Appendix for:
Generalizing toward Nonrespondents: Effect Estimates in Survey Experiments Are Broadly Similar for Eager and Reluctant Participants

April 16, 2024

Contents

A Descriptive Statistics of Eager and Reluctant Respondents

Variable	**Eager**, N = 21,309[1]	**Reluctant**, N = 13,490[1]
Female	12,201 (57%)	7,710 (57%)
Age	49 (17)	44 (16)
Party ID (7-pt)		
Democrat	10,373 (49%)	6,603 (49%)
Republican	7,744 (36%)	4,770 (35%)
Independent	3,192 (15%)	2,117 (16%)
Education		
Some college	8,728 (41%)	5,749 (43%)
BA or higher	8,451 (40%)	4,479 (33%)
HS degree	3,423 (16%)	2,529 (19%)
No HS degree	707 (3.3%)	733 (5.4%)
Race		
White	14,444 (68%)	8,099 (60%)
Hispanic	2,463 (12%)	2,565 (19%)
Black	2,589 (12%)	1,801 (13%)
Multiracial, non-Hispanic	690 (3.2%)	453 (3.4%)
Asian	720 (3.4%)	399 (3.0%)
Other	403 (1.9%)	173 (1.3%)
Income		
Bottom quintile	5,233 (25%)	3,520 (26%)
Fourth quintile	4,975 (23%)	3,139 (23%)
Second quintile	4,213 (20%)	2,573 (19%)
Third quintile	4,187 (20%)	2,494 (18%)
Top quintile	2,701 (13%)	1,764 (13%)
Has Internet	19,246 (90%)	11,927 (88%)
Religiosity	4.07 (2.65)	4.03 (2.60)

[1] n (%); Mean (SD)

B NORC AmeriSpeak Panel Details

The NORC AmeriSpeak Panel uses a two-stage sampling procedure, a full description of which can be found in the NORC white paper "Technical Overview of the AmeriSpeak Panel; NORC's Probability-Based Household Panel".

> AmeriSpeak Panel recruitment is a two-stage process: (i) initial recruitment using USPS mailings, telephone contact, and modest incentives, and (ii) a more elaborate NRFU recruitment using FedEx mailings, enhanced incentives, and in-person visits by NORC field interviewers.
>
> For the initial recruitment, sample households are invited to join AmeriSpeak online by visiting the panel website AmeriSpeak.org or by calling a toll-free telephone line (inbound/outbound supported). Both English and Spanish languages are supported for online and telephone recruitment. The initial recruitment data collection protocol features the following: an over-sized pre-notification postcard, a USPS recruitment package in a 9"x12" envelope (containing a cover letter, a summary of the privacy policy, FAQs, and a study brochure), two follow-up postcards, and contact by NORC's telephone research center for sample units with a matched telephone number.
>
> For the second stage NRFU recruitment, a stratified random sample is selected from the non-respondents of the initial recruitment. Units sampled for NRFU are sent a new recruitment package by Federal Express with an enhanced incentive offer. Shortly thereafter, NORC field interviewers make personal, face-to-face visits to the pending cases to encourage participation. Once the households are located, the field interviewers administer the recruitment survey in-person using CAPI or else encourage the respondents to register online or by telephone.

Standard incentives to participate in the panel are $5 included in an initial recruitment mailing with an offer of $20 for joining the panel (some earlier panelists were offered only $2 initially, and a small number of targets from tough to reach populations were offered $25 for joining the panel). By contrast, initial non-respondents who were selected for NRFU recruitment were sent FedEx packages with more elaborate recruitment materials including $10 in the mailer and an offer of $50 upon joining the panel. Furthermore, the vast majority of NRFU recruits received in person visits (door knocks) from trained recruiters. 84% of NRFU recruits received this in person contact prior to joining with almost all of the remaining 16% having joined prior to in person contact, but after receiving the mailer with enhanced incentives (a very small percentage of these recruits joined after being selected for NRFU but before receiving the enhanced incentives mailer).

B.1 Removing Duplicate Respondents

NORC fields its TESS-funded surveys on a random sample of its AmeriSpeak panel. This means that any two TESS samples may share a small number of the same respondents. For the purposes of reporting sample demographics and learning the random forest models predicting who is a reluctant respondent, then, we must be careful not to count the same respondents multiple times (this is particularly important with regard to out-of-sample prediction). Because the TESS data do not contain unique respondent identifiers from NORC that would allow for identification of respondents who appear in multiple studies, we removed duplicate respondents using demographic data. Specifically, we removed duplicate entries for respondents with the same NRFU status, gender, education, employment status, home type, income, state of residence, marital status, internet status, phone type, religious attendance, metropolitan residence, party identification, housing, household size, and race. We did not use age a unique identifier because panelists' ages get updated throughout their time in the panel. This procedure leaves us with 34,799 unique panelists.

