test variables have missing data rates around 25 percent. We impute the missing post-test data as missing at random given the observed and latent variables and we impute the missing pretest variables as missing at random conditional on the participant's site (Raghunathan, 2004). The model estimates incorporate the additional uncertainty that is due to the missing data, which are imputed as full distributions (Tanner and Wong, 1987). In addition, we conduct sensitivity tests to bound the range of our effect estimates given extreme values of the missing data (Gerber and Green, 2012, 226) in appendix section (A.12).

A.9 Benchmarking the model using simulated data

In the replication materials, we include the code and a tutorial on how to implement the model as well as simulated data that we use to test whether the model yields results that match the benchmark parameters. To create the simulated data we use the following data generating process, which is a slightly simplified version of the model we estimate in the paper. In the simulated data, we draw five pre- and post outcomes from a normal distribution, where the pretest outcomes are standard normally distributed and post-test outcomes are distributed normally with unit standard deviation and conditional mean function:

$$O_{ik}^{1} = \beta_{0k} + \beta_{1k} O_{i}^{0} + \beta_{2k} \theta_{i}^{0} + \beta_{3k} Site_{i} + \Delta\theta_{i} + \Delta\zeta_{ik} + \epsilon_{ik}, \text{ for } k = 1 \text{ to } 5. \tag{1}$$

$$\Delta\theta_i \sim \phi(\Delta\theta_i^*, 1), \tag{2a}$$

$$\Delta\theta_i^* = \alpha_1 H_i + (\delta_1 \cdot \mathit{Liberal}_i + \delta_3 \cdot \mathit{Conservative}_i) \cdot H_i^2, \tag{2b}$$

$$H_i = \text{mean}(\theta_{ij}^0), \tag{2c}$$

$$\theta_{ij}^0 \in \left\{\theta_j^0 : j \text{ is seated at } i\text{'s table}, j \neq i\right\}. \tag{2d}$$

$$\Delta\zeta_{ik} \sim \phi(\Delta\zeta_{ik}^*, 1), \tag{3a}$$

$$\Delta\zeta_{ik}^* = \rho_k \cdot \text{mean}(\Delta\zeta_{ijk}), \tag{3b}$$

$$\Delta\zeta_{ijk} \in \{\Delta\zeta_{jk} : j \text{ is seated at } i\text{'s table}, j \neq i\}. \tag{3c}$$

Table 7 lists the parameters for the DGP and the estimates that result from our model from the single simulated data set that we distribute with the tutorial. For parameters that are indexed by question we report the median of the five estimates.

Table 7: Benchmark and Estimates, Simulated Data

	Benchmark	Estimate	Standard Error
β_0 (median)	1.4	1.42	0.12
β_1 (median)	0.4	0.40	0.04
β_2 (median)	1.0	0.40	0.07
$\beta_{3[1]}$ (median)	-0.5	-0.49	0.15
$\beta_{3[2]}$ (median)	0.5	0.44	0.15
α_1	1.0	0.88	0.14
δ_1	1.0	1.35	0.46
δ_3	-1.0	-1.46	0.52
ρ (median)	–	0.57	0.09

N = 1000, number of groups = 100, 10 participants each

Note the benchmarks are based on regression results using the benchmark latent variable values, rather than estimated scores, using the Stata function "xtreg" and clustering by group.

Given the sample of 1000 units all of the results are statistically different from zero. Overall, the model does well hitting the benchmarks for the quantities of interest in the model with the exception of the scaling coefficient on the pretreatment ideal point score (β_2), which is due only to a change in the scale of the estimated ideal point dimension. This scale does not have an intrinsic scale so the scaling parameter is not of interest, and is not related to the purpose of including the ideal point in capturing the endogenous correlations between the pre- and post-preference measures. The structural parameters of interest all hit the benchmarks well. A bootstrap across two additional replications of the simulated data set yielded identical results.

For comparison, we also estimate these models in Stata using the "xtreg" regression function, clustered by group, but using factor score estimates for the pre- and post treatment ideal points. As Treier and Jackman (2013) explain, using factor scores as if they were observed scores introduces an errors in variables bias to all of the parameters of the regression, since factor scores are estimates and so have inherent uncertainty that is not represented in the score. The estimates for the structural parameters related to the factor score variables, β_2, δ_1 and δ_3 are each about half the magnitude of the benchmark values; attenuated parameters are characteristic of regression coefficients with measurement error in the right hand side variables. In contrast, the parameters in the full Bayesian model that propagate the uncertainty through the model recover the benchmark values, once adjusting for the change in scale for the pretreatment ideal points.

A.10 Code, replication data, and video tutorial

We distribute the full data set and the material to replicate all of the analyses presented in this paper at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/XDDN2Y. In addition, in the replication material we include two videos and a set of files that serve as a tutorial on how to estimate the model, along with the simulated data, a Stata file to generate replicates of the simulated data, and general code to implement the model. The model code we provide accommodates continuous, dichotomous and ordinal outcomes for the pre-post items; it allows any number of outcome items (although there must be at least three items to identify the latent scale); and for the ordinal model it accommodates any number of response categories for the Likert scale greater than three.[13] Finally, the tutorial includes a set of files to preprocess the data to implement each of

[13] If one has items with responses with only three categories, one can either recode the categories into dichotomous variables to use the available code for the dichotomous model, or revise the ordinal model code in the likelihood function directly.

these types of models.

A.11 Correlates of persuasion

Here we drill deeper into the validity of our measure of persuasion, ω_i, as a measure of deliberative persuasion by considering the correlates of the expectations of the two components of ω_i. These correlations are not causal but can provide a descriptive sense of types of interaction that lead to preference change, and hence we can consider whether persuasion occurs within discussions that could be labeled "deliberative."

