Supporting Information for

A randomized trial of a lab-embedded discourse intervention to improve research ethics

Dena K. Plemmons,[1*] Erica N. Baranski,[2] Kyle Harp,[1] David Lo,[1] Courtney K. Soderberg,[3] Timothy M. Errington,[3] Brian A. Nosek,[3,4] Kevin M. Esterling[1*]
[1]University of California, Riverside
[2]University of Arizona
[3]Center for Open Science
[4]University of Virginia
[*]Correspondence regarding the training intervention to: dena.plemmons@ucr.edu; correspondence regarding the methods and other inquiries to: kevin.esterling@ucr.edu

This file includes:

Supporting Information

www.pnas.org/cgi/doi/10.1073/pnas.1917848117

Measures

The pre and post surveys consisted of a series of dichotomous, Likert-type, and open-ended questions assessing attitudes and behaviors relating to RCR in general and data management and authorship in particular. A subset of these questions also assessed lab dynamics and culture. For the complete pre and post surveys, visit https://osf.io/stzgw/.

Note that in the case of each of the variables with Likert responses, the respondent is given five response categories. In each case, on each item only a small handful of respondents chose either the lowest response or the highest response, depending on the item, so we recoded each variable to four categories and collapsed the response category with the fewest responses into the adjacent category. Below we indicate the variables where we recoded the 5 category to 4; for all others 2 is recoded to 1 and 3, 4 and 5 are each decreased by one.

There were two distinct types of questions relating to our approach to subsequent analyses. The first type was direct measures of exposure to the important elements of the training intervention, and so the items themselves are meaningful for analysis. Because the trainings centered on two RCR topics, authorship and data management, in the survey we asked, ”Does your lab have an established authorship plan governing the assignment and order of authors for manuscripts? (1 = yes, 0 = no), and ”Does your lab have a data management policy?” (1 = yes, 0 = no). Since the training was delivered in a face-to-face format that encouraged participation and discussion, we asked ”Have you changed your views about ethical research practices based on discussion within your lab?” (1 = my views haven’t changed at to 5 = my views have changed a lot; recoded 5=4).

Through a factor analysis that we describe in our pre-analysis plan (described below), we established six scales based on batteries of items with high inter-correlation. The first scale measures respondent’s beliefs about the relevance of ethics discourse for their research and work. The items that load on the *Relevance of Ethics Discourse* scale are as follows: (1) “How relevant to your area of work is learning about ethical research practices?” (1 = completely irrelevant to 5 = completely relevant), (2) “To what extent do you agree that seeking others in your department to discuss ethical research practice is your responsibility as a scientist?” (1 = strongly disagree to 5 = strongly agree), (3) “To what extent do you agree that seeking out others in your lab to discuss ethical research practices is your responsibility as a scientist?” (1 = strongly disagree to 5 = strongly agree), (4) “Do you have discussions with members of other labs regarding ethical research practices?” (1 = never to 5 = always; recoded 5=4).

The second scale measures the extent to which the participants perceive the discussion in their lab to be reason-giving, respectful and equal, which are core requirements in the concept of deliberation. The items on the *Respectful Discussion* scale were developed in (37) and are as follows: (1) “If there was a disagreement about research practices, would lab members in your lab make reasonable points and try to make valid arguments?” (1 = never to 5 = very often), (2) “When discussing research with your lab, does everyone have a real opportunity to speak with no one inappropriately dominating the discussions?” (1 = never to 5 = always), (3) “When discussing research with your lab, do lab members listen to one another respectfully and courteously?” (1 = never to 5 = very often), (4) “When discussing research with your lab, do lab members seem to hear and understand your views?" (1 = never to 5 = very often), (5) "How confident are you that your lab co-workers will use ethical research practices when conducting research?" (1 = not confident at all to 5 = extremely confident).

The third scale measures climate and the amount of disagreement in the lab, which is a measure of the constructiveness of within-lab communication. The items on the *Lab Disagreement* scale are as follows: (1) "How often are there disagreements in your lab about research practices?" (1 = never to 5 = very often; recoded 5=4), (2) "How often are there disagreements in your lab about authorship?" (1 = never to 5 = very often; recoded 5=4), (3) "How often are there disagreements in your lab about data management?" (1 = never to 5 = very often; recoded 5=4).

The fourth scale measures the respondent's self-reported understanding of the reasons and rationales for having discussions about a lab authorship plan, which was one of the substantive topics of the training. The *Reasons for Authorship Policy* scale items are as follows: (1) "Do you understand the rationale for having an authorship plan in your lab?" (1 = I don't understand the rationale at all to 5 = I mostly understand the rationale), (2) "Do you understand the importance of having an authorship plan in your lab?" (1 = I don't understand the rationale at all to 5 = I mostly understand the rationale), (3) "Do you understand the implications for having an authorship plan in your lab?" (1 = I don't understand the rationale at all to 5 = I mostly understand the rationale).

The fifth scale measures the respondent's self-reported understanding of the reasons and rationales for discussions about a lab data management policy, which also was a substantive topic of the training. The *Reasons for Data Management* policy scale items are as follows: (1) Do you understand the rationale for having a data management plan/policy in your lab?" (1 = I don't understand the rationale at all to 5 = I mostly understand the rationale), (2) "Do you understand the importance of having a data management plan/policy in your lab?" (1 = I don't understand the importance at all, etc...), (3) "Do you understand the implications for having a data management plan/policy in your lab?" (1 = I don't understand the ethical implications at all, etc...).