C Bias in Estimation of Population Effects

In this section we derive the bias in the difference-in-means estimator, among the survey experimental respondents, relative to the population average treatment effect in the target population. The proof follows that in Miratrix et al. (2018), which provides extensions for Hájek style estimators and equiprobable, non-fixed n designs, in which this proof provides the asymptotic bias. Huang et al. (2021) provides similar proofs for when the target population is an infinite superpopulation.

$$
\begin{aligned}
\text{bias} &= \mathbb{E}[\hat{\tau}_{dim}] - \tau \\
&= \mathbb{E}_{\mathcal{R}}[\mathbb{E}[\hat{\tau}_{dim}] \mid \mathcal{R}] - \tau \\
&= \mathbb{E}_{\mathcal{R}}\left[\frac{1}{n}\sum_{i}^{N} R_i\left(Y_i(1) - Y_i(0)\right)\right] - \frac{1}{N}\sum_{i}^{N}\left(Y_i(1) - Y_i(0)\right) \\
&= \frac{1}{n}\sum_{i}^{N}\pi_i\tau_i - \frac{1}{N}\sum_{i}^{N}\tau_i \\
&= \frac{1}{N}\sum_{i}^{N}\pi_i\frac{N}{n}\tau_i - \frac{1}{N}\sum_{i}^{N}\tau_i \\
&= \frac{1}{N}\sum_{i}^{N}\frac{\pi_i}{\bar{\pi}}\tau_i - \frac{1}{N}\sum_{i}^{N}\tau_i \\
&= \frac{1}{N}\sum_{i}^{N}\left(\pi_i^* - 1\right)\tau_i \\
&= \operatorname{Cov}\left(\pi_i^*, \tau_i\right) \\
&= \rho_{\pi_i^*\tau_i}\sigma_{\pi_i^*}\sigma_{\tau_i}
\end{aligned}
$$

Line 2 follows from the law of iterated expectations, where we denote the outer expectation with a subscript $\mathcal{R}$ to make clear the expectation is over repeated realizations of the survey respondents. Line 3 comes from the well known fact that the difference-in-means estimator is unbiased over repeated treatment assignments, under complete randomization, within a given survey respondent sample. Line 4 follows from the definition of π_i, and lines 5-7 follow from algebraic manipulation. Line 8 follows from the definition of covariance, also outlined on page 9 of the supplementary materials of Miratrix et al. (2018). Line 9 follows from an alternative definition of covariance.

D Survey Experiments Analyzed

Table D2 lists each of these studies including author(s), target population (e.g., general, self-identified partisans, BA degree or higher), the sample size of the entire survey, the sample size used to test the main hypothesis we identified (sometimes these effects were estimated only based on some subset of respondents), and the study's title from the TESS application.

Study No.	Author	Target Pop.	N Full Study	N Re-analysis	Title
1	Shannon	General	2034	1323	Are Americans Willing to Reject a Fiscal Benefit to Exclude Immigrants from Public Entitlements?*
2	Powell, Doan, and Quadlin	General	2034	1023	Factors Affecting Attitudes toward Transgender Bathroom Use
3	Williamson	General	1527	1018	The Taxpayer Gap: Perceptions of the Taxpaying Population and Opposition to Welfare Spending*
4	Tak	General	1280	1160	Gender Inequality in Product Markets
5	Farrow	General	2034	974	Does Misery Love Company? Exploration of a Strategic Intervention to Improve Well-being
6	Geoffrey Wallace	General	2007	1021	International Law, (Non)Compliance, and Domestic Audience Costs*
7	Haaland and Roth	General	1542	1505	Beliefs about Racial Discrimination
8	Mutz	White US Adults	1011	673	The Political Impact of Others' Job Loss: Personifying the Enemy*
9	Baum	General	2930	2930	Crime Reporting and Adjudication in US Rape Culture*
11	Bougher	US Adults self-ID as D or R	1447	440	Issue (Dis)agreement and Intergroup Bias in Affective Polarization*
12	Simas	US Adults with known political party ID	2796	1108	Ambiguous Rhetoric and Legislative Accountability*
13	Ahler and Sood	US Adults self-ID as D or R	2222	1447	The Social Construction of Partisanship: Misperceptions About Party Composition and Partisan Identification*
14	Schnabel	General	2789	2746	Are Religions Gender-Typed? The Perceived Femininity and Masculinity of Christians, Jews, Muslims, and Atheists