First, we can gain a sense of the nature of latent-space persuasion in this context by examining the correlates of the extent of this persuasion for each participant. We measure the extent of each participant's latent persuasion as the estimated individual-level latent persuasion random intercepts, $\Delta\theta_i$, both in direction and in magnitude. We computed the expected persuasion random intercept for each participant (i.e., the point estimate for $\Delta\theta_i$), and used these expected values as a dependent variable in supplemental regressions as a means to assess the descriptive correlation between this measure of latent persuasion and several scales that measure participants' own perception of the nature of the discussion.

To do the supplemental regression, we construct three scales that measure each participant's own perception of the nature of the quality of the discussion at the event.[14] First, we have a set of indicators on the post-test survey that measure how *Informative and Reasoned* each participant perceived the discussion to be. These items ask if the participants "Strongly agree," "Somewhat agree," "Neither," "Somewhat disagree," or "Strongly disagree" to the following questions: "I am more informed about the chal-

[14] We use principal components factor analysis and the full set of discussion-quality items to construct these three scales. The factor model produces this three factor solution (results not reported).

lenges and options for cutting the federal budget deficit;" "The meeting today was fair and unbiased. No particular view was favored;" "I personally changed my views on the budget deficit as a result of what I learned today;" "I personally agree with the voting results at the conclusion of today's meeting;" and "Decision makers should incorporate the conclusions of this town meeting into federal budget policy."

Second, we have a set of post-test indicators that measure how *Civil* each perceived the discussion to be. These questions were, "People at this meeting listed to one another respectfully and courteously;" "Other participants seemed to hear and understand my views;" "Even when I disagreed, most people made reasonable points and tried to make serious arguments;" and "Everyone had a real opportunity to speak today. No one was shut out and no one dominated the discussions."

Third, we have post-test indicators of how *Enjoyable* each found the discussion. These questions were, "I had fun today. Politics should be like this more often;" "I would participate in an event like this one again;" and "Participating today was part of my civic duty as an American to speak out and be heard on this issue."[15]

These scales measure participants' own perceptions of the nature of the discussion at the event, and so are useful in assessing the nature of discussion where ideological persuasion is most prevalent. For example, if participants changed their minds simply because they were intrigued by the charismatic personalities of their co-discussants, we would likely find that preference changes are most likely to occur when participants simply enjoyed the discussion or found the discussion to be civil. In contrast, if participants are most likely to be persuaded when they perceive the session to be informative and reasoned, this would suggest that persuasion occurs in a more rational, evidence-based discourse, and hence, in the presence of deliberation (Barabas, 2004). Note that these correlations are not causal, in that these measures of the nature of the discussion and the outcomes

[15] While the duty item may not fit an enjoyableness factor on its face, the item loads very highly on this scale empirically.

are all taken from the post-test, but they are useful because they are descriptive of the nature of the relevant interactions and in this sense provide a construct validity check of the rationality of persuasion at the event (Cook and Campbell, 1979).

We employ regression models that we describe in appendix section A.6. In the regression modeling the magnitude of latent persuasion, none of the coefficients reached conventional levels of statistical significance. In the model of the direction of latent persuasion we find that the informative discussion rating scale was positively associated with persuasion in the liberal direction for both moderates and conservatives, but not liberals. Specifically, moderates who rated the discussion one standard deviation above average for being informative shifted their latent preference 22 percent ($p = 0.001$) of a standard deviation in the liberal direction. Conservatives who rated the informativeness of the discussion one standard deviation above average shifted their latent preference 37 percent of a standard deviation ($p < 0.001$) *in the liberal direction.* The point estimate for liberals was nearly identically zero and not significant (standard error = 0.06).

By comparison, Republicans who rated the discussion of average informativeness shifted their latent preferences 44 percent of a standard deviation ($p = 0.08$) in the conservative direction, and Democrats who rated the discussion of average informativeness shifted 17 percent of a standard deviation ($p = 0.06$) in the liberal direction. These results suggest that informed liberal arguments at this event tended to have cross-cutting appeal, while less informed arguments tended to drive participants in the direction of their preconceptions.

In contrast, none of the other scales that characterize either the nature of the discussion (civility or enjoyableness) nor the efficacy scales (internal, external) were correlated with this measure of latent persuasion. In addition, none of the other demographic variables were related either to the direction or magnitude of shifts on the ideological dimension, after accounting for partisanship. That the informativeness of the discussion alone is predictive of latent persuasion suggests that, by this self-measured assessment, the persuasion we observe can be characterized as deliberative.

To examine topic-specific persuasion, we compute the expected value for our measure of topic-specific persuasion (the mean of the marginal posterior distribution of $\Delta\zeta_{ik}$) for each participant for each item, and use these measures as dependent variables in six supplemental regressions, with identical specifications to the analogous regressions for the ideological component above. We regress the direction of the policy-specific persuasion on a set of variables and report these results in table 8. Table 9 shows the results for the magnitudes (which is the absolute value of the of the random effect). The cells in each table indicate standardized regression coefficients, which show the association between dependent and independent variables in standard deviation units.