Finally, the sixth scale measures the respondent's perception of the importance of open science practices with respect to the preservation of replication materials, which is a key feature of the OSF and arguably of good data management practice. The *Preserve Replication Materials* scale items are as follows: "How important to your lab is archiving records of earlier versions of data sets? (1 = not at all important to 5 = extremely important), (2) "How important to your lab is archiving records of earlier versions of manuscript drafts?" (1 = not at all important to 5 = extremely important), (3) "How important to your lab is archiving records of earlier versions of lab materials?" (1 = not at all important to 5 = extremely important) (4) "How worthwhile is it to maintain electronic copies of lab materials and methods?" (1 = not worthwhile at all to 5 = extremely worthwhile).

Manipulation check

Each midpoint survey fielded two dichotomous questions: "Have you learned about any on-line data management systems for any course in the past two months?" and "Have you received any ethics training from any source in the past two months?" We use these two dichotomous items to conduct a manipulation check for exposure to the training. We estimate a logit regression conditioning on assignment. The odds ratio exceeds five for each item for both surveys, with each posterior distribution fully to the right of zero.

Statistical Analysis

Because we used random assignment in the RCT, we can identify the causal effects of our training interventions relative to the control condition by comparing follow-up survey responses across the two conditions of the study (39). The RCT design enables a comparison of the post-intervention (follow-up) survey responses between treatment and control arms and is a well-known experimental design.

In addition, since we have access to participants' pre-survey responses to each item, we embedded the RCT inside of a difference-in-differences (DID) design (40). Nesting the two research designs makes the causal effects doubly-robust in that the assumptions for both the RCT and the DID design would need to be violated for the effects to not reflect causality.

In the basic DID design, the researcher conducts a statistical test to see if there is a relationship between group exposure and the *difference* between the pre-treatment O_i^0 and post-treatment O_i^1 survey responses (41). The DID design proceeds by first taking the difference in the outcomes of the intervention and control groups in the time period after the intervention was administered, and then subtracting from that quantity the difference in outcomes between the two groups observed before the intervention. The DID design identifies the causal effect of this difference-in-differences using the "parallel path" assumption (described below) that the intervention group would have had the same over-time trajectory in outcomes as the control group, had it not been exposed to the intervention. Since our intervention and control groups were constructed by randomization, the two groups have no systematic differences in expectation, and so this is a very weak assumption.

Here we extend the DID design to accommodate comparisons across scales and report the results in Figure 3 of the main text. The individual pretest and post-test responses each contain some degree of measurement error in addition to a systematic component that captures respondents' latent opinion on that item at a given time (42-44). To address this, we create scales that measure latent concepts of interest. Mathematically, a "scale" is a weighted sum of individual survey items that are highly intercorrelated, where the weights are recovered from a statistical procedure that we describe below. Conceptually, one can think of these survey items as indicators that help to measure an underlying or latent variable that is of conceptual interest but that cannot be measured directly. For example, many of our concepts of interest focus on the nature of deliberative communication, lab climate, ethics beliefs, and scientific practice, and each of these concepts is difficult to measure accurately with ordinary survey responses. The virtue in focusing on scales is that the scales help to organize the data analysis on the larger concepts (or constructs) regarding changes over time at this more fundamental level.

The individual survey items contain measurement error, which reduces the power of a statistical test. And indeed, very few of the items that are designed to be scaled showed a statistically significant treatment effect, while the treatment effect is more apparent with the scales; this is sensible because the individual items have more statistical noise than do the scales. For the analyses of the scaled outcomes, our estimation method focuses only on changes in the systematic component rather than on the items themselves that contain an unknown amount of noise.

To formalize the statistical model as applied to scale outcomes we rely on (45). To derive the general model, for simplicity assume a continuous, normally distributed opinion response at time *t*, O_i^t, and decompose the opinion response as

$$O_i^t = \beta_o + \lambda_1^t \theta_i^t + \varepsilon_i^t, \qquad t \in \{0,1\} \tag{A.1}$$

where θ_i^t is the scale measuring each respondent's latent opinion and so measures the latent concept or construct of interest, λ_1^t is a scaling structural parameter (sometimes referred to as a factor coefficient or discrimination parameter) and ε_i^t is the idiosyncratic component from measurement error that represents instability in the individual's opinion response, all evaluated at time *t*; *t* = 0 is the pretest and *t* = 1 is the post-test. We identify θ_i^t by nesting questions within participants.

To derive the DID statistical model for latent scales, we can take the difference in equation (A.1) between time *t* = 1 and *t* = 0,

$$\begin{gathered} O_i^1 = \beta_0^1 + \lambda_1^1 \theta_i^1 + \varepsilon_i^1 \\ \beta_1 \left(O_i^0 = \beta_0^0 + \lambda_1^0 \theta_i^0 + \varepsilon_i^0 \right) \end{gathered} \tag{A.2}$$

Subtracting the second row of equation (A.2) from the first and rearranging yields,

$$O_i^1 = \beta_0 + \beta_1 O_i^0 + \Delta\theta_i + \varepsilon_i \tag{A.3}$$

where $\beta_0 = \beta_0^1 - \beta_1 \beta_0^0$ and $\varepsilon_i = \varepsilon_i^1 - \beta_1 \varepsilon_i^0$. With this derivation we have identified a new quantity of interest, $\Delta\theta_i = \lambda_1^1 \theta_i^1 - \lambda_1^0 \beta_1 \theta_i^0$ which is the change in the respondent's pre- to post-discussion opinion in the scaled *latent opinion* space. This derivation allows us to isolate and measure the systematic component of opinion change using the latent scale, and we can test hypotheses regarding the differences in this latent scale between the experimental conditions, rather than relying on the noisily measured changes in the survey response itself.