∗ Denotes study categorized as political science

Table D2: Reanalyzed TESS Studies

Study No.	Author	Target Pop.	N Full Study	N Reanalysis	Title
15	Cheng and Wen	General	3077	1834	Understanding Public Perceptions of Absolute and Relative Social Mobility
16	Dietze and Craig	General; middle-class dropped	1816	1799	How Social Class and the Framing of Income Inequality affect Solidarity Within & Across Groups
18	McCabe	General	2016	792	Public Opinion and Attributions for Health Care Costs*
19	Ryan	General	2056	1722	Are Losers Gullible? A New Test of Ideological Asymmetry in Conspiracy Beliefs*
20	Bandara	General	4064	1400	A Randomized Experiment to Test the Effects of Message Frames on Social Stigma and Support for Punitive Policies towards Individuals with Prior Drug Convictions
21	Chu and Lee	General	3429	3422	Race, Religion, and American Support for Humanitarian Intervention*
22	Mireles	General	2330	1191	Women's College Advantage and Public Perception of College Value in the Labor Market
23	Kennedy and Horne	General	2595	1282	Accidental Environmentalists: Examining the Effect of Income on Positive Social Evaluations of Environmentally-Friendly Lifestyles
24	Hankinson and de Benedictis-Kessner	General	2008	2000	Burden Sharing and Collective Action: A Study of Opinion on Opioid Treatment Funding*
25	Terman	General	1912	766	Human Rights Shaming, Compliance, and Nationalist Backlash*
27	Harbridge-Yong and Paris	General	2101	1366	You Can't Always Get What You Want: How Majority-Party Agenda-Setting and Ignored Alternatives Shape Public Attitudes*
28	Shannon	General	2253	2242	Does Harsh Language Referring to Immigrants Translate into Harsher Preferences for Immigration Policies–Or Is It All Politics?*
29	Busby, Howat, Rothschild, and Shafranek	General	2015	994	Not All Stereotypes Are Equal: Consequences of Partisan Stereotypes on Polarization*
30	Morgan	General	2019	1366	A Question-Wording Experiment on Support for Free Expression

$*$ Denotes study categorized as political science

Table D2: Reanalyzed TESS Studies

Study No.	Author	Target Pop.	N Full Study	N Reanalysis	Title
31	Silverman, Kent, and Gelpi	General	1340	776	Can Factual Misperceptions be Corrected? An Experiment on American Public Fears of Terrorism*
32	Yadon	African American US Adults	1045	202	The Politics of Skin Color: Skin Color as a Politicized Identity for African Americans*
33	Hamilton, Quadlin, and Powell	General	2005	191	Whom Do You Believe? Assessing Credibility of the Accuser and Accused in Sexual Assault
34	Brower	BA degree or higher	1030	205	Reframing Women's Issues: How Intersectional Identity Frames affect Women's Political Attitudes*
35	Krupnikov	General	2005	1982	The Partisan Gender Gap: Genuine Attachment or Social Motivation?*
36	Calarco	General	2005	1293	Public Perceptions of Prenatal Alcohol Consumption
37	Rifkin and Cutright	General	1200	1150	Introducing a Novel Framework for Understanding The Relationships Between Busyness, Idleness, and Happiness
39	Hankinson and de Benedictis-Kessner	General	3112	2374	How Group Identity Shapes Opioid Treatment Policy Opinion*
40	Thorson	General	2118	2084	Effects of Misinformation News Coverage on Media Trust*
41	Melin	General	1682	1676	Testing a Theory of Hybrid Femininity
42	Vogler and Petsko	General	3010	2820	Precarious or Policed Sexualities? How Race and Gender Affect the Categorization of Sexual Behaviors
43	Klar	General	2118	1379	Gender Versus Party? Do Abortion Frames Affect Issue Engagement?*
44	Cohen	General	1610	1528	Social Class, College Debt, and the Purpose of College
45	Blair and Schwartz	General	2342	759	Do Women Make More Credible Threats? Gender Stereotypes and Crisis Bargaining*
46	Margolis	Christian US Adults	2902	2900	Evangelical or Born-Again Christian: Unpacking a Double-Barreled Question*
47	Jakubiak	Married US Adults	1140	569	Do the Benefits of Receiving Affectionate Touch Generalize Beyond Satisfied Couples?
48	Grace and Doan	General	5028	2495	Factors Affecting Public Opinion on Transgender Medical Care Refusal