We find that for both direction and magnitude, the only consistently significant correlate with topic-specific persuasion is the informative and reasoned discussion scale. Some of the items show correlations with the efficacy scales and with race indicators, but these results are not consistently significant (with the exception that African Americans seem to be less persuadable to agree with most items, both liberal and conservative).[16]

Instead, the perceived informativeness of the discussion is the only variable that is consistently associated with non-ideological topic-specific persuasion. In addition, in the direction models, the sign of each coefficient indicates that participants who believe that the discussion is informative tend to be persuaded in the direction of moderation and toward the common goal of reducing the deficit: increasing taxes and reducing spending. That is, if the respondent perceives the discussion to be informative she is more likely to be persuaded to increase taxes and to cut programs and entitlements. In other words, respondents who believed the discussion to be informative tended to move their topic-

[16] We do not have evidence that this effect from this race indicator might be due to an unobserved race ideological dimension structuring the discussion, since interacting the African American indicator with the three discussion quality scales yields results indistinguishable from zero.

Table 8: Correlates of Topic-Specific Persuasion: Direction

	Tax Rich	Cut Programs	Cut Entitlements	Cut Defense	Tax Both	Federal Sales Tax
Discussion Ratings						
Informative	*-0.13***	*0.05***	*0.11***	*-0.02*	*-0.14***	*-0.10***
	(0.03)	(0.02)	(0.04)	(0.03)	(0.04)	(0.02)
Civil	-0.02	0.02	-0.06*	0.02	-0.03	-0.01
	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)	(0.02)
Enjoyable	0.01	0.01	-0.02	-0.05*	0.11**	0.03
	(0.03)	(0.02)	(0.04)	(0.03)	(0.04)	(0.02)
Self-Efficacy Scales						
Internal	0.02	0.03*	-0.03	0.04	-0.02	-0.04**
	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)
External	0.07**	-0.03*	0.06*	-0.05**	-0.04	0.00
	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)
Individual Attributes						
Black	0.13**	0.01	-0.37**	0.14**	0.01	0.13**
	(0.06)	(0.05)	(0.09)	(0.07)	(0.08)	(0.06)
Hispanic	0.08	-0.18**	0.01	0.03	0.28**	0.13
	(0.10)	(0.09)	(0.14)	(0.11)	(0.14)	(0.09)
Asian	-0.12	0.05	0.24	-0.19	0.11	0.01
	(0.13)	(0.11)	(0.18)	(0.14)	(0.18)	(0.12)
Grad School	-0.09*	0.03	-0.01	-0.01	-0.05	0.05
	(0.05)	(0.04)	(0.06)	(0.05)	(0.06)	(0.04)
Constant	-0.08	-0.00	0.11	0.20	-0.05	-0.14
	(0.27)	(0.33)	(0.21)	(0.28)	(0.23)	(0.30)

$^{*} * p \leq 0.05$, $^{*}p \leq 0.10$

Dependent variables are the topic-specific random effect point estimates taken from the corresponding equation in the statistical model; low values of the dependent variable indicate shifts in the liberal direction and high values indicate shifts in the conservative direction. Cell entries are standardized coefficients from a single-equation random effect model in which the clusters are defined by small group discussion tables (OLS estimates give substantively identical results). Fixed effects from income categories not reported (few effects were significant).

N = 1467, number of tables = 327

Table 9: Correlates of Topic-Specific Persuasion: Magnitude

	Tax Rich	Cut Programs	Cut Entitlements	Cut Defense	Tax Both	Federal Sales Tax
Discussion Ratings						
Informative	*0.05**	*0.03*	*0.08***	*0.03*	*0.06**	*0.09***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Civil	0.02	0.03	-0.04	0.03	-0.02	-0.00
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Enjoyable	-0.02	0.02	0.05	0.02	0.02	0.03
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Self-Efficacy Scales						
Internal	0.05*	-0.01	0.02	0.06**	-0.02	0.01
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.02)
External	-0.04	0.05**	-0.01	-0.04	0.01	-0.00
	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)	(0.02)
Individual Attributes						
Black	0.14*	0.06	0.12	0.10	0.03	-0.05
	(0.07)	(0.07)	(0.08)	(0.08)	(0.08)	(0.07)
Hispanic	0.23*	-0.03	0.02	-0.03	-0.13	-0.12
	(0.12)	(0.11)	(0.13)	(0.12)	(0.12)	(0.11)
Asian	-0.15	-0.17	-0.00	0.05	-0.27**	-0.09
	(0.15)	(0.14)	(0.17)	(0.16)	(0.16)	(0.15)
Grad School	-0.01	-0.02	0.02	0.01	-0.11**	0.02
	(0.05)	(0.05)	(0.06)	(0.05)	(0.05)	(0.05)
Constant	0.01	-0.14	-0.08	0.30	0.22	0.38
	(0.21)	(0.28)	(0.15)	(0.22)	(0.16)	(0.25)

$^{*}p < 0.05$

Dependent variables are the absolute value of the topic-specific random effect point estimates taken from the corresponding equation in the statistical model. Cell entries are standardized coefficients from a single-equation random effect model in which the clusters are defined by small group discussion tables (OLS estimates give substantively identical results). Fixed effects from income categories not reported (no effects were significant)

N = 1467, number of tables = 327

specific preferences in the direction of solving the collective problem of the future national debt and deficit.

The magnitudes of these correlations, pooled across liberals, independents, and conservatives, are quite small. Pooling across ideological categories assumes that participants are equally susceptible to opinion change on all of the items, but this might not be sensible in that liberals and conservatives are likely to have different responsiveness to a deliberative exchange depending on the nature of the policy option under consideration.

Table 10 examines the size of the correlation between the informed discussion scale and topic-specific persuasion when disaggregating by ideological subgroups, for both direction and magnitude. We note two findings in this table. First, the size of the correlations increase over the pooled model in those conditions where the correlations are significant. This finding is consistent with the proposition that the persuadability of liberals and conservatives differs depending on the policy under consideration.

