UC RIVERSIDE | UNIVERSITY OF CALIFORNIA | School of Public Policy

UCR School of Public Policy Working Paper Series

February 2018 – SPPWP18-02

TABLE OF CONTENTS

Measuring and Modeling Persuasion within Small Groups, with an Application to a Deliberative Field Experiment on U.S. Fiscal Policy.

Kevin M. Esterling
UC Riverside, School of Public Policy

Archon Fung
Harvard University, Kennedy School of Government

Taeku Lee
UC Berkeley, Department of Political Science

Summary:
We propose a new statistical method to measure persuasion within small groups, and apply this measurement method to a large scale randomized deliberative experiment. We define the construct ``persuasion'' as a change in the systematic component of an individual's preference, separate from measurement error, that results from exposure to interpersonal interaction. Our method separately measures persuasion in latent (left-right) preference space and persuasion in a topic-specific preference space. The functional form of our model accommodates tests of substantive hypotheses found in the small group literature. We illustrate the the measurement method with an application wherein we examine how changes in participants' policy views on U.S. fiscal policy in a large-scale randomized deliberative experiment resulted from the composition of the small discussion groups to which they were randomly assigned.

This project was funded by a very generous grant from the John D. and Catherine T. MacArthur Foundation. We thank Jason Barabas, Peter Esaiasson, Thomas Leeper, Robert Lupton, John Ryan, Georgia Warnke, and seminar participants at the Center for the Study of Stratification and Inequality at Tohoku University, and at UC Center Sacramento, for very valuable comments. We gratefully acknowledge the collaboration with and support from Carolyn Lukensmeyer, Joe Goldman, and all of the staff who worked at AmericaSpeaks.

Modeling Persuasion within Small Groups, with an Application to a Deliberative Field Experiment on U.S. Fiscal Policy *

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

June 21, 2018

*This project was funded by a very generous grant from the John D. and Catherine T. MacArthur Foundation. We thank Jason Barabas, Peter Esaiasson, Thomas Leeper, Robert Lupton, John Ryan, Georgia Warnke, and seminar participants at the Center for the Study of Stratification and Inequality at Tohoku University, and at UC Center Sacramento, for very valuable comments. We gratefully acknowledge the collaboration with and support from Carolyn Lukensmeyer, Joe Goldman, and all of the staff who worked at America*Speaks*.

Abstract

We propose a new statistical method to measure persuasion within small groups, and apply this measurement method to a large scale randomized deliberative experiment. We define the construct "persuasion" as a change in the systematic component of an individual's preference, separate from measurement error, that results from exposure to interpersonal interaction. Our method separately measures persuasion in latent (left-right) preference space and persuasion in a topic-specific preference space. The functional form of our model accommodates tests of substantive hypotheses found in the small group literature. We illustrate the the measurement method with an application wherein we examine how changes in participants' policy views on U.S. fiscal policy in a large-scale randomized deliberative experiment resulted from the composition of the small discussion groups to which they were randomly assigned.

1 Introduction

Persuasion is central to any conception of democratic political communication (Broockman and Kalla, 2016; Minozzi et al., 2015; Mutz et al., 1996). For example, one of the core tenets of deliberative democracy (Gutmann and Thompson, 1996) holds that preferences among debate participants should be responsive to arguments, at least on occasion. The possibility of noncoercieve persuasion is central to Gutmann and Thompson's (1996, 52) conception of "reciprocity," and Habermas's (1984, 9) conception of "communicative action." When debate participants recognize merits in each others' claims, policy agreements possess legitimacy beyond that gained from majority rule voting (Cohen, 1989).

We propose a novel method for modeling persuasion within small-groups, a method that is applicable when assignment to groups is randomized. Our method measures the extent to which individual preference change is caused by exposure to interpersonal interactions within a small group, after netting out measurement error. We partition measured persuasion into two components: *latent persuasion* which is the amount an individual changes on an underlying, left-right dimension that structures preferences across a set of policy items, and *topic-specific persuasion* which is the amount an individual changes preferences on a given topic, such as a policy option, net of latent preferences (similar to Lauderdale et al., 2018). Randomization is the key to identifying both of these components of persuasion; without randomization the model results are likely to be driven by confounding through self-selection processes.

We demonstrate this method in an application where we test for the causal effects of exposure to small-group discussion on persuasion at the "Our Budget, Our Economy" nationwide town hall meetings organized by America*Speaks*, an event where nearly 3,000 participants were randomly assigned to small group discussion tables. The event was held on June 26, 2010 at town halls in 19 separate cities, with between 100 and 500 participants in each town hall. Within each town hall, participants' seating assignments were randomized among small group discussion tables, and we administered opinion surveys both before and after the event. We use this application to demonstrate our novel measurement strategy for persuasion within small groups, and to assess the extent and nature of persuasion that occurred at this event. Substantively, we show that the amount and nature of persuasion we observe meets many of the normative aspirations of deliberative democracy.

2 Measuring and Modeling Persuasion

The standard approach to measuring persuasion in the small group literature evaluates a change in a discussion participant's self-reported preferences from before to after a discussion event (e.g., Grönlund et al., 2015; Schkade et al., 2010; Westwood, 2015). In a basic small-group design, the researcher typically will administer a survey to participants before exposure to the group to measure each participant's pretreatment preference on an item or topic, which we will label O_i^0. Next, the researcher will randomize assignment for each participant to a small group. This randomization varies the composition of the group to which each participant is exposed. For example, randomization will vary the distribution of ideological ideal points within a group, so randomly assigns each participant to a group that is on average either liberal or conservative (or anything in-between), and that is either diverse in ideology or homogeneous.[1] The groups are invited to have a discussion and after the discussion the researcher will measure the respondents' post-treatment preferences, O_i^1.

In the basic design, the researcher will conduct a statistical test to see if there is a

[1]Formally, respondents are assigned to distributions of ideal points; since the mean and variance are properties of a distribution, respondents are randomly assigned to both dimensions. In this small-group design, the variance is a function of the mean, but since this function is nonlinear, the mean and variance of assignments will be uncorrelated in expectation.

relationship between group exposure and the difference between the preteatment and post-treatment response, $(O_i^1 - O_i^0)$. Farrar et al. (2009, 619) is an exemplar of current practices, which models preference change in response to exposure to a small group discussion as

$$O_i^1 = \beta_0 + \beta_1 O_i^0 + \beta_2 H_i + \beta_3 Site_i + \epsilon_i \tag{1a}$$

$$H_i = \frac{1}{n_i - 1} \sum_j O_j^0, \quad j \in \{J_i : j \text{ is seated at } i\text{'s table}, j \neq i\} \tag{1b}$$

where the i^{th} respondent's post-treatment preference on a given topic (O_i^1) is modeled as a function of her own pre-treatment preference (O_i^0), the average (H_i) of the pre-treatment preferences of her $(n_i - 1)$ discussion partners (indexed by $j \in J_i$), and separate intercepts for each location $(Site_i)$ where the discussions were held. One can confirm that Farrar et al. (2009) model preference change as the difference in pre-post survey responses by subtracting $\beta_1 O_i^0$ from both sides of equation (1).[2]

In the Farrar et al. (2009) study, as in our own application, the respondent is randomly assigned to discussion groups so the average of the pre-treatment preferences of her discussion partners (H_i) is also random, and under the normal assumptions for identifying a causal effect in a randomized control trial that we describe in more detail below, β_2 identifies the causal effect on the respondent's change in response to the survey item from the pretest to the post-test, $(O_i^1 - \beta_1 O_i^0)$, that comes from exposure to a discussion group with a given composition of participants (see also Gastil et al., 2008; Klar, 2014).[3]

[2]Including the pre-treatment response in the model as a right-hand-side variable identifies the β_1 coefficient, which allows the scale of the preference item to change over time. One can constrain $\beta = 1$ to set the scales equal.

[3]As we discuss more extensively below, the model tests for the causal effect of exposure to a given composition of participants in the discussion group, which is randomized in the study design, rather than exposure to the discussion itself, which is not randomized. Pre-treatment preference is an instrument for what participants say in discussion and so

The difference in pre-post survey response, however, does not map onto persuasion as a construct because the pretest and post-test responses each contain a stochastic component from measurement error (Achen, 1975; Ansolabehere et al., 2008; Prior, 2010), in addition to a systematic component that captures respondents' preferences at a given time. Only a change in the systematic component that results from some intervention, such as interpersonal interactions within a discussion, should count as a valid measure of persuasion; random noise should not.[4]

To formalize the systematic component for preference change, for simplicity assume a continuous, normally distributed opinion response at time t, O_i^t, and decompose the opinion response as

$$O_i^t = \beta_0 + \theta_i^t + \zeta_i^t + \epsilon_i^t, \quad t \in \{0, 1\} \tag{2}$$

where θ_i^t is the respondent's latent, left-right "ideal point" that structures preferences across a range of issues (Hinich and Munger, 1994),[5] ζ_i^t is a topic-specific preference that remains after netting out latent preferences, and ϵ_i^t is the idiosyncratic component from measurement error that represents instability in the individual's opinion response (Lauderdale et al., 2018), all evaluated at time t; $t = 0$ is the pretest and $t = 1$ is the

the model identifies the complier average causal effect of exposure to a discussion (see Angrist et al., 1996).

[4]Note that this definition of persuasion is not limited to rational persuasion (Habermas, 1984); in the application below we demonstrate methods to assess the nature of persuasion including its rationality using the concept of construct validity.

[5]When the scale has a left-right orientation, the institutional literature labels this latent preference as the respondent's "ideology," and typically ideology can be scaled using a single dimension (Clinton, 2012; Poole and Rosenthal, 1997). The assumption of unidimensionality is not necessary and the below model can accommodate an arbitrary number of dimensions through a more elaborate design.

post-test. If θ_i^t and ζ_i^t are invariant or fixed over time, then opinion change is driven only by the idiosyncratic component and is essentially noise.

The statistical task is to separate out systematic preference change in θ_i^t and ζ_i^t from random noise through a measurement strategy, and then to model the two systematic components directly. To derive a model of preference change over time from first principles, we can take the difference in equation (2) between time $t = 1$ and $t = 0$,

$$O_i^1 = \beta_0^1 + \theta_i^1 + \zeta_i^1 + \epsilon_i^1 \tag{3a}$$

$$\beta_1(O_i^0 = \beta_0^0 + \theta_i^0 + \zeta_i^0 + \epsilon_i^0). \tag{3b}$$

Subtracting equation (3b) from equation (3a) and rearranging yields,

$$O_i^1 = \beta_0 + \beta_1 O_i^0 + \Delta\theta_i + \Delta\zeta_i + \epsilon_i \tag{4}$$

where $\beta_0 = \beta_0^1 - \beta_1\beta_0^0$ and $\epsilon_i = \epsilon_i^1 - \beta_1\epsilon_i^0$. With this derivation we have identified two new quantities, $\Delta\theta_i = \theta_i^1 - \beta_1\theta_i^0$ which is the change in the respondent's pre- to post-discussion preferences in the *latent preference* space, and $\Delta\zeta_i = \zeta_i^1 - \beta_1\zeta_i^0$ which is the change in the respondent's *topic-specific preference* for the outcome represented by O_i after accounting for changes in latent preferences. This derivation allows us to focus on these more substantively interesting preference changes, rather than only on the noisily measured changes in the survey response itself. We define *measured persuasion* as the change in the systematic components of the respondent's expressed preference.