∗ Denotes study categorized as political science

Table D2: Reanalyzed TESS Studies

Study No.	Author	Target Pop.	N Full Study	N Reanalysis	Title
50	Headley, Blount-Hill, and St. John	General	732	706	Affective Architecture: Isolating the Influence of Physical Environment on Perceptual and Behavioral Attitudes toward Police
51	Zhu and Yzer	US adults 21+ who drink alcohol	789	254	Does Self-affirmation Influence Health Message Processing through Changing Construal Level?
52	Hollin	General	2138	1414	Price Disclosure for Direct-to-Consumer Pharmaceutical Advertising: Price Transparency, Information Asymmetry and Consumer Behavior
53	Stoker, Lerman, and Sahn	General	3576	1786	Equivalency Framing of Societal Problems and Policy Solutions*
54	Weisshaar	Employed US Adults	1814	896	An Imperfect Match? How Gender and Race Influence Perceptions of Job Applicants by Qualification Levels
56	Bai	General	1501	1490	Mechanical Asians and Animalistic Blacks: The Political Implications of The Symmetry of Two Forms of Dehumanization in Racial Perceptions*

∗ Denotes study categorized as political science

Table D2: Reanalyzed TESS Studies

E Coding Scheme

As discussed in the paper, our goal was to extract one average treatment effect per TESS study to obtain a sample of "typical" survey experiments in social science. Though any of the ATEs discussed in the TESS proposals may have counted as "typical," we focused on what could be considered each study's primary analysis. To ascertain which condition and outcome variable constituted a primary analysis, we consulted a series of sources and deferred to the most authoritative one. The top source was the researcher's response to our survey, if we received one. Requests were sent out twice, about three weeks apart, to the authors listed on each study's publicly available TESS proposal on the OSF (Open Science Foundation, osf.io) website. We received 12 (24%) survey responses in total. The survey responses (see questionnaire below) told us which conditions constituted their primary treatment and control and which variable was their primary outcome, as well as how they coded these variables.

For those without a survey (38 of the 50 studies), the next source we referred to was a published article or a working paper that used the TESS data. We looked for which ATE most closely reflected their primary research question. Sometimes this was clear and unambiguous. Other times, an argument could be made for more than one ATE. In those cases, we chose the one that appeared first in the text. Also, whenever possible, we collapsed over treatment conditions so as to maximize sample size. For example, if a study were comparing the effect of a reading about a Black political candidate versus a white candidate and there were two conditions for each, a male and female one, then we would collapse across gender and code the two Black conditions as treatment and the two white conditions as control. Also to avoid under-powered tests, whenever possible we chose analyses that did not involve any moderation effect or interaction terms.

The next and last source to which we referred was the publicly available TESS proposal for the study. We followed the same process as with the published papers, focusing on the primary research question and the experiment that most closely matched it, and when there was ambiguity, we chose the first-mentioned ATE.

To illustrate this process with a slightly more complicated example than the one used in the main body of the paper, we can review the decision-making process used for Study 11. The study's main question, indicated in the TESS proposal, is whether voters feel warmer toward political candidates from the opposing party if they share policy positions. If so, then partisan affective polarization is partly driven by differences over policy. The experimental design contains several conditions pertaining to within-party contests, which the researcher included for other research questions. Since the study was about partisan affective polarization, we focused on the "general election" conditions in which a Democrat ran against a Republican and dropped the other conditions. Within these "general election" conditions were a control group and two possible treatment groups, one using salient policies and another using less salient ones. We coded both of the latter into the treatment group. Following the TESS proposal, we used the absolute value of the difference in feeling thermometer scores toward each candidate as the outcome measure. Here and in the other studies, we divided the outcome variable by its standard deviation in the control group. The ATE among eager respondents was -0.79 (SD = 0.12, DF = 268) and among reluctant was -0.63 (SD = 0.15, DF = 168). The average effects of both groups are similar in magnitude and not statistically different from each other, suggesting both responded in a similar fashion to the treatment. They both felt warmer toward out-party candidates when they shared issue positions by about two-thirds of a standard deviation. The data point from this study falls somewhere near the diagonal in the lower left quadrant of Figure 2.