Second, a very interesting pattern emerges in terms of which ideological category is most susceptible to non-ideological, topic-specific persuasion across the full set of policies. In considering the correlation between informativeness and the direction of preference change, notice that liberals are most likely to be persuaded in an informed discussion to agree with conservative policies (cut programs and cut entitlements), conservatives are most persuaded to agree with a liberal policy (tax rich) and liberals and conservatives are equally persuaded on the two policies that are orthogonal to the ideology scale (tax middle class and rich, and the federal sales tax) in the direction of raising taxes. This table strongly indicates that the dynamics at these events are consistent with deliberative expectations, in that 1) topic-specific persuasion was most likely to occur when participants perceived the discussion to be informative, and 2) that within these discussions, liberals and conservatives were each persuaded to moderate on, and accept the merits in, policies that are favored by the other side and that would contribute to solving a pressing national problem.

Table 10: Correlation of Topic-Specific Persuasion with Informative Discussion, by Ideology

	Tax	Cut	Cut	Cut	Tax	Federal
	Rich	Programs	Entitle-	Defense	Both	Sales Tax
			ments			
Direction						
Liberal	-0.05	0.10**	0.24**	0.05	-0.18**	-0.10**
	(0.05)	(0.04)	(0.07)	(0.05)	(0.06)	(0.04)
Moderate	-0.02	0.03	0.12	-0.05	-0.14*	0.08
	(0.05)	(0.05)	(0.08)	(0.06)	(0.07)	(0.06)
Conservative	-0.20**	0.04	0.02	-0.03	0.09*	-0.09**
	(0.04)	(0.03)	(0.05)	(0.04)	(0.05)	(0.03)
Magnitude						
Liberal	0.05	0.16**	0.14**	0.13**	0.07	0.16**
	(0.05)	(0.06)	(0.05)	(0.05)	(0.05)	(0.06)
Moderate	-0.00	0.10	0.14**	-0.11*	0.11	0.06
	(0.08)	(0.07)	(0.07)	(0.06)	(0.07)	(0.07)
Conservative	0.12**	0.06	0.02	-0.09	0.05	0.09*
	(0.05)	(0.05)	(0.05)	(0.05)	(0.04)	(0.05)

$^{*} * p < 0.01$ $^{*}p < 0.05$
Dependent variables are 1) the topic-specific random effect point estimates and 2) the absolute value of these estimates, taken from the corresponding equation in the statistical model. Cell entries are standardized coefficients from a single-equation random effect model in which the clusters are defined by small group discussion tables (OLS estimates give substantively identical results), identical to the previous except adding main and interactive effects of ideological ideal point categories. Fixed effects from income categories not reported (few effects were significant)
N = 1467, number of tables = 327

We have three reasons by which we can assert this topic-specific persuasion is outside of ideology. First, the model controls for both the individual's own ideological ideal point as well as ideological influences from interacting with co-discussants at a table. Second, as figure 3makes clear, the degree of dependence does not vary among liberals, moderates and conservatives for any of the items. Third, we analyze the expected degree of topic-specific persuasion ($\widehat{\Delta\zeta_{ik}}$) for each participant and for each of the six preference items, and in this analysis we find that these random effect estimates have only a minuscule correlation with the ideological ideal points scale, ranging from -0.08 to 0.06. In addition, we do not observe any site-level factors that explain this measure of topic-specific persuasion; regressing the site dummies on each estimated $\widehat{\Delta\zeta_{ik}}$ vector shows only one site (Silicon Valley) that had a non-zero relationship with the random effect for only two items. This site was very small (n=87), however, and is only one of 19 sites and thus its deviation from zero is consistent with sampling variability.

A.12 Missing data sensitivity checks

In the model we impute missing post-test data as distributions under an assumption of missing at random (taking each missing data point as a parameter to estimate with uncertainty) conditional on the respondents' pretest response, her ideology, site fixed effects, and the ideal points of other participants seated at her table. In the analysis, if a subject has a missing post-test but filled out a pretest, we impute a posterior distribution for their post-test response as missing at random conditional on their pretest policy preference responses for that same item and their latent ideal point. Since the pretest response and the latent preferences are extremely predictive of post-test responses, a missing at random assumption is well justified for this imputation. The pretest response on the item as well as the respondent's ideal point are extremely predictive of the post-test response and hence make the missing at random assumption strongly defensible for those who filled out a pretest but failed to fill out a post-test. Imputing the missing data as full distributions incorporates the full uncertainty in the estimate, under the missing at random assumption, into the statistical model (Tanner and Wong, 1987).

There are a handful of respondents, however, who filled out a post-test but not a pretest. Since the model requires estimates of the latent ideal points of each respondent in order to calculate the table-level mean and standard deviation, we cannot drop these respondents from the sample. We cannot fully rely on a missing at random assumption for missing pretest data, however, as the only prior information we have on these respondents is their site, which is not highly predictive of pretest responses. Hence we conduct a sensitivity analysis to identify bounds for extreme assumptions regarding the distributions for these missing observations.

In the main analysis, we present results that treat these respondents as missing at random, conditional on site fixed effects. In addition, we conduct a sensitivity analysis of the missing at random assumption. To do this we re-estimate the model twice. In the first re-estimation, we impute the missing pretest data under the assumption that the respondents who failed to fill out a pretest were drawn from an unusually liberal distribution (with mean of this distribution set to one standard deviation below the mean for all respondents). In the second re-estimation, we do the same but set the missing data distribution to unusually conservative. This supplemental analysis identifies the bounds for the results reported in the main paper (which imputes missing pretest responses at random conditional on the site indicators) under 1) the assumption that the missing responses were drawn from an underlying extreme liberal distribution (i.e., only liberals failed to fill out the pretest) and 2) were drawn from an underlying extreme conservative distribution (only conservatives failed to fill out the pretest).