Consider the two systematic components of preference change in turn. First, one can consider latent preferences to be a heuristic, such as left-right ideology, that enables individuals to make sense of and engage in policy debates involving complex matters even with limited information (Eatwell, 1993; Hinich and Munger, 1994). In this interpretation, $\Delta\theta_i$ captures changes in the latent structuring of their preferences that organizes their views across a range of policies. In the American context, by-and-large ideology reduces to a single, latent dimension (Poole and Rosenthal, 1997). For example, in the context of our application on U.S. fiscal policy that we describe below, as an empirical matter all preferences load exclusively on a single latent dimension captured by θ.

At the same time, the structure of preferences within specific policy topics can be complex (Feldman and Johnson, 2014; Treier and Hillygus, 2009), and particularly at the elite level or within deliberative communication (Gutmann and Thompson 1996, 56; Habermas 1984, 99) reasoning about policy topics is not strictly constrained by ideology or to any other single latent dimension (see Tausanovitch and Warshaw, 2017). Such an assumption would be overly restrictive and indeed a gross oversimplification of human cognition. For example in the town hall event we study, participants were provided policy reading material and expert testimony to inform discussions, and so had the capacity to give reasons and exchange rationales that go beyond a heuristic defined by ideology. In this view, the $\Delta\zeta_i$ measure of topic-specific persuasion captures the amount of persuasion that occurs "outside" of the latent scale.

Thus, within a small-group event, persuasive processes can operate at these two different levels. Note that this partitioning between $\Delta\theta_i$ and $\Delta\zeta_i$ does not create a hierarchy among latent and topic-specific reasoning. In the statistical model, the relative amount of each can vary freely across individuals.

As is common practice (e.g., Farrar et al., 2009), we allow the scale of the response space to vary over time by multiplying both sides of equation 3b by β_1. For example, $\beta_1 < 0$ implies a plenary shift in preferences toward moderation and $\beta_1 > 0$ implies a plenary shift toward extremity. Note that in the case of both $\Delta\theta_i$ and $\Delta\zeta_i$, the change in systematic preferences is based on the underlying preference space rescaled by β_1. One can fix the scales across the two time periods by setting $\beta_1 = 1$, which is equivalent to modeling the difference $(O_i^1 - O_i^0)$ as an outcome (such as in Westwood, 2015).

In general, including an outcome response variable measured pretreatment, such as O_i^0, on the right hand side will lead to endogeneity bias since many of the individual-level determinants of an outcome in the pretreatment period also determine the outcome in the post-treatment period. To see why in the case of modeling preference change, define $\omega_i^t = \theta_i^t + \zeta_i^t$, and note that $cov(\omega_i^0, \omega_i^1) \neq 0$, since $\theta_i^1 = \theta_i^0 + \Delta\theta_i$ and so θ_i^0 is contained in both ω_i^0 and ω_i^1. In the statistical model below we correct for this by including θ_i^0 in the outcome equations. In essence, we guard against endogeneity bias under the assumption that the latent preference scale is a strong predictor of both pre- and post-discussion preferences, and that the remaining variation in preferences O_i^0 and O_i^1 is random once a respondent's ideal point is accounted for.[6] Thus, the equation we estimate is,

$$O_i^1 = \beta_0 + \beta_1 O_i^0 + \beta_2 \theta_i^0 + \Delta\theta_i + \Delta\zeta_i + \epsilon_i \quad (5)$$

This model differs from the standard practice for modeling persuasion (e.g., Farrar et al., 2009) in two ways. First, our model includes pretest preferences O_i^0 in a way that does not induce endogenous variable bias. Second, our model focuses on the change in the respondent's latent and topic-specific preferences, represented by $\Delta\theta_i$ and $\Delta\zeta_i$, rather than the raw opinion change that are at best noisy measures of persuasion.

A research design that would enable this statistical strategy to measure the systematic component of preference change has several requirements. First, the respondent must express preferences on three or more topics in both the pre- and the post-discussion survey in order to identify the underlying latent preference space in $\Delta\theta_i$. If multiple outcomes

[6] By adding θ_i^0 to the model, we further change the mapping of the scale of the underlying ideological spaces in $\Delta\theta_i$ from β_1 to $(\beta_1 + \beta_2)$. This is only a mathematical transformation and highlights that scales do not have a ratio level of measurement and so require a transformation to bridge one space into the other. If one had substantive reasons to assume the two scales are identical in a specific application, one can choose instead to estimate a restricted model with $\beta_1 = 1$, $\beta_2 = 0$, and then assume (and hope) endogeneity bias does not exist in the application.

do not exist then only $\Delta\zeta_i$ is identified. Second, the standard assumptions in evaluating randomized control trials must be met (Angrist et al., 1996; Gerber and Green, 2012) in order to identify the effect of group composition rather than confounds to each group's discussion. We discuss these assumptions below in section 4.2.

3 The OBOE Town Halls

We apply our measurement strategy in a test of small group persuasion using a dataset from a randomized, large-scale deliberative field experiment. On June 26, 2010, nearly 3,000 individuals in 19 different cities convened in town hall meetings to discuss America's long term fiscal future.[7] The event, entitled "Our Budget, Our Economy," brought together diverse citizen-deliberators, armed with background reading material, to discuss and prioritize policy options that would help put the nation's budget on a more sustainable long term fiscal path. To recruit participants, the event organizer, America*Speaks*, worked with hundreds of local groups in each of the 19 cities, from all walks of life, to create a group of participants that closely mirrors the demographic composition of each community (see the appendix for a description of the event, recruitment, and the re-

[7]The event was held simultaneously in 19 sites in 19 different cities, and the sites were coordinated via videoconferencing technology. Six of the sites were designated "large sites" with approximately 500 participants each: Albuquerque, Chicago, Columbia (SC), Dallas, Philadelphia, and Portland (OR). The remaining sites were smaller and had 100 or fewer participants: Los Angeles, Des Moines, Overland Park (Kansas City), Louisville, Augusta (ME), Detroit, Jackson (MS), Missoula, Portsmouth (NH), Grand Forks, Richmond, Caspar, and Palo Alto. A table in the appendix gives the number of participants at each site.

spondents' characteristics).[8] In addition, America*Speaks* worked with over 30 national organizations that research and advocate budget policies, both liberal and conservative, to develop technical background reading material that was factual, balanced, and that represented the views of diverse perspectives.

On the day of the event, participants were randomly assigned to small group discussion tables, with the randomization occurring within each site.[9] They spent the entire day reading the materials, watching some instructional videos, and discussing their policy views with others seated at their table. Given the diversity of the participants in the town halls, randomization served two purposes. First, randomizing participants to small group discussions helped to assure that many participants were exposed to the views of citizens who were very different from themselves. In the absence of predetermined seating assignments, participants are likely to seek out other participants that are like themselves (Fowler et al., 2011), or to sit with other participants with whom they arrived at the event, which in turn would minimize the diversity of viewpoints available at each table. Randomization washes out any existing social ties among participants and diversifies the views to which participants are exposed. Since the groups were small in number, typically 10 participants, sampling variability under randomization assured that the composition of preferences would vary across tables, ranging from homogeneous to heterogeneous groups.

[8]The recruitment is similar to Barabas (2004). Since America*Speaks* could not compel a truly representative sample of citizens to participate in the experiment (see Fishkin and Luskin, 2005; Luskin et al., 2002), we can only state the in-sample group dynamics. The in-sample results remain interesting since they test for dynamics among those who have a propensity to show up to a deliberation.

[9]Prior to the event, the organizers printed up cards with table numbers, and then shuffled the cards before handing them to participants as they arrived. Randomization and balance tests show that the quality of the randomization was very good. See the appendix for a detailed analysis.

Second, random assignment allows us to identify the causal effects of exposure to different group compositions (Farrar et al., 2009) and so enables us to identify our measure of persuasion. In the present case, the mix of pre-discussion viewpoints among participants at a given table is exogenous to the analysis. One might believe a better measure of persuasion would rely on the arguments actually made in the course of the discussion, say from a transcript of the session (e.g., Karpowitz and Mendelberg, 2007; Westwood, 2015). This measurement strategy however cannot test for causal effects as the arguments offered during a discussion occur post-treatment; that is, since arguments are not randomly assigned, a statistical test based on arguments will lack internal validity.[10] Instead, we rely on the composition of pretest ideal points of the other participants at the respondent's table as an instrument of exposure to viewpoints during the discussion, since we can take the pretest ideal points of the discussion partners as an exogenous and randomly assigned encouragement to create the mix of arguments made in the discussion under an encouragement design (as in Farrar et al., 2009). To connect the group compositions to persuasion from discussion, our strategy must assume that 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.

The institutional context in which deliberation occurs can affect the nature of discussion. In the OBOE deliberation, America*Speaks* assigned a moderator to each table. The moderator did not participate substantively in the discussion and was trained by the event organizers in techniques to ensure that everyone at the table had the chance to speak, to encourage everyone to participate, and to enforce a set of rules (written on cards located at the center of each table) that were designed to make each table a "neutral, safe space"

[10]Using post-treatment argument as a causal variable would require the much stronger assumption "sequential ignorability" assumption from mediation analysis (Imai et al., 2011).

for expressing diverse views. We expect this careful structure to induce deliberative exchanges within the small groups (Barabas, 2004; Gastil et al., 2008; Gerber et al., 2016; Grönlund et al., 2015; Luskin et al., 2007), and so our findings might well depart from those of non-deliberative small group studies (see Isenberg, 1986).

4 Data and Model

The statistical model tests for the presence of persuasion within the small groups regarding various policy proposals considered at the event. At each of the 19 town halls, we asked participants to complete a short survey as they arrived, before the event began, and to complete another survey at the conclusion of the event. We refer to the former as the pretest survey, and the latter as the post-test survey. A total of 2,793 participants, seated at 339 tables across 19 different sites, filled out one or the other or (for the vast majority) both of these surveys.[11]

The pretest and post-test surveys each had a block of items asking participants their policy preferences on a set of proposals. The block of six questions is preceded with "Here are several things the government could do to cut the budget deficit. Please tell us what you think about each approach to reducing the deficit." The response categories each have a five point scale: "Strongly disagree," "Disagree," "Neither," "Agree," "Strongly agree." The items are (labels for items shown below in bold font were *not* in the survey):

Q1: Tax Rich Raise income taxes on the very wealthy – individuals making $250,000 ore more and households making $500,000 or more.