E.1 Coding Scheme Robustness Check

To see how robust our results are to the coding scheme we followed, we can compare the Deming regression estimates from the coding we did before we received the researchers' survey responses to those we obtained after incorporating their responses. In other words, would our results have changed had we never incorporated researchers' feedback on how they conducted their primary analyses? In 7 out of the 12 of the surveys we received, responses indicated that our initial definition of the main treatment effect of interest was consistent with what the researcher viewed as their main treatment effect of interest. In the remaining 5, researchers' survey responses generally suggested that they viewed a different condition or dependent variable as defining the main treatment of interest, typically for studies in which the TESS proposal included several dependent variables and/or conditions.

The estimated intercept and slope from the Deming regression using our 50 pre-survey codings are -0.020 (s.e. = 0.005) and 0.995 (s.e. = 0.0843), respectively. As in the main text, these estimates also suggest strong correspondence between eager and reluctant ATEs.

The robustness of our results suggests they generalize to other treatment arms and ATEs we could have analyzed but did not because they were not the "primary" analysis. It's important to not stretch this generalization too far, however. Some experiments very well could exhibit significant heterogeneity. Party cue treatments, for instance, in which subjects receive an argument for or an endorsement of a policy from a party elite, should be more persuasive when coming from an in-party elite as opposed to an out-party elite. In that case, respondents' party ID should moderate the effect of treatment. In many cases, and in the studies using party cues in our data set, researchers code the treatment for whether it matches respondents' party identification (e.g., a respondent is considered treated if the source of the argument they hear is someone from the political party the respondent themselves identifies with). Then we would find treatment effect homogeneity if in-party cueing affects Republicans and Democrats alike.

F Deming Regression

To estimate ATEs for eager and reluctant respondents, we fit linear regressions predicting the outcome variable (divided by its standard deviation in the control group) with binary treatment variables and heteroskedasticity-robust standard errors using `lm_robust` in the `estimatr` R package (Blair et al. 2022). We then estimated a Deming regression to assess how similar the eager and reluctant ATEs were. Deming regression is a special case of linear regression that minimizes the squared residuals in both the vertical and horizontal directions (weighted by the respective variances of the eager and reluctant ATEs in our case). It is often used to test the similarity of two measurement strategies when there might not be a clear dependent variable and independent variable. With the `deming` function in the `deming` R package (Therneau 2018), we regressed the reluctant ATEs on the eager ATEs, and used the standard error from the linear model as our estimate of the standard deviation of each ATE. Coefficient variances were estimated using a block bootstrap in which we cluster studies conducted on the same respondents to account for a few sets of studies being fielded together on common surveys.

F.1 Deming Regression Estimates

Subgroup	Intercepts	SE	Slopes	SE	N
All	-0.020	0.010	1.026	0.063	50
Men	-0.018	0.015	1.169	0.089	47
Women	-0.019	0.014	0.925	0.091	50
Age: 18-39	-0.010	0.020	0.843	0.109	50
Age: 40-59	0.006	0.031	1.282	0.163	49
Age: 60+	-0.066	0.017	1.075	0.132	49
Democrats	-0.017	0.017	0.879	0.087	50
Independents	-0.052	0.043	1.539	0.475	46
Republicans	-0.024	0.026	1.307	0.200	49
HS or less	0.003	0.032	1.385	0.339	48
Some college	-0.025	0.015	0.985	0.090	49
College or more	-0.025	0.016	1.055	0.084	50
Low-Income	-0.023	0.016	0.904	0.149	50
Mid-Income	-0.028	0.019	1.193	0.125	50
High-Income	-0.005	0.028	1.213	0.110	49
Whites	-0.018	0.013	1.148	0.086	48
Non-Whites	-0.013	0.022	0.864	0.158	47
Metro	-0.018	0.012	1.015	0.061	50
Non-Metro	-0.018	0.041	1.286	0.331	45
Landline	-0.051	0.026	1.302	0.630	46
Cellphone	-0.016	0.013	1.040	0.074	50

Table F3: Coefficient Estimates and Bootstrapped Standard Errors from Deming Regressions of Eager ATE on Reluctant ATE

G Analyses Using Only Political Science Studies

We remade Figures 2 and 3 from the main text using only studies from political science to assess whether there may have been different types or degrees of treatment effect heterogeneity lurking in the politically charged experiments. Because the Deming regression slope for independents was the largest of any subgroup in the main paper's results (using all 50 studies), we also wanted to see if eager and reluctant independents responded differently to political treatments in particular.