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 opinion change, note that $\operatorname{cov}\left(\theta_i^0,\theta_i^1\right)\neq 0$, since $\lambda_1^1\theta_i^1=\lambda_1^0\beta_1\theta_i^0+\Delta\theta_i$, and so $\operatorname{cov}\left(O_i^0,O_i^1\right)\neq 0$ by equation A.2. In the statistical model below we correct for this by conditioning on 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 responses, and that the remaining variation in O_i^0 and O_i^1 is conditionally independent.[†] Thus, the equation we estimate is,

$$O_i^1=\beta_0+\beta_1 O_i^0+\beta_2\theta_i^0+\Delta\theta_i+\varepsilon_i \tag{A.4}$$

We note too that modeling the θ_i^0 and $\Delta\theta_i$ parameters jointly with the structural parameters $\boldsymbol{\beta}$ correctly propagates the uncertainty that comes from estimating these parameters through the statistical model. That is, the estimates of the $\boldsymbol{\beta}$ parameters are the marginal distributions having integrated over the sample space underlying θ_i^0 and $\Delta\theta_i$, a technique known as the "method of composition" (46).

To identify the θ_i^0 and the $\Delta\theta_i$ parameters, we nest the set of questions for each scale within individual respondents. The likelihood for a single categorical outcome is summarized in the second row of equation A.5, which is a non-linear implementation of equation A.4.

$$\begin{aligned} O_{ik}^0 &\sim OrderedLogit\left(\lambda_{1k}\theta_i^0\right)\\ O_{ik}^1 &\sim OrderedLogit\left(\beta_{1k}O_{ik}^0+\beta_{2k}\theta_i^0+\Delta\theta_i\right) \end{aligned} \tag{A.5}$$

We estimate this model simultaneously for each of the opinion items within a scale, and we model each of the six scales separately. In this equation, *i* indexes *N* participants and *k* indexes *K* questions for a given scale.

Finally, we note that the participants are each nested in a lab and so we model the dependence of responses within labs using a conditional autoregressive prior (see *47*

[*] When taking the differences, the standard DID design implicitly sets the coefficient on the pretreament outcome to one, a fixed constant. The model we develop here relaxes this constraint by modeling the outcome endogeneity. The results are substantively the same when we impose the constraint (results not reported).

[†] Adding θ_i^0 to the model subtracts $\beta_2\theta_i^0$ from $\Delta\theta_i$. As a result, we change the mapping of the scale of the underlying latent opinion spaces in $\Delta\theta_i$ from $\lambda_1^0\beta_1$ to $\left(\lambda_1^0\beta_1+\beta_2\right)$. 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. The results are substantively the same when we apply these two constraints (results not reported).

chapter 7). We define $\Delta\theta_i$ in the first row of equation A.6 as a normally-distributed random effect with conditional mean $\Delta\theta_i^*$ and variance equal to one,[‡]

$$\begin{aligned} &\Delta\theta_i \sim \phi\left(\Delta\theta_i^*, 1\right) \\ &\Delta\theta_i^* = \alpha_1 Z_i + \eta_i \end{aligned} \tag{A.6}$$

where Z_i is the randomized treatment assignment indicator. Since $\Delta\theta_i$ is the difference in the pre- and post- treatment scale for a respondent, the structural parameter α_i is the difference-in-differences estimand for each scale. η_i is the lab-level cluster random effect that accommodates spatial dependence of responses within a lab and

$$\begin{aligned} &\eta_i \sim \phi\left(\eta_i^*, 1\right) \\ &\eta_i^* = \frac{\rho\sum_j \eta_{ij}}{N_i}, \text{ where } j \in \left\{i\text{'s lab, } i \neq j\right\} \end{aligned} \tag{A.7}$$

and N_i is the number of lab members not counting the respondent. The ρ parameter captures the degree of autoregressive dependence of the latent scale responses within a lab and so η_i serves as a level-2 random effect that nets out lab-specific dependence from the treatment effect equation (see 47, chapter 7).

In the single equation models that we report in Figure 2 of the main text we use an identical approach to estimate the difference-in-differences with nesting in labs, except the outcome of interest is individual items instead of scales. We estimate the single equation models using this same nesting approach, using a logit likelihood for the dichotomous responses and an ordered logit model for the model with an ordinal response.

Identification Assumptions

We can take exposure to the training as a causal effect under the standard assumptions for identifying causal effects within RCTs (48,49). The first assumption is randomization, which is met by the study design and the administration of the randomization, and because labs were not able to re-assign themselves to a different treatment arm.

For all colleges other than the engineering college, we block randomized within departments; as we mention above in engineering we use simple randomization. For the block randomization, the first lab from a given department was randomized to one of the conditions; each subsequent lab after the first lab is randomized based on the first lab's random assignment. In this sequential assignment there is a possibility for labs to self-

[‡] 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 could be estimated.

select into experimental conditions but this is highly unlikely. Self-selection would require the PI 1) knows our assignment protocol is alternating within departments, which we did not reveal, 2) knows if the first lab was assigned to treatment or control, 3) knows that at least one other lab in the department is planning to enroll, and 4) decided to time her enrollment in the study to either go next (and ensure they get the opposite assignment) or wait until the other lab enrolls to get the same assignment as the first. And also 5) that the other labs in the department are not doing 1-4. This also would require that there are more than two labs in the department that participated.