Q2: Cut Programs Cut discretionary federal programs and services by 5% across the board.

[11]Because the analysis depends on table-level summary statistic functions, we drop all tables with fewer than five participants. This omits 46 participants who were seated at 20 tables which is less than 2 percent of the sample.

Q3: Cut Entitlements Cut the growth of spending on entitlement programs such as social security and Medicare benefits.

Q4: Cut Defense Cut the spending on national defense and the military.

Q5: Tax Both Raise taxes on the middle-class as well as the wealthy.

Q6: Federal Sales Tax Create a new federal consumption tax, which would be like a federal sales tax that would be on top of any state and local sales tax.

In the American context, the first four items have a clear left-right orientation: to solve the deficit, liberals prefer to tax the rich and cut defense; conservatives prefer to cut discretionary programs and entitlements. The remaining two items that advocate taxing the middle class and a federal sales tax cut across liberal-conservative ideology. The statistical model makes use of pretest and post-test values of these items; an indicator of whether the pretest is missing (9 percent of pretests are missing);[12] a variable indicating a unique table identification number (among 339 tables total); and dummy variables indicating the site (out of the 19 sites, omitting one site) for each participant. The appendix provides summary statistics for all of the variables.

4.1 Statistical model

Our statistical model estimates the effect of small group composition on persuasion, making use of random assignment to groups and a measurement model. The full statistical model is given in the appendix. In this section we "walk through" the elements of the likelihood function in order to show how we measure persuasion, and how the parameters and functional form specifications allow us to test a variety of substantive hypotheses

[12] See appendix section A.9 for sensitivity tests that assess the possible range of estimates that would result under different extreme distributions of missing pretest data. Among those who filled out a pretest, 22 percent failed to fill out a post-test. We impute missing post-test data as missing at random conditional on the respondent's pretest response on the policy item, her ideology, and the ideological composition of her table.

regarding persuasion that are found in the literature on small group dynamics. The likelihood for a single categorical outcome is summarized in equation 6a, which is a non-linear implementation of equation 5.

$$O^1_{ik} \sim \text{OrderedLogit}(\boldsymbol{\beta_{1k}}\mathbf{O^0_{ik}} + \beta_{2k}\theta^0_i + \boldsymbol{\beta_{3k}}\mathbf{Site_i} + \omega_{ik}), \tag{6a}$$

$$\omega_{ik} = \Delta\theta_i + \Delta\zeta_{ik}. \tag{6b}$$

We estimate this model simultaneously for each of six policy preference items. In this equation, i indexes N participants (each i is a potential "persuadee") and k indexes $K = 6$ policies, which are labeled Q1 to Q6 above. The post-test policy preferences for each item and for each individual, O^1_{ik}, are modeled as a function of her pretest policy preference $\mathbf{O^0_{ik}}$, her pretest left-right ideal point θ^0_i, an indicator of the $\mathbf{Site_i}$ (city) of her event, and a random effect ω_{ik} that varies across individuals and policies. We describe each of these four elements in turn, noting for now that our main interest will focus on ω_{ik}.

The first component (O^0_{ik}) is the respondent's pretreatment response on the respective policy preference survey item. Including the pretreatment opinion on the right-hand side ensures that the structural parameters in the model estimate the individual's *change* in preference that occurs between the pre- and the post test (Farrar et al., 2009). As we describe above, including the pretreatment outcome on the right-hand side and estimating the $\boldsymbol{\beta_{1k}}$ parameter allows the the scale of the post-treatment outcome to vary. Since O^0_{ik} is categorical, we include a set of dummy variables indicating each of the first four response categories for the pretest item (omitting the fifth category), and hence $\mathbf{O^0_{ik}}$ is a matrix and $\boldsymbol{\beta_{1k}}$ is a vector. Using these dummy variables enables us to relax an assumption that each response category predicts the post-test response equally and in the same direction, and also allows the degree of scale compression and expansion to vary across the response options.

For the second component, we include θ_i^0 in the likelihood function to capture the endogenous dependence between the pretest and post-test response on the outcome, and so corrects for any endogenous variable bias that comes from including the pretest item in the outcome equation (see Skrondal and Rabe-Hesketh, 2004, 107-8). We use pretest responses to the tax rich, cut programs, cut entitlements, and cut defense (Q1 to Q4) items to estimate each participant's pretreatment latent ideal point preference scale, since these items have a clear liberal-conservative orientation.[13] We estimate each participant's latent ideal point θ_i^0 dynamically within the model, as in a structural equation model, and hence the estimation uncertainty inherent in θ_i^0 is included in the likelihood.

For the third component we condition on the *Site* or city in which the participants' event took place. Since randomization took place within sites these fixed effects allow us to control for any site-specific influences.

The fourth component of the likelihood function is a random effect, ω_{ik}, that varies across individuals and policies.[14] ω_{ik} measures the amount of dependence among the preference changes of participants in communication with each other (Anselin, 1988), for both the latent preferences ($\Delta\theta_i$) and the topic-specific preferences ($\Delta\zeta_i$) and hence represents the amount of a respondent's systematic preference change that is due to exposure to the discussion. In our application, since participants are randomly assigned to tables, we can state that any relationship between group composition and respondents' preferences we observe is caused by interpersonal interactions, rather than due to confounding, omitted variables or homophily.[15]

[13]We demonstrate in a separate analysis that there is a one factor solution for this set of items, where the first and last items had negative loadings and the other two positive, results not reported.

[14]We estimate the components of ω_{ik} using a nonlinear spatial auto-regression model, as described in Congdon (2003, chapter 7).

[15]As we discuss below, the ω_i parameter captures any within-group dependence, and

Because we estimate this model for multiple items simultaneously, and since the policy items contain an underlying latent structure, we are able to decompose ω_{ik} into two components, shown in equation 6b as a random effect that varies across individuals, $\Delta\theta_i$, and a second random effect that varies across both individuals and policies $\Delta\zeta_{ik}$. $\Delta\theta_i$ is a random effect parameter nested jointly within the full set of policy items and hence captures a systematic shift in preferences along the underlying latent, liberal-conservative dimension that structures preferences across topics. $\Delta\zeta_{ik}$ is specific to each policy item and captures dependence in the preference changes among participants seated at a table for that item, net of the latent component.

Since we define *persuasion* as the component of pre-post preference change that is due to interpersonal interactions, our interests lie in modeling variation across individuals and policies in ω_{ik} and hence variation in $\Delta\theta_i$ and $\Delta\zeta_{ik}$. We model these two dimensions separately. We define $\Delta\theta_i$ in equation 7a as a normally-distributed random effect with conditional mean $\Delta\theta_i^*$ and variance equal to one.[16]

$$\Delta\theta_i \sim \phi(\Delta\theta_i^*, 1), \tag{7a}$$

$$\begin{aligned}\Delta\theta_i^* = {} & \alpha_1 H_i + (\delta_1 \cdot \mathit{Liberal}_i + \delta_2 + \delta_3 \cdot \mathit{Conservative}_i) \cdot H_i^2 \\ & + (\gamma_1 \cdot \mathit{Liberal}_i + \gamma_2 + \gamma_3 \cdot \mathit{Conservative}_i) \cdot S_i \\ & + \kappa_1 \cdot \mathit{Liberal}_i + \kappa_2 \cdot \mathit{Conservative}_i.\end{aligned} \tag{7b}$$

We model the conditional mean for $\Delta\theta_i$ in equation 7b as a function of the latent ideal points of others seated at the respondent's discussion table (H and S, defined next) as well as the respondent's own ideology (liberal, moderate, or conservative). Equation 7b contains four distinct variables. To create the $\mathit{Liberal}_i$ and $\mathit{Conservative}_i$ variables, we

hence one must be careful in the study design not to introduce confounding group-specific interventions or influences that some groups are exposed to but not others.

[16]In an ordered logit model, the scale of the linear index is not identified and hence we must set this variance parameter to a constant. In other applications this variance should be estimated.

retrieve the pretreatment ideal point for each participant and trichotomize this scale into three equally sized groups. H_i is defined in equation 8a as the estimated mean of the pre-discussion ideal points of the discussants seated at i's table, excluding i's own ideal point. S_i is the variance of the ideal points of the other discussants at i's table, again not including i's own ideal point. These functions of ideal point estimates, H_i and S_i, are estimated dynamically within the structural equation model.

$$
\begin{aligned}
H_i &= \text{mean}(\theta^0_{ij}), && \text{(8a)}\\
S_i &= \text{mean}([\theta^0_{ij}]^2) - \text{mean}(\theta^0_{ij})^2, && \text{(8b)}\\
\theta^0_{ij} &\in \left\{\theta^0_j : j \text{ is seated at } i\text{'s table}, j \neq i\right\}. && \text{(8c)}
\end{aligned}
$$

In equation 8c, j indexes i's discussion partners, and the two mean functions are

$$
\begin{aligned}
\text{mean}(\theta^0_{ij}) &= \sum_j (\theta^0_{ij})/(N_i^-), && \text{(9a)}\\
\text{mean}([\theta^0_{ij}]^2) &= \sum_j ([\theta^0_{ij}]^2)/(N_i^-). && \text{(9b)}
\end{aligned}
$$

N_i^- is the number of participants sitting at i's table, not including i.

The parameterization and functional form of equation 7b is designed to test substantive hypotheses from the literature on small group dynamics. We have labeled each set of parameters with a different Greek letter (α, β, or γ) to indicate the hypothesis each set of parameters tests. The parameter α_1 estimates the degree to which person i's preferences depend on the ideal point composition of others seated at her table, which is the instrument for discussion using the pretreatment ideological orientation of the participants seated at the respondent's table (Farrar et al., 2009; Gastil et al., 2008; Klar, 2014). As Farrar et al. (2009) notes, since respondents ideological ideal points are measured pretreatment, we can take these as exogenous, and since group compositions are randomly assigned, we can take effects of this exposure to this measure of group composition to be causal.

The signs for the parameters δ_1 and δ_3 test whether there is polarization (Furnham et al., 2000; Isenberg, 1986; Schkade et al., 2010; Sunstein, 2002, 2008) evident in the respondents' latent-space persuasion, separately for liberals and conservatives; we include δ_2 corresponding to moderates for completeness and we do not have expectations for its sign. To state expectations for the signs of δ_1 and δ_3, note that we code the policy-preference items so that high values indicate a conservative response and low values indicate liberal, so higher scores on the latent ideal point scale indicate a conservative leaning. If liberals become more liberal, as the table becomes more liberal, then under a law of group polarization δ_1 should be negative as this would indicate that as a liberal respondent's table becomes more liberal, her latent preferences will become polarized and even more liberal (see the hypothetical curve in the left panel of figure 1). The patterns should be symmetric for conservatives and so under polarization δ_3 should be positive. If polarization is not evident, then these parameters will not differ from zero. We note the empirical deliberation literature proposes that structured deliberation inoculates groups from polarization (Barabas, 2004; Gerber et al., 2016; Grönlund et al., 2015; Klar, 2014; Luskin et al., 2007), and hence do not expect to see small group polarization to emerge in this context.