We classified 29 of the 50 studies as coming from the field of political science. To decide which studies to classify as political science, we relied first on the official proposal documents—35 of which stated the disciplines from which the study came. Of these 35 study proposals, 17 contained the term "political science" or "political psychology" on their title page. Of the remaining 15, 12 were determined to be from political science, based on the study's title or authors' occupation. If the dependent variable was an attitude or behavior pertaining to politics, we categorized the study as political science. Studies categorized as political science are denoted with an asterisk (∗) after their title in Table D2.

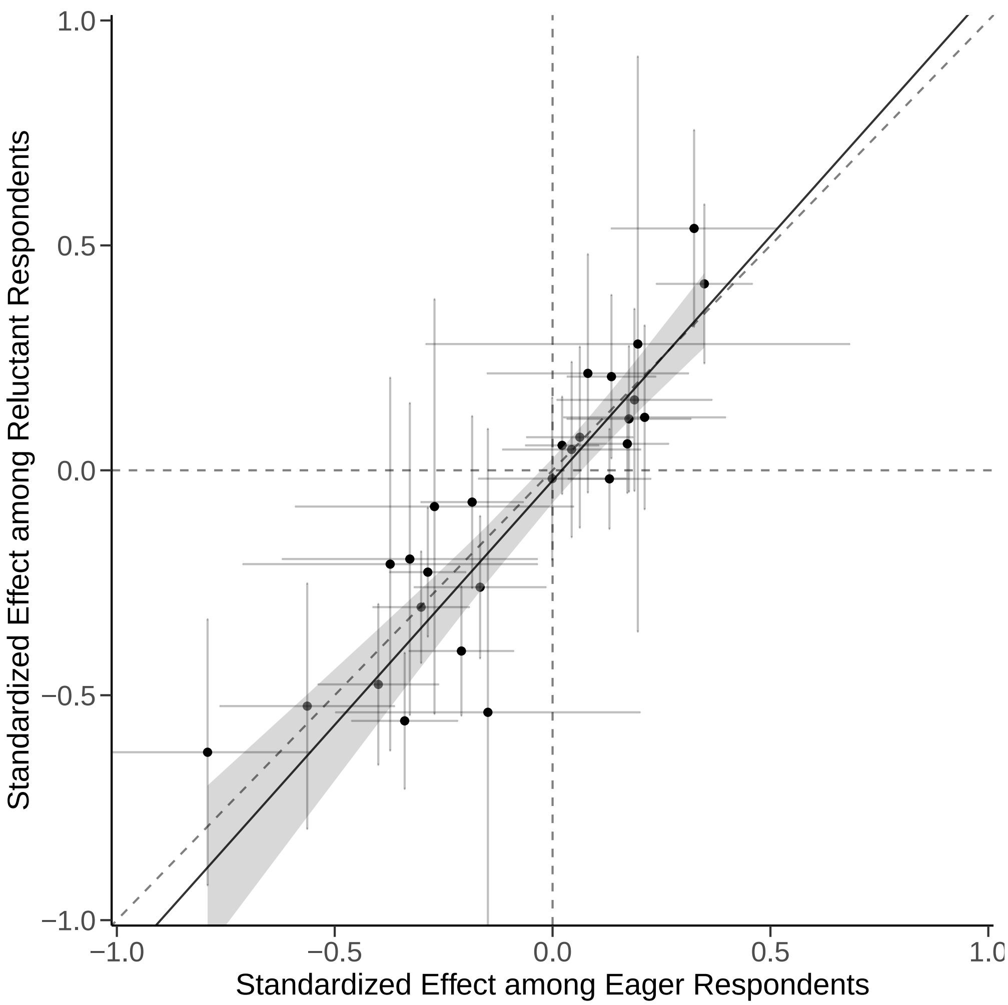

Figure G1: Eager and Reluctant ATEs (Political Science Studies Only)