The second assumption is the stable unit treatment value assumption (SUTVA), which has two requirements: there is no communication across labs and no alternate versions of the treatment. The assumption of no communication across labs is somewhat strong for our application in that labs are nested in departments, but typically there were only 1-3 labs per department. The assumption of no alternate versions of treatment requires that no information similar to the training was introduced during the study period to some labs but not to others. The final assumption is the exclusion restriction, which requires that the random assignment process itself does not influence respondents' opinion responses other than through the training. This assumption is not testable but it is difficult to think of ways that our random assignment procedures would have any direct effect on the opinions that we measure.

The assumption required to identify a causal effect in the DID framework is the "parallel path" assumption, which is that the respondents in the treatment arm would have had the same change over time as observed in the control arm, if the treatment respondents had not been exposed to the treatment. This assumption is reasonably assured by the randomization, which helps to ensure the intervention and control groups are similar.

Finally, we note that our design nests the RCT design within the DID design, so the causal inferences are doubly-robust; only one set of assumptions needs to hold in order to identify causal effects.

Estimation

We estimate all models in OpenBUGS version 3.2.3 using Bayesian MCMC methods (50). All OpenBUGS code to implement these models can be found at (51). The model likelihood for the scale outcomes is shown above in equations A.5, A.6 and A.7. We assign Γ(0.1, 0.1) vague directional priors for all factor coefficient parameters (where the direction is assigned for identification), a log normal prior for the ρ parameter with log mean zero and log standard deviation 1, and flat N(0,1000) priors for all other parameters. We run the MCMC simulation from overdispersed initial values until the chains are stationary and then sample until the posterior distributions are smooth. The replication materials (described below) provides the traceplots and postestimation analysis for stationarity as well as all of the sampled results we report in the paper.

Missing Data Imputation

We impute the missing all data as missing at random given the observed data and model-implied latent variables (52). The Bayesian model propagates the estimation uncertainty in the missing data imputations, and so the estimates of the structural parameters incorporate the additional uncertainty that is due to the missing data; that is, all parameters are marginalized over the full distribution of missing data (53).

Randomization Balance Tests

We block randomized labs to treatment and control conditions in order to identify the causal effects from our estimates.[§] We checked the realized randomization by conducting a balance test of the similarities of the joint distribution of the available covariates across the conditions. We estimated the omnibus test statistic of (54) which is distributed χ^2. We observed that the pairwise distributions of non-white, male and PI role have standardized differences of 0.08, 0.17, and -0.09, respectively, and none of these differences are pairwise significant. The omnibus test isχ^2=2.68 with *df* = 4 (*p* = 0.61) and so we cannot reject the null hypothesis that the covariates are balanced.

Factor model results

We use the pre-survey responses to evaluate the convergent validity of the scales that measure the latent outcomes of interest to our study. In our preanalysis plan (described below) we established a principled train-test procedure in which we first use the presurvey outcomes to determine empirically the existence of a latent dimension common to a set of items that are intended to measure a latent variable of interest, and then we use the post-survey to confirm the results from the presurvey. The post-survey results confirm the factor model results from the presurvey for each scale.

In table S1, we report two sets of statistics from the presurvey factor analyses that demonstrate the convergent validity of each scale constructed from ordinal outcomes.[**] First, the statistical model that we use in the main analysis incorporates measurement models for each scale as a part of the data likelihood. We report the factor coefficients from each model in the first column of Table S1. Since all outcomes are ordinal, each factor coefficient measures the impact of the latent variable on changes of the relevant observed outcome across a given threshold in the log-odds scale. With few exceptions, each log-odds estimate is large, and all are statistically significant. Second, we use descriptive principal components factor analysis (PCA) based on polychoric correlations (estimated in Stata 15) as a second descriptive test of each scale. Each scale yields a single factor solution (with first eigenvalues greater than one and second less than one) and the second column reports the factor coefficients for each resulting factor. With few exceptions, the factor coefficients are large and indicate a good fit.

[§] Through an oversight we used simple randomization among the six labs enrolled from the UCR engineering college, two were assigned to control and four to treatment.

[**] Standard measures of reliability such as Cronbach's Alpha do not apply to ordinal data, since the difference in values on an ordinal scale does not correspond to a distance measure.

Table S1. Factor Model Results

	Model Log Odds Ratio[a] (SE)	PCA Factor Coef.
Relevance of Ethics Discourse scale		
How relevant to your area of work is learning about ethical research practices?	1	0.54
To what extent do you agree that seeking others in your department to discuss ethical research practice is your responsibility as a scientist?	1.30 (0.31)	0.76
To what extent do you agree that seeking out others in your lab to discuss ethical research practices is your responsibility as a scientist?	1.55 (0.34)	0.81
Do you have discussions with members of other labs regarding ethical research practices?	0.42 (0.27)	0.50
Respectful Discussion scale		
If there was a disagreement about research practices, would lab members in your lab make reasonable points and try to make valid arguments?	1	0.65
When discussing research with your lab, does everyone have a real opportunity to speak with no one inappropriately dominating the discussions?	0.94 (0.28)	0.59
When discussing research with your lab, do lab members listen to one another respectfully and courteously?	1.39 (0.30)	0.79
When discussing research with your lab, do lab members seem to hear and understand your views?	0.87 (0.30)	0.58
How confident are you that your lab co-workers will use ethical research practices when conducting research?	1.21 (0.29)	0.67