The parameters γ_1, γ_2, and γ_3 test whether the dispersion of ideal points at a table – a pretreatment measure of the extent of disagreement among the discussion participants – itself has an effect on preference change, separately for liberals, moderates and conservatives. While we do not have strong priors regarding the direction of this dynamic, it is possible that as the group becomes more divided (as the standard deviation of ideal points increases) participants will tend to selectively attend to the arguments that match their predispositions (see, e.g., Bolsen et al., 2014; Edwards and Smith, 1996; McGarty et al., 1994; Nyhan and Reifler, 2010; Tabor and Lodge, 2006) and hence increase the within-group polarization. In this case, γ_1 should be negative, γ_3, should be positive, and we have no prior expectations for γ_2.

The second, policy-specific component of persuasion, $\Delta\zeta_{ik}$, is defined in equation 10a as a normally-distributed random effect with mean $\Delta\zeta^*_{ik}$ and variance one. $\Delta\zeta^*_{ik}$ is a function of the respondent's own ideology and the policy-specific random effects $\Delta\zeta_{jk}$ of the other participants that are seated at i's table. Nesting this random effect within the participants of a given table enable us to assess the extent of dependence in the preference changes on the specific policy topic among table co-discussants, after netting out the covariates in the model as well as $\Delta\theta_i$. Methodologically, this random effect accommodates remaining spatial dependence within clusters (Congdon, 2003, chapter 7).

$$\Delta\zeta_{ik} \sim \phi(\Delta\zeta^*_{ik}, 1), \tag{10a}$$
$$\Delta\zeta^*_{ik} = (\rho_{1k} \cdot \mathit{Liberal}_i + \rho_{2k} + \rho_{3k} \cdot \mathit{Conservative}_i) \cdot \text{mean}(\Delta\zeta_{ijk}). \tag{10b}$$
$$\Delta\zeta_{ijk} \in \{\Delta\zeta_{jk} : j \text{ is seated at } i\text{'s table}, j \neq i\}, \tag{10c}$$

where

$$\text{mean}(\Delta\zeta_{ijk}) = \sum_j (\Delta\zeta_{ijk})/(N_i^-). \tag{11}$$

The parameters ρ_{1k}, ρ_{2k}, and ρ_{3k} estimate the degree of dependence on each policy preference item among table participants for liberals, moderates, and conservatives (respectively) after netting out each respondent's pretreatment preference, her own ideal point, both before and after the discussion. If a ρ_{2k} is positive and significant, this indicates that if everyone else at the table has a shift in their expected post-test preference on policy k, then person i also can be expected to have a shift in the same direction on issue k; conversely, if everyone else's preferences stay put, so does person i's. (Negative rhos are very unusual in this type of model.) If this dependence is net of ideology, then ρ_1 and ρ_3 should test to zero.

We assert that the two components of ω_{ik} (that is, $\Delta\theta_i$ and $\Delta\zeta_{ik}$) capture spatial dependence that comes from the respondents' exposure to her co-discussants. In particular, $\Delta\theta_i$ measures the extent to which the respondent's preferences change along a latent ideological dimension that results from exposure to discussion groups of varying preference compositions; $\Delta\zeta_{ik}$ measures the extent to which a respondent's post-discussion preferences are dependent on her co-discussant's post-discussion preferences on that specific topic for any other reasons. In both of these ways, the model measures dependence that comes from interpersonal interactions.

These 12 structural parameters, α_1, $\boldsymbol{\delta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\kappa}$, and $\boldsymbol{\rho}$ capture the effects of exposure to small group discussion partners on persuasion within the small group, with each set of parameters evaluating the specific mechanisms for persuasion for both ideological persuasion, measured by $\Delta\theta_i$, and for topic-specific persuasion for each policy, measured by $\Delta\zeta_{ik}$.

4.2 Interpretation and Assumptions

We can take exposure to the discussion group composition as a causal effect provided the standard assumptions for identifying causal effects within randomized control trials (RCTs) are met (see Angrist et al., 1996; Gerber and Green, 2012). The first assumption is randomization, which is met by the study design in that the event organizers used a random assignment procedure to assign table numbers, and because the number of participants at each table was fixed, a participant physically could not reassign herself to a different table (the appendix describes an extensive randomization check and balance tests for the table assignments).

The second assumption is the stable unit treatment value assumption (SUTVA), which has two requirements: there is no communication across tables and no alternate versions of the treatment. The assumption of no communication across tables is somewhat strong for our application in that tables were adjacent to each other, but one important design feature was that the tables were round, and as a physical configuration of the discussion space for each group the round shape strongly tended to focus discussion within a table and discouraged communication across tables. In addition, with hundreds of people in the event room, the discussion at other tables was mostly background noise. The assumption of no versions of treatment is met since no information relevant to the decision was introduced by a third party to some discussion groups during the discussion, but not to others. Otherwise, this information could create a group-specific dependence that would confound the effect of interpersonal interactions. The final assumption is the exclusion restriction, which requires that the random assignment process itself does not influence respondents' policy preferences other than through the group composition. This assumption is not testable but it is difficult to think of ways that our random assignment procedures would have any direct effect on preferences.

Given these three assumptions, we can use group composition as an instrument for exposure to the randomly assigned small group composition. While ideally we would like to measure persuasion from the arguments and statements made during the discussion (as in Westwood, 2015), we are only able to randomize assignment to group compositions and not to arguments. The arguments that are made and not made during the discussion may mediate persuasion, but we are unable to identify causal mediating effects given our research design (Imai et al., 2011). To the extent participants do not express their preferences and views, and assuming SUTVA and the exclusion restriction hold, the intention-to-treat estimand is a conservative estimate of the average treatment effect. It may be, for example, that some participants are conflict averse or shy and hence do not express their ideological dispositions within a discussion, but by randomization these personality traits are randomly distributed across tables.

Our proposed measurement of persuasion does not generalize to non-randomly assigned small groups or social networks, since the RCT assumptions are unlikely to hold in these situations. In naturally-occurring discussion groups or networks, within group dependence can occur due to confounding or homophily in addition to any influences from interpersonal interactions.

5 Results

We estimate the model in `OpenBUGS` using Bayesian MCMC methods (Lunn et al., 2009) and provide details in the appendix. We report the estimates for latent-space persuasion in figure 1 (using $\widehat{\alpha_1}$, $\widehat{\delta_1}$, $\widehat{\delta_2}$, and $\widehat{\delta_3}$) to estimate the degree of persuasion conditionally on the meann of the group ideal points, separately for liberals, moderates and conservatives. Figure 2 shows the effect of (pretest-measured) ideological diversity on latent persuasion (estimated by $\widehat{\gamma}_1$, $\widehat{\gamma}_2$, and $\widehat{\gamma}_3$). The results for topic-specific persuasion ($\widehat{\rho_{1k}}$, $\widehat{\rho_{2k}}$ and $\widehat{\rho_{3k}}$ for each of the six outcomes) are in figure 3.

5.1 Latent-Space Persuasion

The curves in figure 1, moving from left to right, show the effect of increasing the proportion of the participant's co-discussants that have ideal points at the conservative end of the latent preference scale, H_i, on the participant's change in the latent space ($\Delta\theta_i$). The middle panel of the figure (moderates) shows that $\widehat{\alpha}_1$ is positive, substantively quite large, and statistically significant, indicating that moderates' latent preferences respond to exposure to the mix of arguments they hear in the discussion. The left hand (liberals) panel indicates that $\widehat{\delta}_1$ is relatively small, positive in sign, and not significantly different from zero, and the right hand panel (conservatives) shows that $\widehat{\delta}_2$ is small, negative in sign, and also not significant.

These results for both liberals and conservatives are consistent with a linear pattern or even (by their point estimates) a diminishing return response to the table's ideal point composition when moving in the direction of the respondent's own ideological leaning. For example, as a table grows more conservative in composition, all participants tend to move in the conservative direction on the latent preference scale; but the right hand panel shows that conservatives themselves do not become especially more conservative; this pattern is symmetric for liberals.

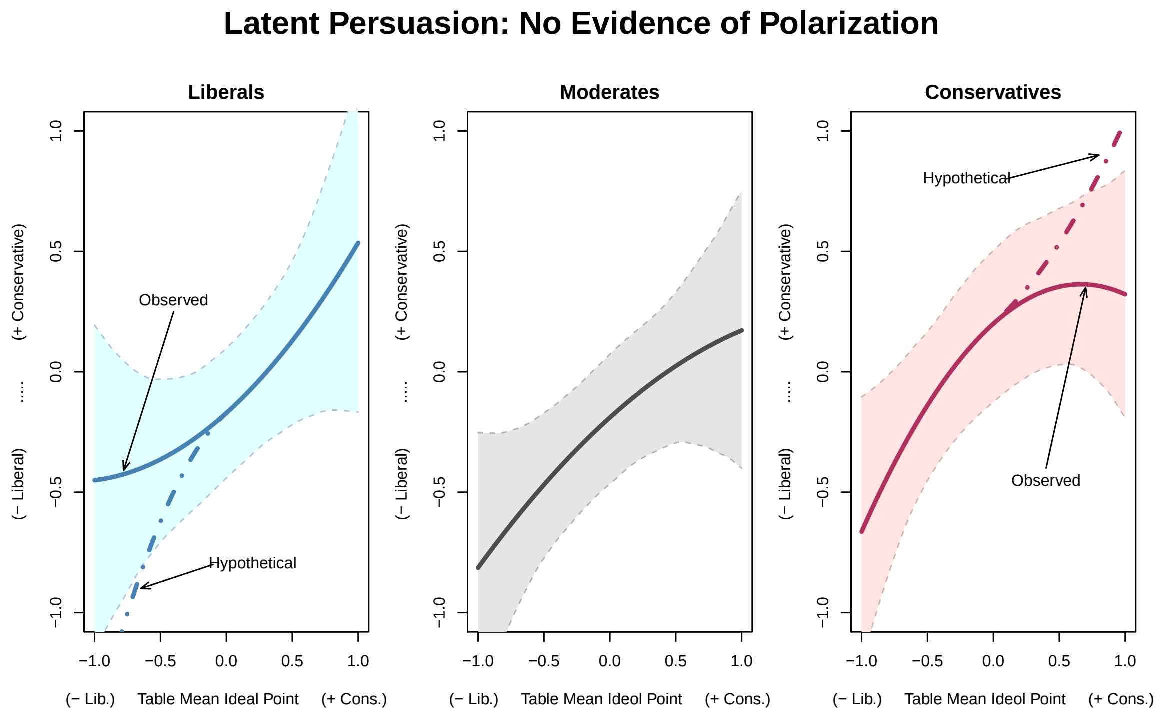

Figure 1: Latent-Space Persuasion. If the "law" of small group polarization held true, then we would expect to see liberals becoming even more liberal as the table grew more liberal (a concave pattern) and vice-versa for conservatives (a convex pattern). Instead we observe a linear relationship or diminishing returns, which is consistent with a mechanism of persuasive arguments within cross-cutting discourse. The confidence bands indicate 95 percent highest posterior density intervals.