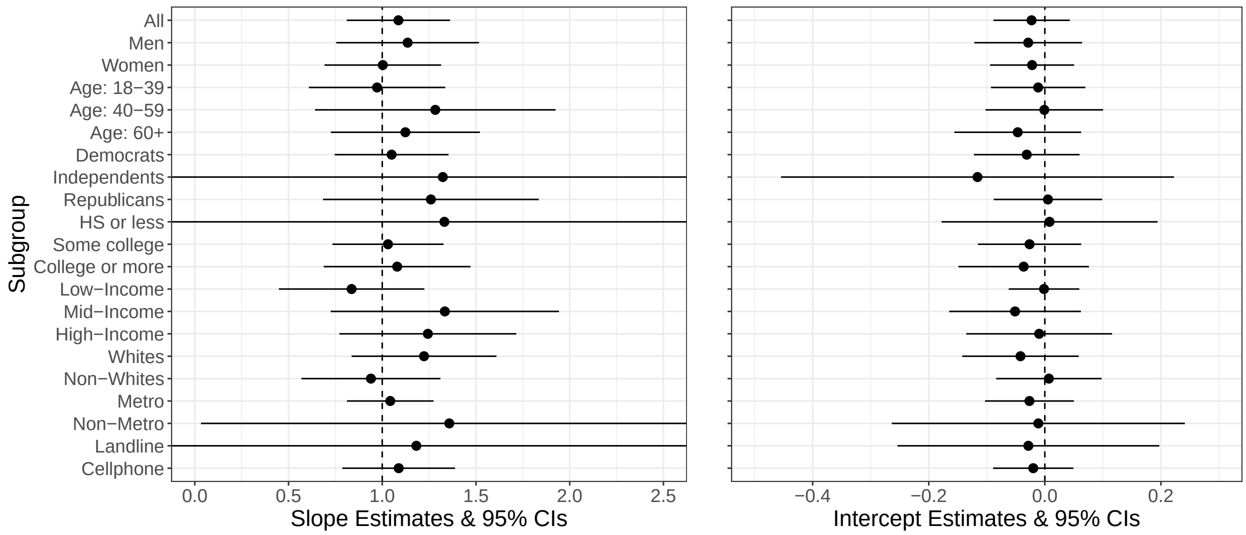

Figure G2: Deming Regression Estimates for Eager and Reluctant Respondents by Subgroup (Political Science Only)

Subgroup	Intercept	SE	Slope	SE	N
All	-0.023	0.025	1.086	0.108	27
Men	-0.029	0.039	1.135	0.140	24
Women	-0.022	0.029	1.003	0.149	27
Age: 18-39	-0.011	0.036	0.972	0.174	27
Age: 40-59	-0.001	0.050	1.283	0.238	26
Age: 60+	-0.047	0.037	1.123	0.159	26
Democrats	-0.031	0.033	1.050	0.143	27
Independents	-0.116	0.106	1.323	0.625	23
Republicans	0.005	0.043	1.260	0.208	26
HS or less	0.008	0.075	1.331	0.645	25
Some college	-0.026	0.038	1.031	0.139	26
College or more	-0.036	0.044	1.080	0.170	27
Low-Income	-0.001	0.029	0.836	0.144	27
Mid-Income	-0.051	0.035	1.334	0.192	27
High-Income	-0.010	0.056	1.243	0.157	26
Whites	-0.042	0.030	1.223	0.113	25
Non-Whites	0.007	0.041	0.940	0.168	24
Metro	-0.027	0.028	1.043	0.100	27
Non-Metro	-0.011	0.273	1.358	2.671	25
Landline	-0.028	0.132	1.182	5.293	25
Cellphone	-0.020	0.030	1.088	0.123	27

Table G4: Coefficient Estimates and Bootstrapped Standard Errors from Deming Regressions of Eager ATE on Reluctant ATE (Political Science Only)

H Comparing Subgroup Effects across Studies

Figure H3 plots the subgroup-specific treatment effect estimates for eager and reluctant respondents. For example, each point in the top-left pane plots the treatment effect estimated among men who were eager respondents (on the horizontal axis) against the treatment effect estimated among men who were reluctant respondents (on the vertical axis), with each point representing one of the 50 studies in our data.

As implied by the Deming regression estimates presented in Figure 2 of the main text, these subgroup-specific plots show little if any evidence of systematic differences between these two sets of estimates for any of the subgroups considered. Note that some of these subgroups tend to have relatively small sample sizes (and correspondingly large confidence intervals for their estimates). Where there are somewhat more precise estimates, however, we see these points lining up close to the 45 degree line indicating similar effects on average among eager and reluctant respondents within a given subgroup.