Figures 2 and 3 show the results of these sensitivity tests. As is apparent, there results are unchanged and so robust to different distributions of the missing data. The likely reason is that there are simply not enough missing observations to affect the results in any way, even if the missing data really had been drawn from extreme distributions.

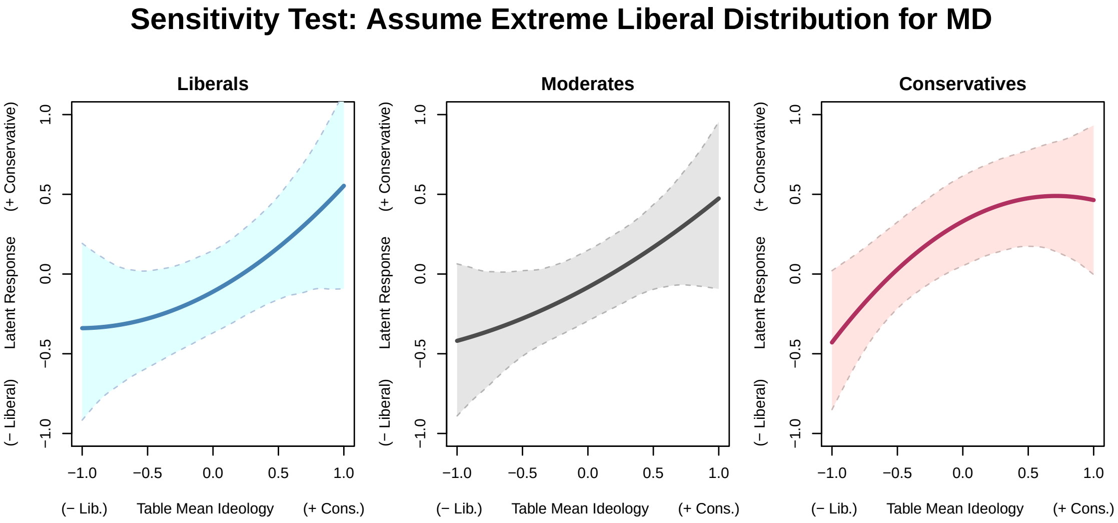

Figure 2: Sensitivity Test: Assume a Liberal Distribution for Missing Data

Finally, any subject who refused to fill out either survey is not available for the analysis. Since the total number of respondents who filled out at least one survey is virtually exactly the number of people who attended the event, we can reasonably ignore this possibility.

A.13 Replication study

As we describe above, the sample in this study is entirely self-selected and hence is not representative of a known population. Self-selection is not a threat to the internal validity of the findings, but does raise questions regarding the study's external validity. Fortuitously, America*Speaks* hosted a similar event in California in 2007, on the topic of health care reform. The design of the event was very similar to the OBOE event[17] and the data are very useful as a replication as 1) the California study occurred three years prior to

[17] One exception is that instead of using a simple randomization for seating assignments the organizers used a variant of sequential systematic sampling. We describe elsewhere (results not shown) that the sequential assignment method resulted in complete balance in a manner similar to the simple randomization used in the present design.

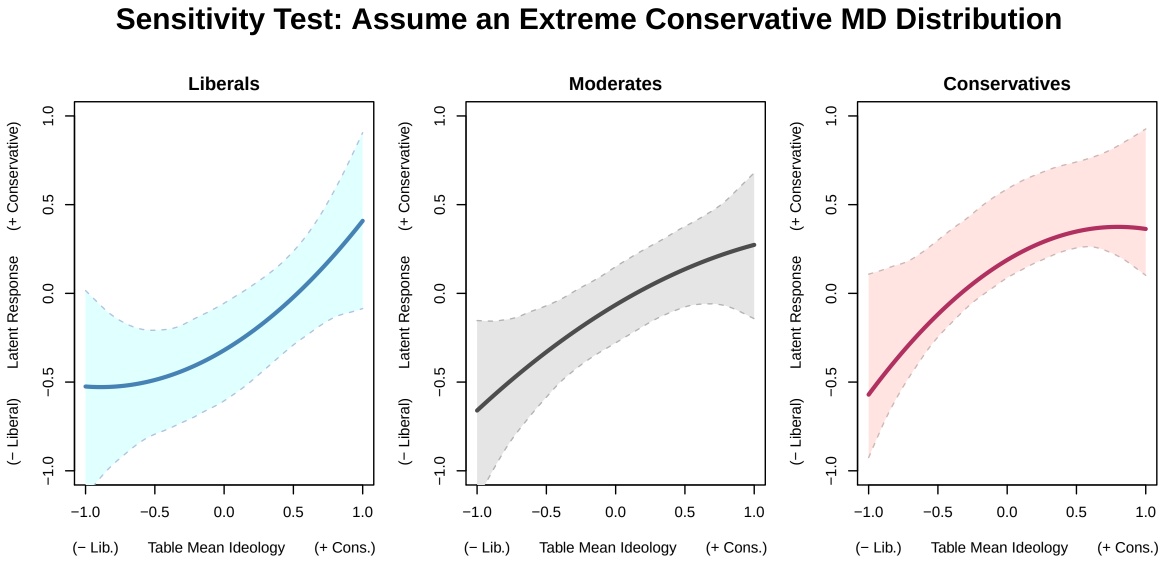

Figure 3: Sensitivity Test: Assume a Conservative Distribution for Missing Data

the OBOE study, 2) was limited to eight cities in California,[18], 3) were recruited in part through a survey research firm using randomized methods, and 4) the topic was on health policy instead of fiscal policy.