Given these patterns, we do not observe ideological polarization within these small groups, findings that are similar to Barabas (2004), Gerber et al. (2016) and Grönlund et al. (2015) who show that deliberative institutions can inoculate small groups against polarization. Under a law of polarization (Sunstein, 2002) we would expect to see the curve in the right hand panel to be convex or upward-bending and the curve in the left hand panel to be concave or downward bending, patterns indicated by the hypothetical (dashed) curves in figure 1. The figure shows that the effect of latent-space persuasion is large, but similar for each group. That is, assuming that participants' pre-discussion ideal points are a good instrument for the quantity of ideologically-informed arguments they make, these results show that the participants are persuaded by fellow co-participants' ideological appeals, but that ideologues are not especially persuaded by co-ideologues to become extreme.

Recall that participants were randomly assigned to tables, and as a result the effects of table composition can be taken as causal persuasion. Under a counter-argument, one might worry that the linear increasing effect we observe is simply driven by a conformity process, in that a liberal seated at a mostly conservative table might simply conform to conservative positions under social pressure and vice versa. We can argue that conformity is not at work, however, in that the respondents filled out their post-test surveys privately as their final activity of the day and they had no reason to reveal their post-test responses to their co-discussants. Thus, participants completed the post-test in an environment that lacked social monitoring (for elaboration, see Boster and Cruz, 2003, 478).

One might also counter-argue that the diminishing effect we observe is due to a ceiling effect, in that liberals and conservatives might already be located near the endpoints of the latent preference scale with little additional room to move. This concern is mitigated in that, as we demonstrate in the appendix, the distribution of ideal points follows a normal distribution so there are very few respondents who are located near the endpoint of the scale. Indeed, only 8.4 percent of liberals chose the lowest category for each pretest preference item, and no conservatives chose the highest category for each.

5.2 Within-group polarization

In addition to the mean ideal point of the group, the statistical model for latent persuasion also includes a second function that characterizes the dispersion of ideal points within each table: the standard deviation of pre-discussion ideal points among participants at each table. This function is an instrument for the diversity of viewpoints available at a given table. Participants might respond to diverse viewpoints by combining those views and so provide a response on the post-test that is closer to the center (Druckman and Nelson, 2003). Alternatively, participants might use motivated reasoning to selectively attend to the arguments that tend to support their own preconceptions (see, e.g., Bolsen et al., 2014; Edwards and Smith, 1996; McGarty et al., 1994; Nyhan and Reifler, 2010; Tabor and Lodge, 2006) and so increase in their polarization through a form of confirmation bias. We do not have strong prior expectations regarding either of these patterns.

In the model the $\gamma_{.}$ parameters test for any effect from a diversity of viewpoints at a table, evaluating the effect of increased ideological diversity on liberals, moderates and conservatives. Figure 2 shows the results. Considering first the point estimates, we find that with greater diversity of views, liberals (and moderates) tend to become more liberal, while conservatives show no change.

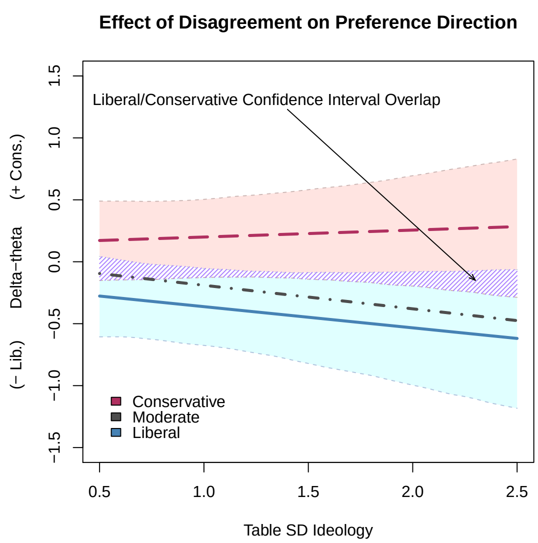

Figure 2: Disagreement and Persuasion. As the table becomes more ideological diverse, ideologues tend to reinforce their pre-existing views, although the effects are not statistically significant. The confidence bands indicate 95 percent highest posterior density intervals (not shown for moderates).

These point estimates suggest that it is diversity among discussants rather than ideological homogeneity that may increase polarization in a deliberative context. We note, however, that these point estimates are not statistically different from each other at standard levels. Thus the evidence for polarization from high levels of within-group disagreement at this event is relatively weak.

5.3 Topic-specific persuasion

Statements made in deliberation need not be constrained by any heuristic such as left-right ideology (Gutmann and Thompson 1996, 56; Habermas 1984, 99). We are able to assess the amount of persuasion that occurs outside the constraints of the latent preference scale in small group discussions by examining the degree of dependence of respondents' post-treatment topic-specific preferences ($\Delta\zeta_{ik}$) within a group on each policy preference item.

Figure 3 shows the estimates of the $\rho_{.k}$ correlation parameters assessing the degree of dependence in the topic-specific preference changes among table co-participants, separately by the ideology of the participant and the item. Overall, the figure indicates a very strong dependence of topic-specific preferences within tables since the $\rho_{.k}$ parameters are large and significantly different from zero for the cut social programs, cut defense, tax rich, and federal sales tax items, and the probability that ρ is different from zero is very large for the cut entitlements and tax both items. Remembering that assignment to tables is random, these results make a strong case for the existence of topic-specific persuasion.

The results of figure 3 show that participants' topic-specific preferences are responsive to interactions that occur within the small group discussions, and since the dependence is uniform between liberals, moderates and conservatives, we show that these preference changes are unrelated to the participant's left-right ideology. This finding is consistent with the aspirations of deliberative democracy in that participants appear to be responsive to reasons and rationales regarding policies that go beyond ideological appeals.

5.4 Evaluating the nature of persuasion

The Bayesian approach we use estimates a full posterior distribution for latent-space persuasion ($\Delta\theta_i$) and topic-specific persuasion ($\Delta\zeta_i$) as a separate parameter for each individual, and hence the posterior distributions are available for post-estimation analysis. Examining the correlates of each type of response change can help to illuminate the nature

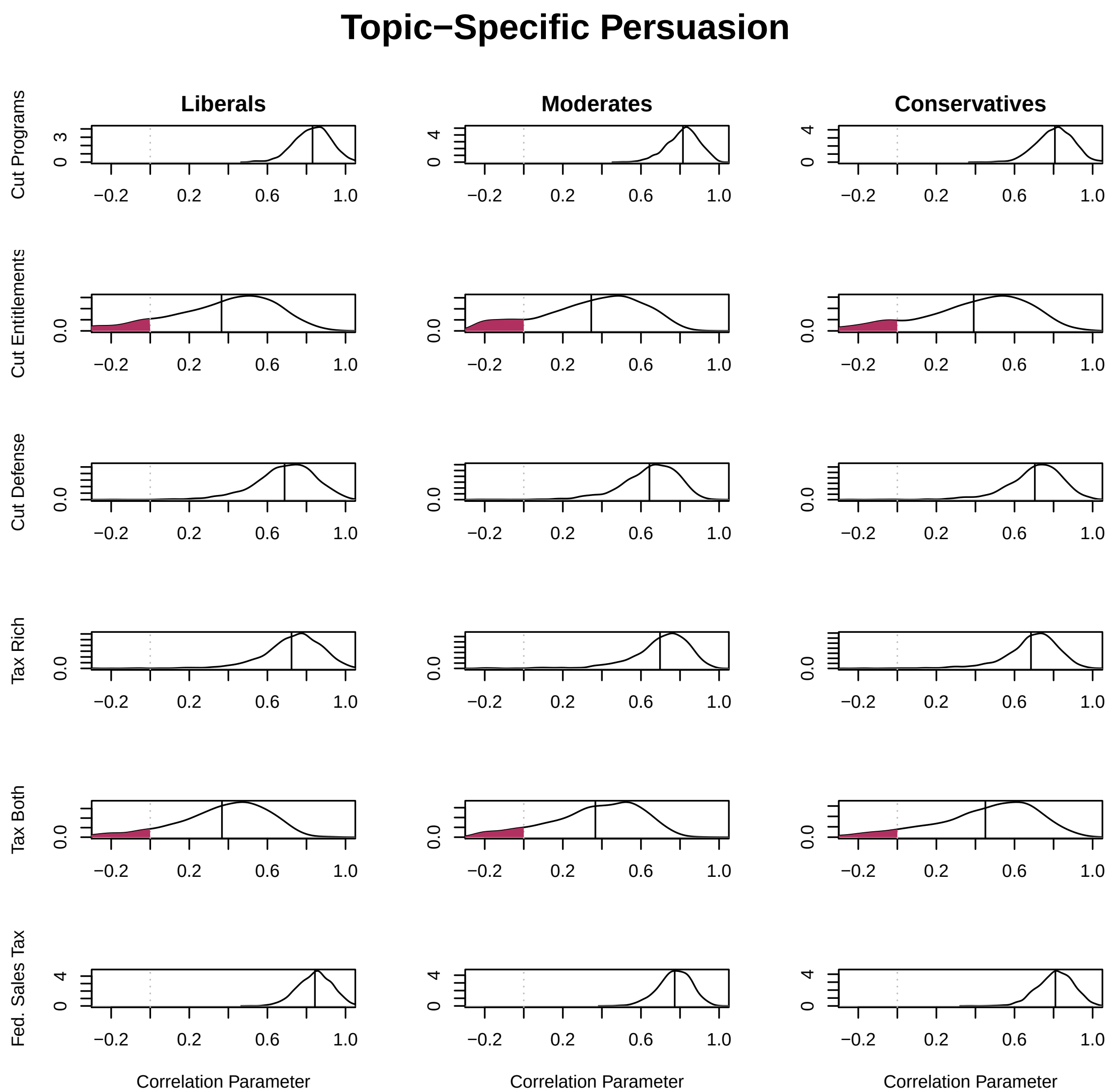

Figure 3: Topic-Specific Persuasion. This figure shows the posterior distributions for the $\rho_{.}$ correlation parameters, which test for spatial dependence in respondents' changes in topic-specific preferences for each item. Note that the dependence is identical for liberals, moderates and conservatives across all items.

and characteristics of persuasion, and in particular establish the construct validity (Cook and Campbell, 1979) of measured persuasion as an indicator of rational discourse.