Table S1 Continued

Lab Disagreement scale		
How often are there disagreements in your lab about research practices?	1	0.79
How often are there disagreements in your lab about authorship?	0.98 (0.27)	0.74
How often are there disagreements in your lab about data management?	1.56 (0.37)	0.82
Reasons for Authorship Policy scale		
Do you understand the rationale for having an authorship plan in your lab?	1	0.90
Do you understand the importance of having an authorship plan in your lab?	2.80 (0.40)	0.95
Do you understand the implications for having an authorship plan in your lab?	1.93 (0.33)	0.85
Reasons for Data Management Policy scale		
Do you understand the rationale for having a data management plan/policy in your lab?	1	0.95
Do you understand the importance of having a data management plan/policy in your lab?	3.19 (0.48)	0.96
Do you understand the implications for having a data management plan/policy in your lab?	2.64 (0.42)	0.91
Preserve Replication Materials scale		
How important to your lab is archiving records of earlier versions of data sets?	1	0.87

Table S1 Continued

How important to your lab is archiving records of earlier versions of manuscript drafts?	2.54 (0.43)	0.88
How important to your lab is archiving records of earlier versions of lab materials?	2.35 (0.41)	0.87
How worthwhile is it to maintain electronic copies of lab materials and methods?	0.57 (0.20)	0.53

N = 184. [a]Posterior means and standard deviations for the factor coefficient structural parameters from the six scaling models.

Statistical Dependence among Respondents within Labs

The statistical model that we use for all of the analyses we report in the main text uses a conditional autoregressive prior to explicitly model the dependence in responses to our surveys that occurs among respondents who are co-located within labs. The model is from (47, chapter7). See (45) for an application of the model to deliberative persuasion in groups. Any dependence that occurs could be due to self-selection of homophilous participants into labs, or to confounding factors within the lab or in the lab environment, or to causal effects within groups in the presence of communication such as persuasion. The model does not distinguish between these sources of dependence nor does it need to for any statistical reasons.

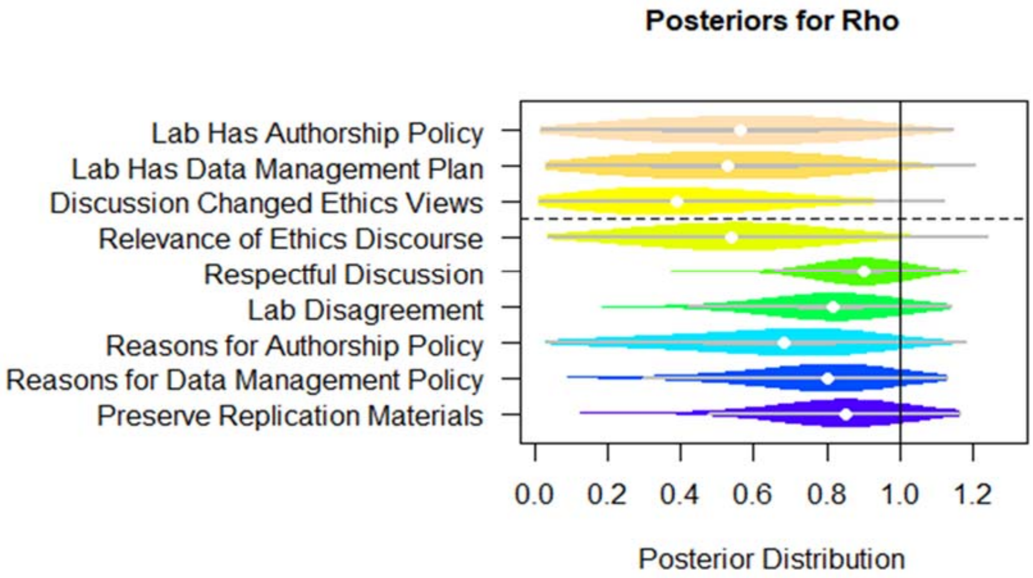


Figure S2. Violin plot (created using 38) for the ρ parameter in the conditional autoregressive model that measures response dependence within labs. N=184.

As we describe above, to model within-lab dependence, the likelihood includes a parameter η_i that is a random effect that nests participants within labs, which in turn is modeled with a conditional autoregressive (spatial dependence) parameter ρ that tests for the degree of within-group dependence in the random effects. We include this spatial dependence structure in each model, both for the models of single outcomes and of scales.

Figure S2 shows the posterior distributions of the ρ parameters for each of the scale outcomes. The vertical line shows the mean of the standard log-normal prior (log-mean of zero) and the rows show the posterior distribution for the dependence parameter for each outcome as a violin plot. As is evident from the figure, the posterior in each case is largely to the left of the prior, indicating very little empirical dependence within labs.

Table S2: WAIC Information Scores, Restricted and Unrestricted Models

	Unrestricted Model WAIC	Restricted Model WAIC
Relevance of Ethics Discourse	625.48	628.28
Respectful Discussion	628.90	640.89
Lab Disagreement	635.90	641.44
Reasons for Authorship Policy	678.09	688.86
Reasons for Data Management Policy	687.76	694.58
Preserve Replication Materials	643.66	660.97

N = 184. Unrestricted model nests respondents in labs using the conditional autoregressive prior (46, chapter 7); restricted model assumes independence between respondents.