The health policy data are somewhat more complicated in that the five outcome items do not load on a single dimension. Two of the policy preference items load on the same scale as ideology and party self-reports, so these two items fit into the standard left-right ideological space. These are:

- Limit government's role to providing insurance coverage for the low income or unemployed, or those who can't get insurance on their own (five point agree/disagree scale)
- Fundamental change to insure all Californians through a state-administered system that all Californians and their employers pay into (five point agree/disagree scale)

[18] Only four of the sites in the California study had complete compliance with seating assignments: Riverside, San Luis Obispo, Sacramento and Eureka, so we limit the replication to these sites.

Three policy preference items did not load on this dimension, so these items to not fit in the ideological space. These are:

- Expand coverage by working with employers to cover more working people and families (five point agree/disagree scale)
- All Californians should receive a health care voucher or tax credit, to be used to purchase their own coverage (five point agree/disagree scale)
- Health insurance companies should be required to offer affordable coverage plans to everyone, regardless of their health condition (five point agree/disagree scale)

Because there were two distinct dimensions to these data, we modify the statistical model to estimate "ideological persusion" on these two dimensions. For simplicity, in the replication study we label the first dimension the "ideological" dimension and the second "non-ideological."

Figures 4 to 7 show the results for the causal portion of the statistical model. Note that the results are virtually the same, particularly showing no evidence for small group polarization at this event and instead the same linear or diminishing effect of increasing the number of co-ideologues at one's table. We do not observe the same pattern of motivated reasoning, however, primarily because the results are not statistically significant. The signs of the slope change across the two figures, but this is consistent with sampling error. These results suggest that the pattern regarding motivated reasoning in the OBOE sample, which also failed to reach significance, is also likely a result of mere sampling error.

These results for the replication study strongly demonstrate the external validity of the causal results we obtain in the OBOE study.