As we mention above, one could reasonably assert that not all opinion persuasion that is caused by interpersonal interactions should be labeled rational or deliberative (Habermas, 1984). Instead, one might be persuaded by co-participants' arguments based on their personal characteristics rather than the substance of their arguments (Petty and Cacioppo, 1986), and this non-deliberative persuasion also can induce dependence among responses that is due to interpersonal interaction at a table. While we do not have objective measures of the quality of discourse at each table (such as Steiner et al., 2004), we can gain a sense of the nature of topic-specific persuasion by examining the correlates of the estimated topic-specific random effects, $\mathrm{E}[\widehat{\zeta_{ik}}]$, both in their direction and in their magnitude. We detail these supplemental tests in appendix section A.8.

The appendix details a battery of correlations we test using covariates from our survey. In short, the only consistently significant correlation with each $\widehat{\zeta_{ij}}$ we uncovered, both in direction and magnitude, is the perceived *informativeness* of the discussion. In addition, the signs of the correlations indicate that respondents' perceptions of the informativeness of the discussion covaries with movement toward favoring policies that solved the collective problem of the national debt, that is, toward increasing taxes and toward reducing spending. We find that among those who found the discussion to be informative, liberals tended to move toward conservative policies (cut programs and cut entitlements), conservatives moved toward favoring a liberal policy (tax rich) and both liberals and conservatives moved toward favoring the policies that do not load on the latent scale (tax middle class and the rich, and the federal sales tax).

These results are consistent with deliberative aspirations, in that participants who moderated their positions toward the collective goal of solving the debt crisis also perceived the discussion to be informative. While self-perceptions of informativeness do not measure the objective amount of rationality in discourse (Gerber et al., 2016), the correlation establishes the participants' subjective beliefs about the merits of the discussion, which in turn are likely to influence their views of the legitimacy of the event (Cohen, 1989).

5.5 Missing Data Sensitivity Analysis and Full Replication Study

The appendix provides analyses that examine the robustness and external validity of these results. First, the appendix analyzes the sensitivity of our findings to different assumptions regarding missing pretest responses.

Second, the appendix reports the results of a replication study that uses data collected from a separate event to test the external validity of the causal findings we report in this paper. The data come from the 2007 California*Speaks* health care policy event that also was conducted by America*Speaks*. These data are useful as an external validity test in that the 2007 and 2010 events were substantively very similar in design, but the 2007 study was conducted 1) three years earlier, 2) entirely within the state of California rather than nationally, 3) relied partially on randomized survey methods for recruitment, and 4) was on health policy rather than fiscal policy. We show that all of the causal results hold up under the replication, including the inconsistency of the observed preference changes with any law of small group polarization.

5.6 Discussion of the Application

The findings in our application differ with the vast bulk of the social psychology literature on non-deliberative small group discussion, which typically finds that small groups tend to polarize to extremes (Sunstein, 2002). While we did not vary the institutional setting in order to do a comparative institutional analysis, we can speculate that the dynamics of small group persuasion are likely to be responsive to the institutional context within which discussion occurs, and it is likely that deliberation is more constructive when the institutional setting is well-designed in a way that induces deliberative exchanges (Barabas, 2004; Grönlund et al., 2015).

If this institutional hypothesis were true, there would be two blocks of explanations, which are complementary and not mutually-exclusive. First, deliberative fora might attract a "deliberative class" of citizens who wish to engage in constructive discourse. Second, deliberative institutions might be designed to induce informed and constructive discourse compared to the non-deliberative institutions that are the typical focus of small-group research. For example, in the America*Speaks* event, the organizers established norms to govern the conversations, provided trained moderators for each table who were instructed to facilitate the conversation but not interject their own opinions, and the conclusions reached at this event did not directly cause redistribution of economic or social resources. Discussions that occur in contexts that do not contain these selection and design elements likely will differ from our findings regarding polarization.

6 Conclusion

This paper develops a novel measurement strategy to evaluate persuasion within small groups at a large-scale deliberative town hall. We wish to underscore the importance of measurement when testing hypotheses about persuasion. Our methods help to focus the statistical test on the systematic components of preference change that is due to interpersonal interactions, rather than the total variance in preference change that includes some unknown random or noise component. This explicit focus on measurement also allows us to identify substantively important dimensions of persuasion, where in our case the distinction between latent-space and topic-specific persuasion is important to understand the full dynamics of deliberative interaction. The methods we propose are very general and can be applied to any small group interaction where participants have been randomly assigned to small groups (as in Farrar et al., 2009).

In our application, we find that the persuasion we observe met many of the normative aspirations for deliberative democracy. Participants are responsive to their co-discussants' ideological appeals, but within the deliberative setting we do not observe a tendency toward ideological polarization (in contrast to Sunstein, 2002). In addition, we find that liberals and conservatives tend to be responsive to non-ideological appeals, which we label "topic-specific persuasion;" that the extent of both ideological and topic-specific persuasion covaries with the participant's perception that the discussion was informative; and the correlation between informativeness and persuasion was most evident for liberals on conservative policies and conservatives on liberal policies. Given the polarized nature of contemporary political discourse, particularly on national fiscal matters, we believe that reinforcing deliberative institutions might prove an effective way forward to address many of our pressing common problems.

References

Achen, C. H. (1975). Mass Political Attitudes and the Survey Response. *American Political Science Review 69* (Dec.), 1218–1231.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of Causal Effects using Instrumental Variables. *Journal of the American Statistical Association 91* (June), 444–455.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models.* Boston, Mass.: Kluwer Academic.

Ansolabehere, S., J. Rodden, and J. M. J. Snyder (2008). The Strength of Issues: Using Multiple Measures to Guage Preference Stability, Ideological Constraint, and Issue Voting. *American Political Science Review 102* (2), 215–232.

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

Bolsen, T., J. N. Druckman, and F. L. Cook (2014). The Influence of Partisan Motivated Reasoning on Public Opinion. *Political Behavior 36*, 235–262.

Boster, F. J. and M. G. Cruz (2003). Persuading. In J. Dill (Ed.), *The Persuasion Handbook: Development in Theory and Practice*, pp. 477–494. Sage Publications.

Broockman, D. and J. Kalla (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science 352*(6282), 220–224.

Clinton, J. D. (2012). Using Roll Call Estimates to Test Models of Politics. *Annual Review of Political Science 15*, 79–99.

Cohen, J. (1989). Deliberation and Democratic Legitimacy. In A. Hamlin and P. Pettit (Eds.), *The Good Polity: Normative Analysis of the State*, pp. 17–34. New York, N.Y.: Basil Blackwell.

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.

Druckman, J. N. and K. R. Nelson (2003). Framing and Deliberation: How Citizens' Conversations Limit Elite Influence. *American Journal of Political Science 47*(Oct.), 729–745.

Eatwell, R. (1993). Introduction: What Are Political Ideologies? In R. Eatwell and A. Wright (Eds.), *Contemporary Political Ideologies*, pp. 1–20. London: Pinter.

Edwards, K. and E. E. Smith (1996). A Disconfirmation Bias in the Evaluation of Arguments. *Journal of Personality and Social Psychology 71*(July), 5–24.

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.

Feldman, S. and C. Johnson (2014). Understanding the Determinants of Political Ideology: Implications of Structural Complexity. *Political Psychology 35*(3), 337–358.

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

Fowler, J. H., M. T. Heaney, D. W. Nickerson, J. F. Padgett, and B. Sinclair (2011). Causality in Political Networks. *American Politics Research 39*(March), 437–480.

Furnham, A., K. Simmons, and A. McClelland (2000). Decisions Concerning the Allocation of Scarce Medical Resources. *Journal of Social Behavior and Personality 15*(2), 185–200.

Gastil, J., L. Black, and K. Moscovitz (2008). Ideology, Attitude Change, and Deliberation in Small Face-to-Face Groups. *Political Communication 25*, 23–46.

Gastil, J., L. W. Black, E. P. Deess, and J. Leighter (2008). From Group Member to Democratic Citizen: How Deliberating with Fellow Jurors Reshapes Civic Attitudes. *Human Communication Research 34*, 137–169.

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

Gerber, M., A. Bachtiger, S. Shikano, S. Reber, and S. Rohr (2016). Deliberative Abilities and Influence in a Transnational Deliberative Poll (EuroPolis). *British Journal of Political Science Online*, 1–26.

Grönlund, K., K. Herne, and M. Setälä (2015). Does Enclave Deliberation Polarize Opinions? *Political Behavior 37*, 995–1020.

Gutmann, A. and D. Thompson (1996). *Democracy and Disagreement: Why Moral Conflict cannot be Avoided in Politics, and What Should be Done about It.* Princeton, N.J.: Princeton University Press.

Habermas, J. (1984). *The Theory of Communicative Action, Volume One: Reason and the Rationalization of Society* (McCarthy ed.). Boston, Mass.: Beacon Press.

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

Hinich, M. and M. C. Munger (1994). *Ideology and the Theory of Political Choice.* Ann Arbor, Mich.: University of Michigan Press.

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.

Isenberg, D. J. (1986). Group Polarization: A Critical Review and Meta-Analysis. *Journal of Personality and Social Psychology 50*(6), 1141–1151.

Karpowitz, C. F. and T. Mendelberg (2007). Groups and Deliberation. *Swiss Political Science Review 13*(4), 645–662.

Klar, S. (2014). Partisanship in a Social Setting. *American Journal of Political Science 58*(3), 687–704.

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.

Lunn, D., D. Spiegelhalter, A. Thomas, and N. Best (2009). The BUGS Project: Evolution, Critique, and Future Directions. *Statistics in Medicine 28*, 3049–3067.

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.

McGarty, C., S. A. Haslam, K. J. Hutchinson, and J. C. Turner (1994). The Effects of Salient Group Memberships on Persuasion. *Small Group Research 25*(2), 267–293.

Minozzi, W., M. A. Neblo, K. M. Esterling, and D. M. Lazer (2015). Field Experiment Evidence of Substantive, Attributional, and Behavioral Persuasion by Members of Congress in Online Town Halls. *Proceedings of the National Academy of Sciences 112*(13), 3937–3942.

Mutz, D. C., P. M. Sniderman, and R. A. Brody (1996). Political Persuasion: The Birth of a Field of Study. In D. C. Mutz, P. M. Sniderman, and R. A. Brody (Eds.), *Political Persuasion and Attitude Change*, pp. 1–14. University of Michigan Press.

Nyhan, B. and J. Reifler (2010). When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior 32*(2), 303–330.

Petty, R. E. and J. T. Cacioppo (1986). The Elaboraiton Likelihood Model of Persuasion. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology, Vol. 19*, pp. 123–205. Academic Press.

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

Prior, M. (2010). You've Either Got It or You Don't? The Stability of Political Interest over the Life Cycle. *Journal of Politics 72*(3), 747–766.

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.

Schkade, D., C. R. Sunstein, and R. Hastie (2010). When Deliberation Produces Extremism. *Critical Review 22*(2-3), 227–252.