This lack of econometric dependence is confirmed in that we find nearly identical results when we re-estimate each model under the assumption of no dependence.[††] In this constrained model we simply set η_i to zero for all i. Under this constraint, we find that the ratio of the treatment effect parameters for all scales under the constraint are either equal to or greater than the unconstrained model, and all of the estimates of these structural parameters are statistically indistinguishable. We use the WAIC statistic (55),

[††] We assign log-normal priors to ρ in order constrain ρ to be positive because it would be very unlikely to see negative correlation within a lab. To evaluate the sensitivity of the results to the prior for ρ, we re-estimated the model for the Relevance of Ethics Discourse scale (the first scale in Figure 3) with a uniform [0,10] prior and again with a uniform [-5,5] prior, and recover identical results in all cases. The estimate for ρ under the log-normal prior is 0.53; under [0.10] is 0.53; and under [-5,5] is 0.38 (the point estimate under [-5,5] is smaller since the posterior has a long tail that dips below zero). But under all of these priors for ρ the treatment effect of interest, α_1, is the same in all cases (0.56 under log-normal, 0.53 under [0,10] and 0.54 under [-5,5], each with the same probability to the left of zero).

which is an assessment of the accuracy of a model in the sense of its ability to fit out of the sample used to estimate the model and so penalizes an unconstrained model if it the additional parameters simply overfit the data (56). We report the WAIC statistic for each of the six scale outcome models in Table S2. Notice that the unrestricted WAIC statistic is lower for the unrestricted model for each scale, indicating a better fit. As a result, we report the unconstrained model in the main text, which as we note above are the conservative estimates.

One possible reason for the absence of dependence in our data is that there are only an average of five respondents per lab and so there is simply not enough respondents per lab to observe a correlation. Another possible reason is that participants with Ph.D. training, or who are receiving Ph.D. training, are not like typical survey respondents and may have greater independence in perceptions and beliefs than is typical.

Maximum Likelihood Estimates for Single Item Responses

The scale outcomes require a bespoke multilevel model implemented in Bayesian MCMC, and for consistency we also analyze the single item models that correspond to Fig. 2 in the main text using the same Bayesian MCMC. Models for these single items also can be implemented using well-established maximum likelihood models in most statistical packages. Table S3 gives the results for the maximum likelihood estimates for the single item responses using the multilevel procedures for logit and ordered logit implemented in Stata 15 in log odds. We summarize these estimates in the main text and provide the full results in table S3.

Interpreting the null effect on data management policy. Note in Table S3 (and in Figure 2 of the main text) that the log odds for the effect of exposure to the training on the respondent's perception that the lab has a data management policy is centered at zero, and we cannot reject the null hypothesis. We did not expect this finding; one of the topics of the training was specifically data management. Although we did not require labs to create a data management policy, the contrast is strong with our other topic in that respondents were very likely to perceive their lab to have an authorship policy as a result of the training.

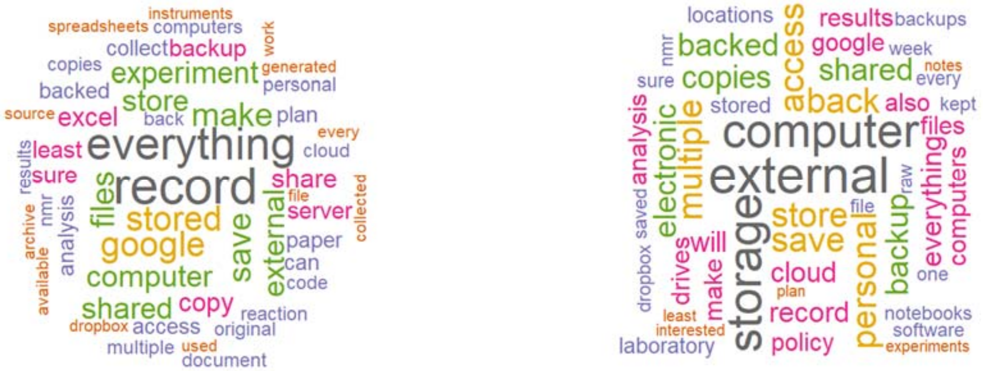
To gain some insight for why there is no apparent treatment effect for data management policy we conducted an exploratory text analysis of open-ended text responses from the survey. For this analysis, we focus on the follow up survey responses only, and we sort by whether the respondent indicated she did or did not perceive the lab to have a data management policy. Within these groups we compare the open-ended text responses to a question asking respondents to type in what they understand to be their own personal data management practices. We then conducted a wordcloud analysis that discarded words that were frequently used by both groups ("data," "lab," "hard," "drive," "notebook," "keep") and generated the wordclouds (using 57) for each group to visualize the comparison. The results are in Figure S3.

While this analysis is entirely exploratory, it is possible that the wordclouds are suggestive of an interesting difference, in that respondents who did not perceive the lab to have a data management policy seem to be less systematic and want to record everything, while those who perceive their lab to have a data management policy seem to focus on technology to implement data storage. Since many scientists develop habits for data management beginning with their initial training, it is possible that these different styles are deeply rooted and changing these habits might require additional focus in the training.

Table S3: Survey Item Results

	Dichotomous Model		Ordered Model
	Data Management Policy	Authorship Policy	Discussion Changed Views
Pretest Response	1.987*	2.006*	0.938*
	(0.431)	(0.534)	(0.221)
Training	0.039	2.137*	0.761*
	(0.423)	(0.490)	(0.362)
Constant	-0.880	-1.740*	
	(0.339)	(0.418)	
Cutpoint 1			0.670
			(0.558)
Cutpoint 2			2.363*
			(0.587)
Cutpoint 3			4.293*
			(0.688)
ρ	0.0	0.0	0.0
N	113	113	108

$^{*}p < 0.05$.



a) Lab Has No DM Plan

b) Lab has DM Plan

Figure S3: Wordcloud frequencies for post-treatment personal data management practices for a) respondents reporting has no data management plan and b) respondents reporting lab has data management plan (created using 56). N=184.