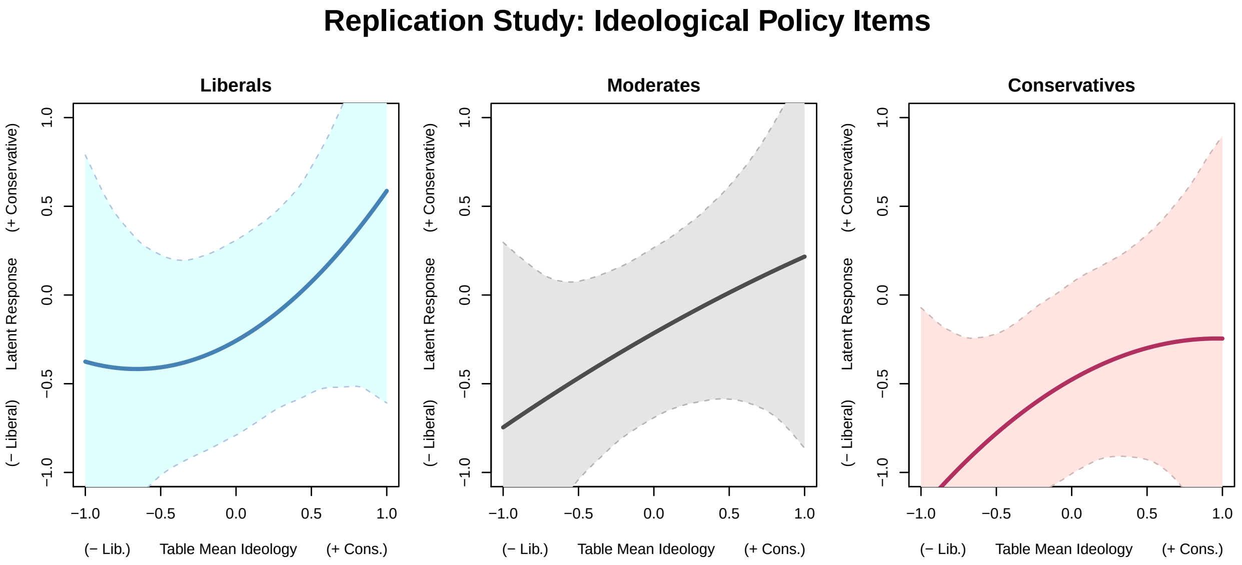

Figure 4: Replication Study: Mean Composition Effect on Ideological Dimension

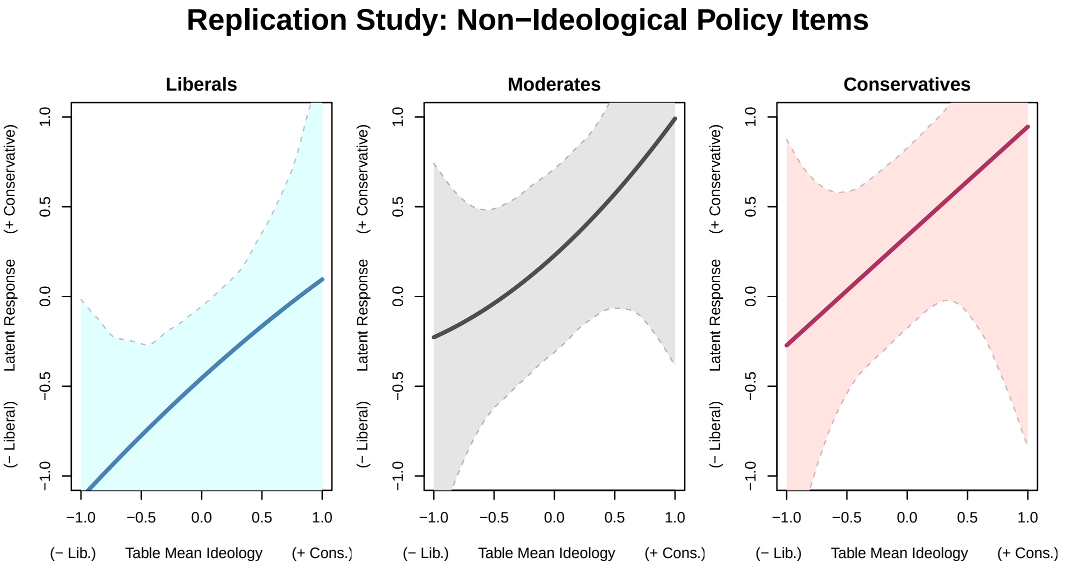

Figure 5: Replication Study: Mean Composition Effect on Non-Ideological Dimension

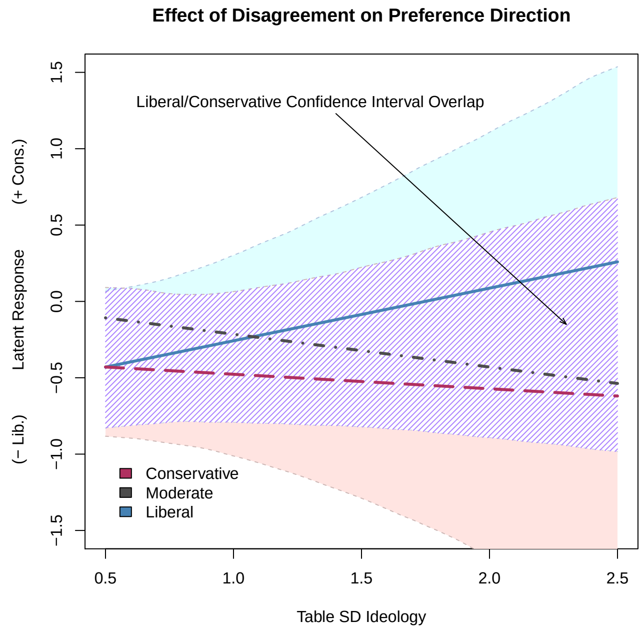

Figure 6: Replication Study: Disagreement Effect on Ideological Dimension

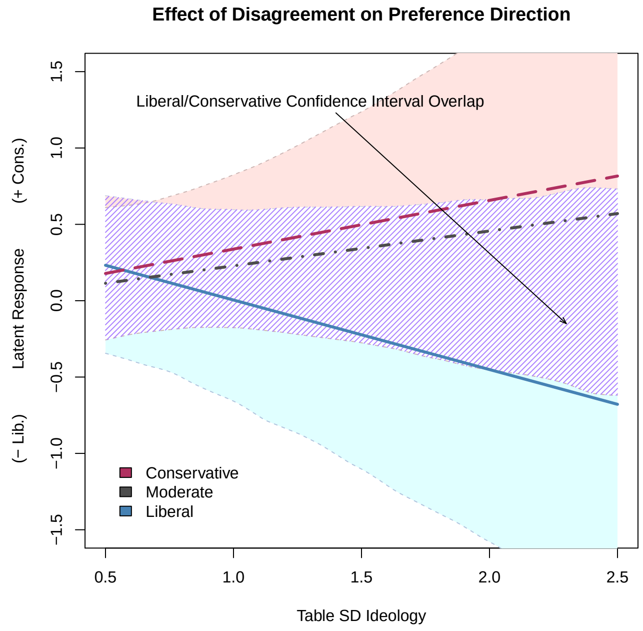

Figure 7: Replication Study: Disagreement Effect on Non-Ideological Dimension

A.14 Methodological FAQs

This section gives brief answers to questions we have encountered regarding the statistical model.

What is the benefit of this modeling approach to measuring persuasion over simpler approaches such as a pre-post difference in the opinion response?

We argue that simpler approaches to modeling preference change, such as relying on a pretest-posttest opinion difference, is not methodologically or conceptually defensible given the extent of noise contained within a survey response. As we describe in the text, the raw difference between the posttest and pretest opinion contains an unknown amount of measurement error and hence the raw difference score does not map onto persuasion as a construct. We derive this thesis from a fundamental statement of the survey response itself, which parallels the decomposition in the recent (Lauderdale et al., 2018) paper. Our modeling approach improves on the Lauderdale paper in that it demonstrates how to model preference change in response to an intervention (such as a group discussion) and identifies new quantities for measuring preference change that are likely to be of interest to the small group literature, as well as to any study that examines preference change over time in response to a randomized intervention.

Why does this model rely on pretreatment ideal points of the discussion partners as the intervention rather than the arguments that are actually made during the discussion?

Our research design assigns participants to compositions of groups at the discussion tables, and in the language of experimental design, the table composition is a randomized encouragement design to expose participants to different types of arguments. The discussion that happens over the course of the event occurs post-treatment, that is, after participants are seated at their table for the interaction. We cannot identify the causal effect of the discussion itself since this is a "mechanism" or "mediator" that occurs post-treatment beyond the assignment to the table composition, and hence a statistical test based on some measure of arguments will lack internal validity (Imai et al., 2011). Instead, our paper limits its findings to those statements that we designed to be internally valid, and the encouragement design is well-understood in the experimental methods literature. In our application, the "encouragement" only needs to assume that the pre-discussion ideal points is predictive of the kinds of arguments participants are likely to make in the discussion. This assumption would be satisfied, for example, if there is a larger mix of conservative arguments made at a discussion where most of the participants have conservative pre-discussion ideal points compared to tables where most of the participants have liberal ideal points, and vice versa.