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

Steiner, J., A. Bächtiger, M. Spörndli, and M. R. Steenbergen (2004). *Deliberative Politics in Action: Analysing Parliamentary Discourse.* New York, N.Y.: Cambridge University Press.

Sunstein, C. R. (2002). The Law of Group Polarization. *Journal of Political Philosophy 10*(2), 175–195.

Sunstein, C. R. (2008). *Republic.com 2.0.* Princeton, N.J.: Princeton University Press.

Tabor, C. S. and M. Lodge (2006). Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science 50*(3), 755–769.

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.

Tausanovitch, C. and C. Warshaw (2017). Estimating Candidates' Political Orientation in a Polarized Congress. *Political Analysis 25*(2), 167–187.

Treier, S. and D. S. Hillygus (2009). The nature of political ideology in the contemporary electorate. *Public Opinion Quarterly 73*(4), 679–703.

Westwood, S. J. (2015). The Role of Persuasion in Deliberative Opinion Change. *Political Communication 32*(4), 509–528.

A Appendix

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 the appendix 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

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

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 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

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.

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 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 4 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 4 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

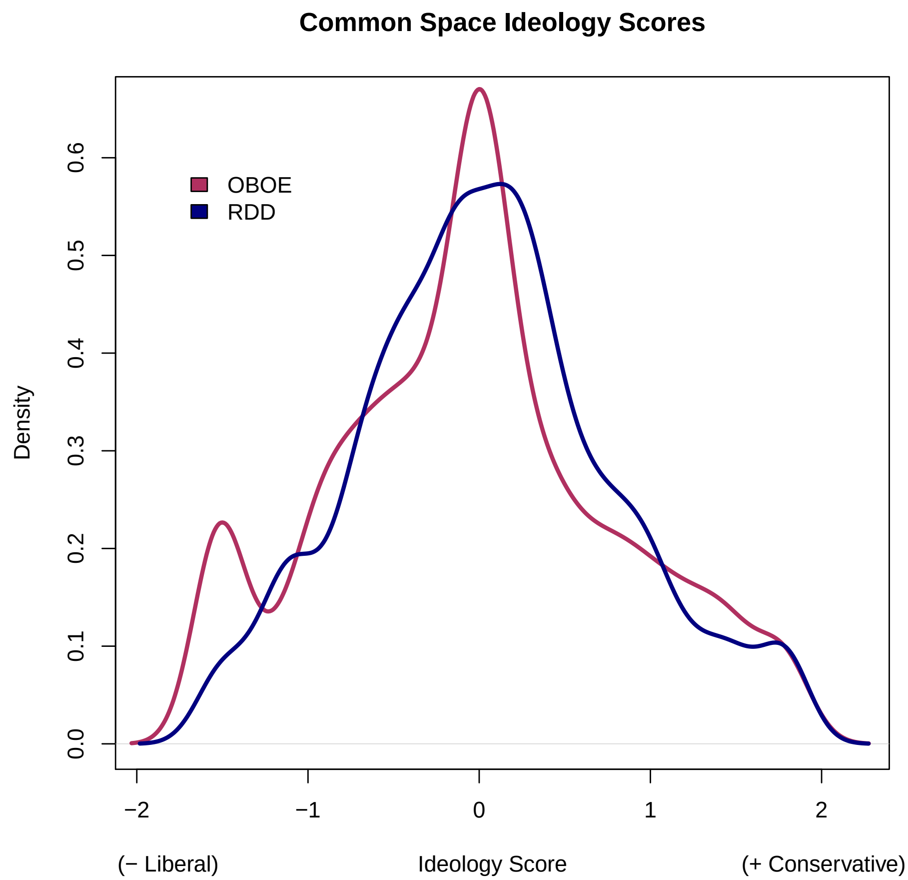

Figure 4: Ideological Common Space Comparison: RDD and OBOE Samples

liberals. Overall, however, Figure 4 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 Descriptive sample statistics

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

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 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 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 re-coded 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.[17]

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

[17] 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 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.5.1 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.[18] Under 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

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

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.[19] 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.[20] 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 variables 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[21] and external[22] efficacy, as well

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

[20]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.

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

[22]"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

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.[23] 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.[24] Note however that virtually no respondent chose the extreme category for the full set of items,[25] and as we demonstrate in the appendix,

government in Washington to do what is right."

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

[24]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 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.

[25]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.

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.5.2 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 participants "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.[26] Because of randomization to groups of size 10, we had no tables that con-

[26]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

tained 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.[27] These

this descriptive analysis, we disregard missing observations.

[27]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 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).

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

A.6 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 and diagrammed in figure 5. 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. 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.[28] 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.

Figure 5 diagrams the statistical model we use for a single outcome equation, including the variables, parameters and functional form we use to specify the model; in our application we estimate this model for six outcome measures simultaneously. In this figure, variables listed in squares are observed and variables in ovals are latent and hence

[28]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.

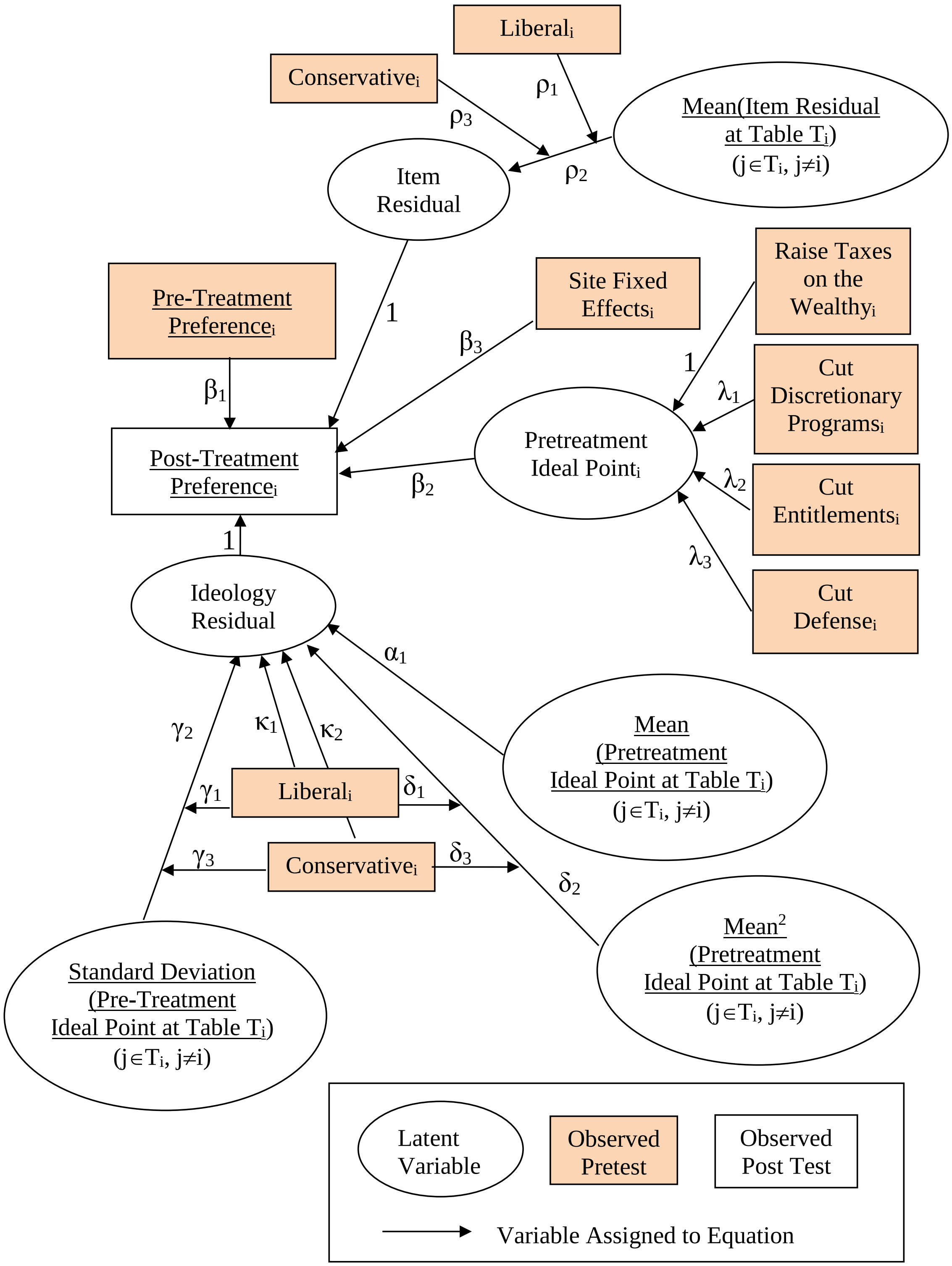

Figure 5: Persuasion Statistical Model

estimated. Arrows assign variables to equations. The shaded rectangles list the pretest variables and the unshaded rectangle indicates the post-test outcome variable.

A.7 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.9).

A.8 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.[29] 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 challenges 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

[29] 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).

duty as an American to speak out and be heard on this issue."[30]

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 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.5. 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).

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

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 7. Table 8 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

Table 7: Correlates of Topic-Specific Persuasion: Direction

	Tax Rich	Cut Programs	Cut Entitle-ments	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 described in figure 5; 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 8: Correlates of Topic-Specific Persuasion: Magnitude

	Tax Rich	Cut Programs	Cut Entitle-ments	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 described in figure 5. 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

to be less persuadable to agree with most items, both liberal and conservative).[31]

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-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 9 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

[31]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 9: Correlation of Topic-Specific Persuasion with Informative Discussion, by Ideology

	Tax Rich	Cut Programs	Cut Entitle-ments	Cut Defense	Tax Both	Federal Sales Tax
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 described in figure 5. 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

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.

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 3 makes 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.9 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 6 and 7 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.

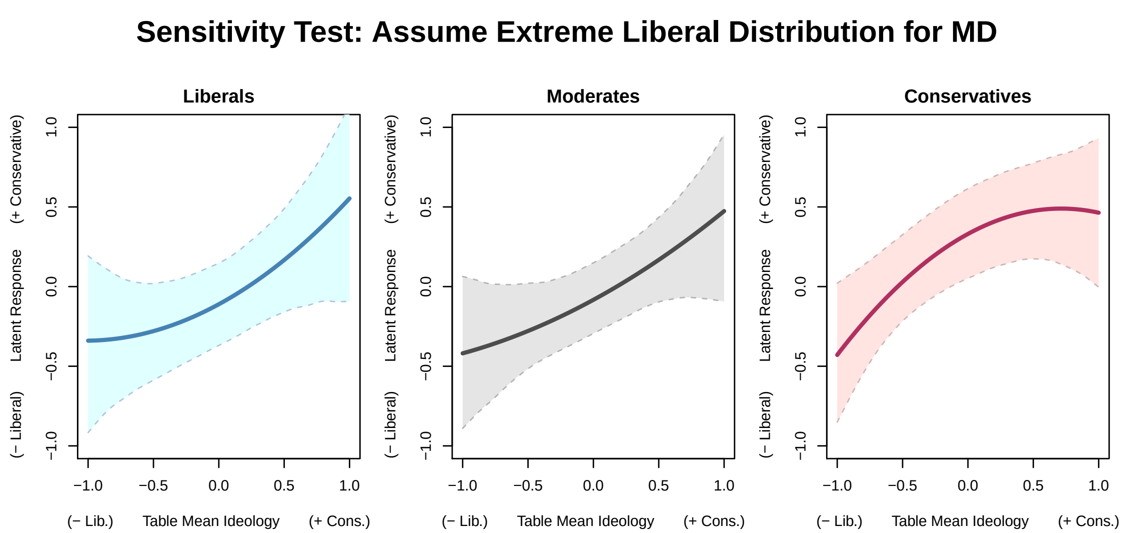

Figure 6: Sensitivity Test: Assume a Liberal Distribution for Missing Data

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.