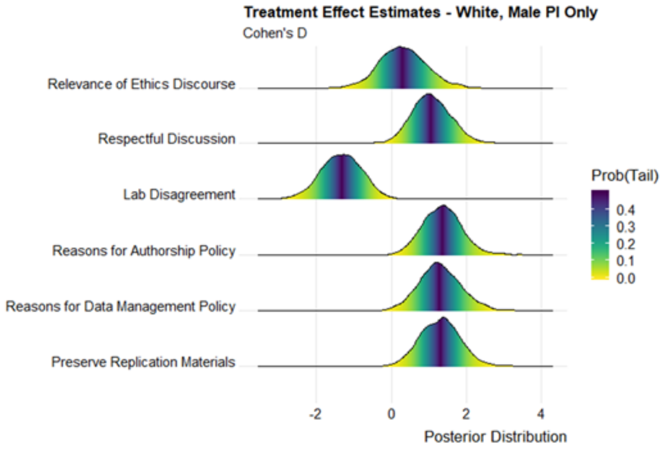
Subgroup Analysis

Our training is motivated by the normative aspirations of deliberative communication, and as we note in the text deliberation is harder to achieve in the presence of inequality. To assess whether inequality affected or moderated our treatment intervention for this sample, we conducted subgroup analyses, where we estimated the treatment effect within subgroups that are underrepresented in STEM and in comparison to PIs.

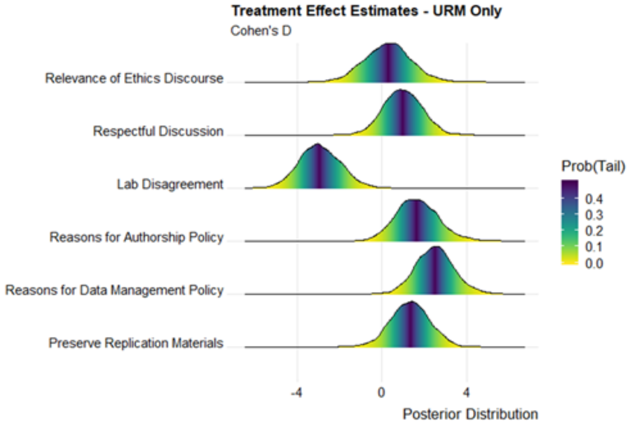
To implement the model to include subgroup conditional effects, we include indicator variables for subgroups as additive terms to equation A.6 and as interactions with the treatment assignment variable. We include indicators for underrepresented minority (self-reported Latino or African American), not male (not all respondents gender identified) and not-PI, with the omitted category being white, male PI. These analyses were not planned and we do not present the results as having validity external to our sample; instead, we report these results to better understand the dynamics that may have happened within our sample. We report the subgroups treatment effect results in Figure S4.

In these results we find no variation across any of the subgroups for any of the scales. Specifically, for each of the scales except for one, the subgroup treatment effects increase both the magnitude and the uncertainty of the estimate while preserving statistical significance, yielding similar posterior distributions across the different categories. For one scale, Relevance of ethics discourse, the results preserve the point estimate for the magnitude while increasing the standard error, and so the model is not sufficiently powered to assess the effect for this dimension among subgroups. All of these results however are identical across subgroups.

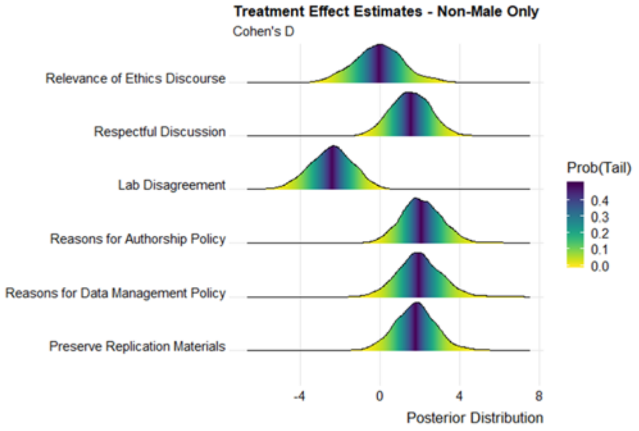
These subgroup results were not planned and so we offer the results as an in-sample diagnostic only. The results would be interesting, for example, if one had priors beliefs that there should be large observed differences between the subgroups. The results would have been especially informative if, contrary to our findings, the null could be rejected for any group in comparison to the baseline PI.



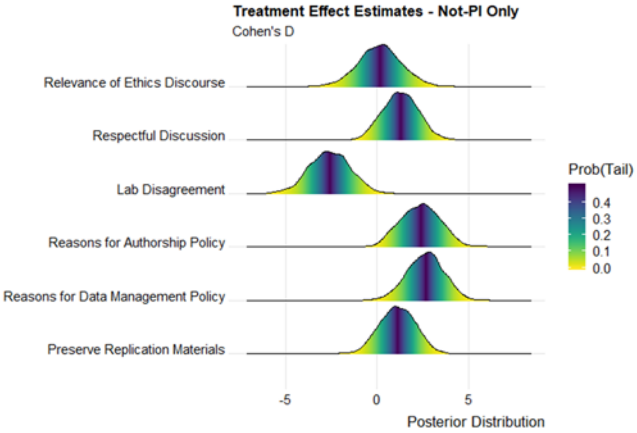
(a)



(b)



(c)



(d)

Figure S4: Conditional effects for four subgroupings: a) White, male PIs (the baseline category), b) under-represented minorities, c) non-male, and d) not-PIs. The model is the same as the one used for Figure 3 of the main text with subgroup main and interactions included in the scale outcome equation (created using 38). N=184.