There was only one randomization, but participants are randomly assigned to two things: the mean and the standard deviation of ideal points.

Randomization to groups assigns participants to the distribution of ideal points at that table, and distributions are characterized both by a mean and a standard deviation. This is no different from a random assignment that assigns participants to two different factors in a two-factor model. The mean and SD as randomized quantities enter the model as ordinary regressors that predict change in the latent preferences. As we note in the paper, the mean and standard deviation are independent both theoretically and empirically so there is no identification problem either in assigning both through the randomization, or by including both as regressors on the right-hand side.

Doesn't including the pretreatment opinion response create endogeneity bias in the model?

In the model description in the text that leads up to equation (5), we show how the model accommodates possible endogenous dependence between the pretreatment and the post-treatment outcome measures by modeling the latent correlation. This is a standard method to allowing for dependence between pretreatment and post-treatment responses and we provide cites in the text to support that. We note too that this problem is often ignored and pretreatment opinion is often taken as exogenous in the experiments literature, which is not methodologically defensible.

How do you handle missing data?

As we describe in the text, we use multiple imputation to impute posterior distributions for missing post-test responses, where the imputation is conditioned on both the pretreatment opinion response for the item as well as the respondent's pretreatment ideal point, both of which are extremely predictive of post-treatment response. Our method incorporates the estimation uncertainty in the post-test response imputations and propagates that uncertainty to the statistical model. Imputation is standard in the modeling literature and is superior to alternative methods such as listwise deletion. We note though that a small number (nine percent) of respondents failed to fill out a pretest, and we do not have data to reasonably impute these responses. In the main text we simply use the mean response at the participants' site to condition the imputation, and then in appendix section A.12 we provide sensitivity tests to show that this imputation does not affect the estimated quantities of interest at all.

How do you separately identify the latent space parameter (θ_i^t) and the topic-specific parameter (ζ_i^t)?

The complex structure of group data separately identifies the θ_i^t and ζ_i^t parameters. For example, in our application, we identify θ_i^t by nesting questions within participants, and ζ_i^t by nesting participants within discussion groups.

Does the shrinkage in the Bayesian estimator of θ_i^t attenuate the treatment effect estimates?

Overall, shrinkage (also known as partial pooling) of the latent space parameter estimate should have little effect on the treatment effect estimate, since the treatment effect in the latent space is given in standard deviation units and those units are the same (under randomization) across the different group compositions. It is possible that shrinkage reduces the influence of observations that are in the tails since the scale compression will be greater the farther an observation is from the mean, but as we demonstrate in appendix A.2, there are few observations located in the tails, and any bias that could result would only lead to more conservative estimates of the treatment effect.

References

Barabas, J. (2004). How Deliberation Affects Public Opinion. *American Political Science Review 98*(Nov.), 687–701.

Congdon, P. (2003). *Applied Bayesian Modelling.* Hoboken, N.J.: John Wiley & Sons, Ltd.

Cook, T. D. and D. T. Campbell (1979). *Quasi-Experimentation: Design and Analysis for Field Settings.* Chicago, Ill.: Rand-McNally.

Farrar, C., D. P. Green, J. E. Green, D. W. Nickerson, and S. Shewfelt (2009). Does Discussion Group Composition Affect Policy Preferences? Results from Three Randomized Experiments. *Political Psychology 30*(4), 615–647.

Fishkin, J. S. and R. C. Luskin (2005). Experimenting with a Democratic Ideal: Deliberative Polling and Public Opinion. *Acta Politica 40*(Sept.), 284–298.

Gerber, A. S. and D. P. Green (2012). *Field Experiments: Design, Analysis and Interpretation.* New York, N.Y.: W.W. Norton.

Hansen, B. B. and J. Bowers (2008). Covariate Balance in Simple, Stratified and Clustered Comparative Studies. *Statistical Science 23*(2), 219–236.

Imai, K., L. Keele, D. Tingley, and T. Yamamoto (2011). Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies. *American Political Science Review 105*(Nov.), 765–789.

Lauderdale, B. E., C. Hanretty, and N. Vivyan (2018). Decomposing Public Opinion Variation into Ideology, Idiosyncracy, and Instability. *Journal of Politics 80*(2), 707–712.

Luskin, R. C., J. S. Fishkin, and K. S. Hahn (2007). Consensus and Polarization in Small Group Deliberation. Technical report, Presentation at the Midwest Political Science Association, Chicago, Ill. https://www.researchgate.net/publication/253165745.

Luskin, R. C., J. S. Fishkin, and R. Jowell (2002). Considered Opinions: Deliberative Polling in Britain. *British Journal of Political Science 32*, 455–487.

Poole, K. T. and H. Rosenthal (1997). *Congress: A Political-Economic History of Roll Call Voting*. New York, N.Y.: Oxford University Press.

Raghunathan, T. E. (2004). What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data. *Annual Review of Public Health 25*, 99–117.

Skrondal, A. and S. Rabe-Hesketh (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, Fla.: Chapman and Hall.

Tanner, M. A. and W. H. Wong (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association 82*(398), 528–540.

Treier, S. and S. Jackman (2013). Democracy as a Latent Variable Democracy as a Latent Variable. *American Journal of Political Science 52*(1), 201–217.