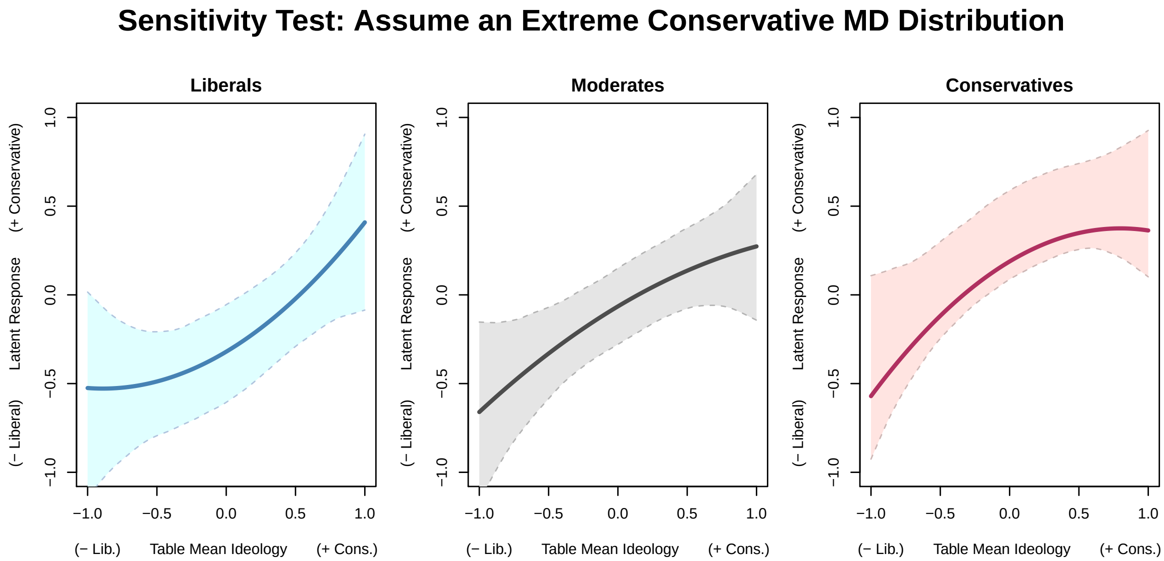

Figure 7: Sensitivity Test: Assume a Conservative Distribution for Missing Data

A.10 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[32] and the data are very useful as a replication as 1) the California study occurred three years prior to the OBOE study, 2) was limited to eight cities in California,[33], 3) were recruited in part

[32]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.

[33]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.

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)

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 8 to 11 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

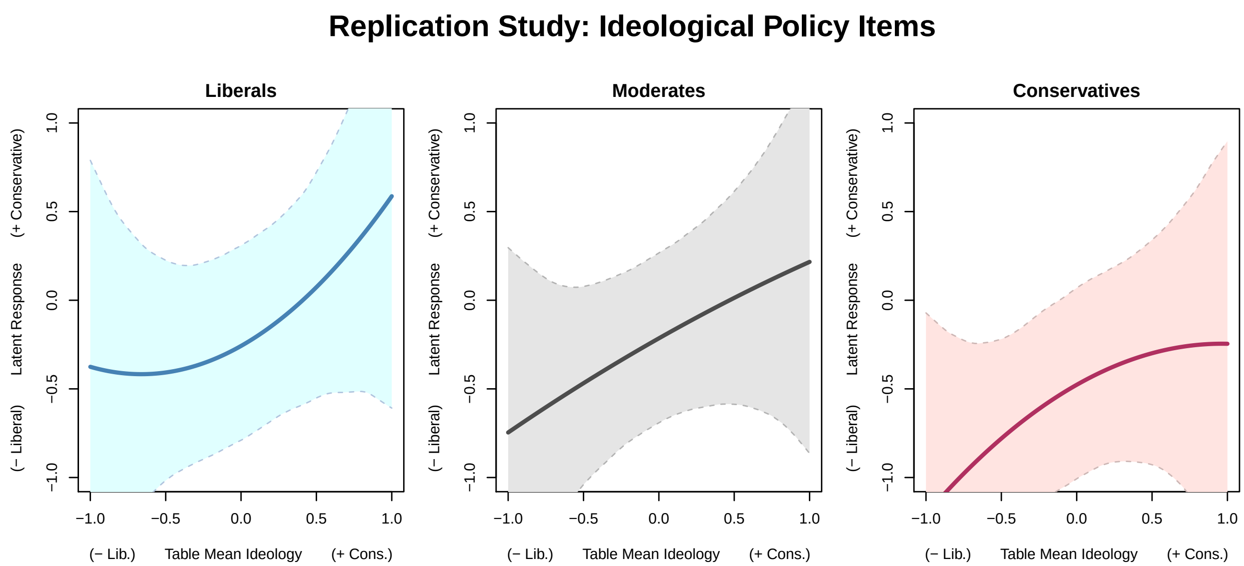

Figure 8: Replication Study: Mean Composition Effect on Ideological Dimension

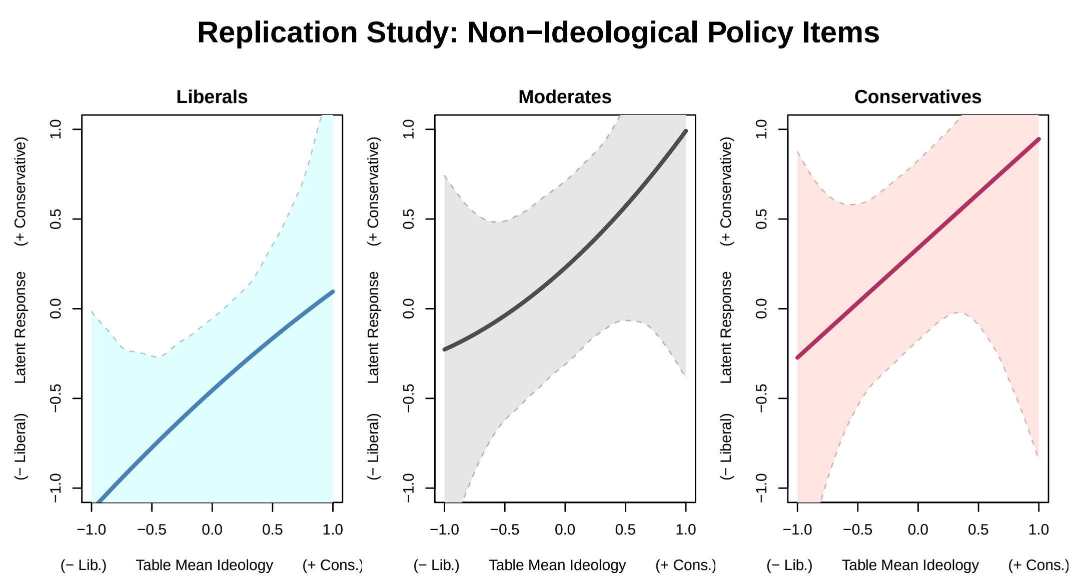

Figure 9: Replication Study: Mean Composition Effect on Non-Ideological Dimension

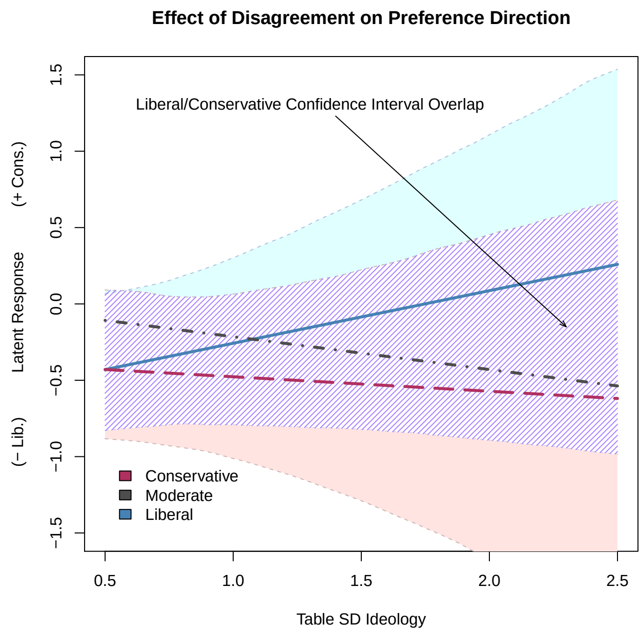

Figure 10: Replication Study: Disagreement Effect on Ideological Dimension

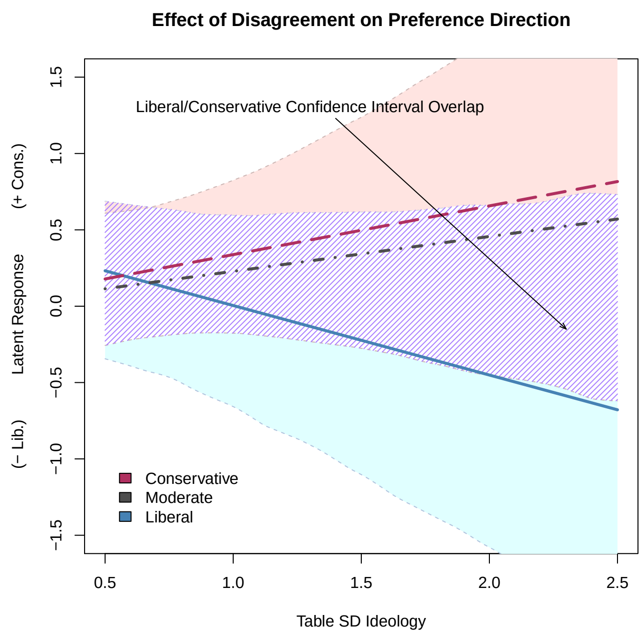

Figure 11: Replication Study: Disagreement Effect on Non-Ideological Dimension

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.

A.11 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 the appendix we provide sensitivity tests to show that this imputation does not affect the estimated quantities of interest at all.