Preregistration and Constraints on Researcher Degrees of Freedom

Before observing the data we preregistered the study on the OSF at https://osf.io/pfxat. In the preregistration we offer ex ante predictions for the direction of the treatment effect for each of the items we analyze in this paper and our results are consistent with the prespecified predictions, with the one exception of the null finding for the presence of a lab data management plan. A PDF file with a summary of the expectations grouped in the order they appear in the manuscript is included in (51).‡‡

Beyond this prespecification, our design allows for very few researcher degrees of freedom (58). The analytical design we use is an RCT nested in a difference-in-differences model. Since our model has only one covariate specification available, we are not at risk of using covariate selection or searched specifications to yield results. We also report results of constrained models as alternative functional specifications and none yield different results from what we report. Although we analyze a large number of survey items across the six scales, we are not at risk of underestimating standard errors from multiple comparisons since we estimate the treatment effects using a scaling (multilevel) framework (59).

The preanalysis plan gives expectations for the direction for most of the items in the full survey, including those that we do not analyze in this paper. In this paper we analyze most but not all of the items from the survey. The preanalysis plan summary shows the items that we have not yet analyzed. To choose items to focus on in this paper, we used a principled, train-test approach to scale construction where we evaluated the presence of

‡‡ A direct link to the summary preregistration is here: https://www.dropbox.com/s/u75ta9myzjwgjsi/iREDS%20PREANALYSIS%20PLAN.pdf?dl=0

latent factors on the pretest survey, and then confirmed the presence of those factors using the post-treatment survey.

The remaining items are not analyzed but are freely available in the replication material described below. Most of the remaining items are centered on the individual psychology of the respondent, which is only tangentially related to the deliberative outcomes that are the focus of this paper. There are also sets of items that measure respondents' attitudes toward authorship and data management policies, conditional on the lab having such a policy; these items must be analyzed in a selection model and are outside of the scope of this paper.

Replication Materials

We have made all replication materials available at (51); this includes the full project data including the unanalyzed items, as well as all code to implement all analyses and to create the figures. The Bayesian MCMC analyses require the ability to use OpenBUGS which is a programming platform. To facilitate replication for those who are not already familiar with OpenBUGS once can see our training video on how to use OpenBUGS at https://youtu.be/_44_RXTWpRw .

SI References

39. D.B. Rubin, Estimating Casual Effects of Treatments in Randomized and Nonrandomized Studies. *J. Educ. Psychol.* **66**, 688–701 (1974).

40. J. Angrist, J. Pischke, *Mostly Harmless Econometrics* (Princeton University Press, Princeton, NJ, 2008).

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

42. C. H. Achen, Mass Political Attitudes and the Survey Response. *Am. Polit. Sci. Rev.* **69**, 1218–1231 (1975).

43. S. Ansolabehere, J. Rodden, J. M. Snyder, The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting. *Am. Polit. Sci. Rev.* **102**, 215–232 (2008).

44. M. Prior, You've either got it or you don't? The stability of political interest over the life cycle. *J. Polit.* **72**, 747–766 (2010).

45. K.M. Esterling, A. Fung, T. Lee. When Deliberation Produces Persuasion Rather Than Polarization: Measuring and Modeling Small Group Dynamics in a Field Experiment. *Br. J. Polit. Sci.* **forthcoming** (2019).

46. S. Trier, S. Jackman, Democracy as a Latent Variable. *Am. J. Pol. Sci.* **52**, 201–217 (2008).

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

48. J. D. Angrist, G. W. Imbens, D. B. Rubin, Identification of Causal Effects using Instrumental Variables. *J. Am. Stat. Assoc.* **91**, 444–455 (1996).

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

50. D. J. Lunn, A. Thomas, N. Best, D. Spiegelhalter, WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Stat. Comput.* **10**, 325–337 (2000).

51. K. Esterling, Replication Data for: Enhancing Ethical Discourse in STEM Labs: A Randomized Trial of Project-Based Ethics Training (2019), doi:10.7910/DVN/AIPWNU, Harvard Dataverse, V2, UNF:6:OYCOhw9pmqePMLywLuJNRA== [fileUNF].

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

53. M. A. Tanner, W. H. Wong, The Calculation of Posterior Distributions by Data Augmentation. *J. Am. Stat. Assoc.* **82**, 528–540 (1987).

54. B.B. Hansen, J. Bowers, Covariate Balance in Simple, Stratified and Clustered Comparative Studies. *Stat. Sci.* **23**, 219-236 (2008).

55. A. Vehtari, A. Gelman, J. Gabry, Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*. **27** (2017), pp. 1413–1432.

56. R. McElreath, *Statistical rethinking: A Bayesian course with examples in R and Stan* (Chapman and Hall/CRC, 2018).

57. J. Silge, D. Robinson, *Text Mining with R: A Tidy Approach* ("O'Reilly Media, Inc.," 2017).

58. J. P. Simmons, L. D. Nelson, U. Simonsohn, False-Positive Psychology: The Way We Report Studies Privileges False Findings. *PsycEXTRA Dataset* (2012), , doi:10.1037/e636412012-001.

59. A. Gelman, J. Hill, M. Yajima, Why We (Usually) Don't Have to Worry About Multiple Comparisons. *J. Res. Educ. Eff.* **5**, 189–211 (2012).