Figure H4 plots, for each study separately, the estimated treatment effects for eager and reluctant respondents among each of the subgroups shown in Figure 2 in the main paper. For most of these studies there is little evidence of heterogeneity between these subgroups. The studies with the most variable estimates across subgroups also tend to be the ones with the largest confidence intervals for the effect estimates (which are typically those with smaller sample sizes and/or dependent variables with more random variability). There is little overall evidence of notable differences between subgroup-specific effects between eager and reluctant respondents for these studies.

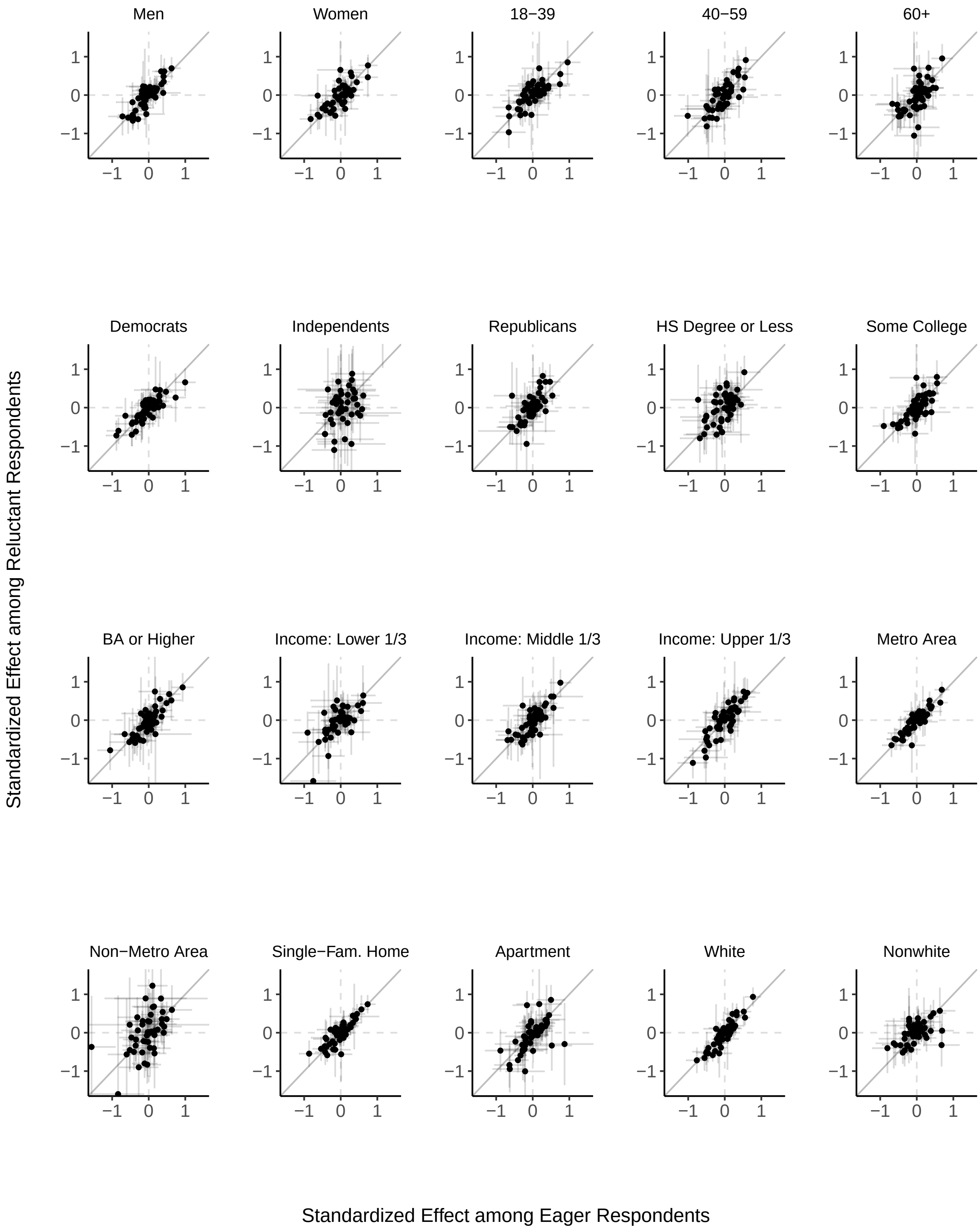

Figure H3: Eager and Reluctant ATEs for Subgroups

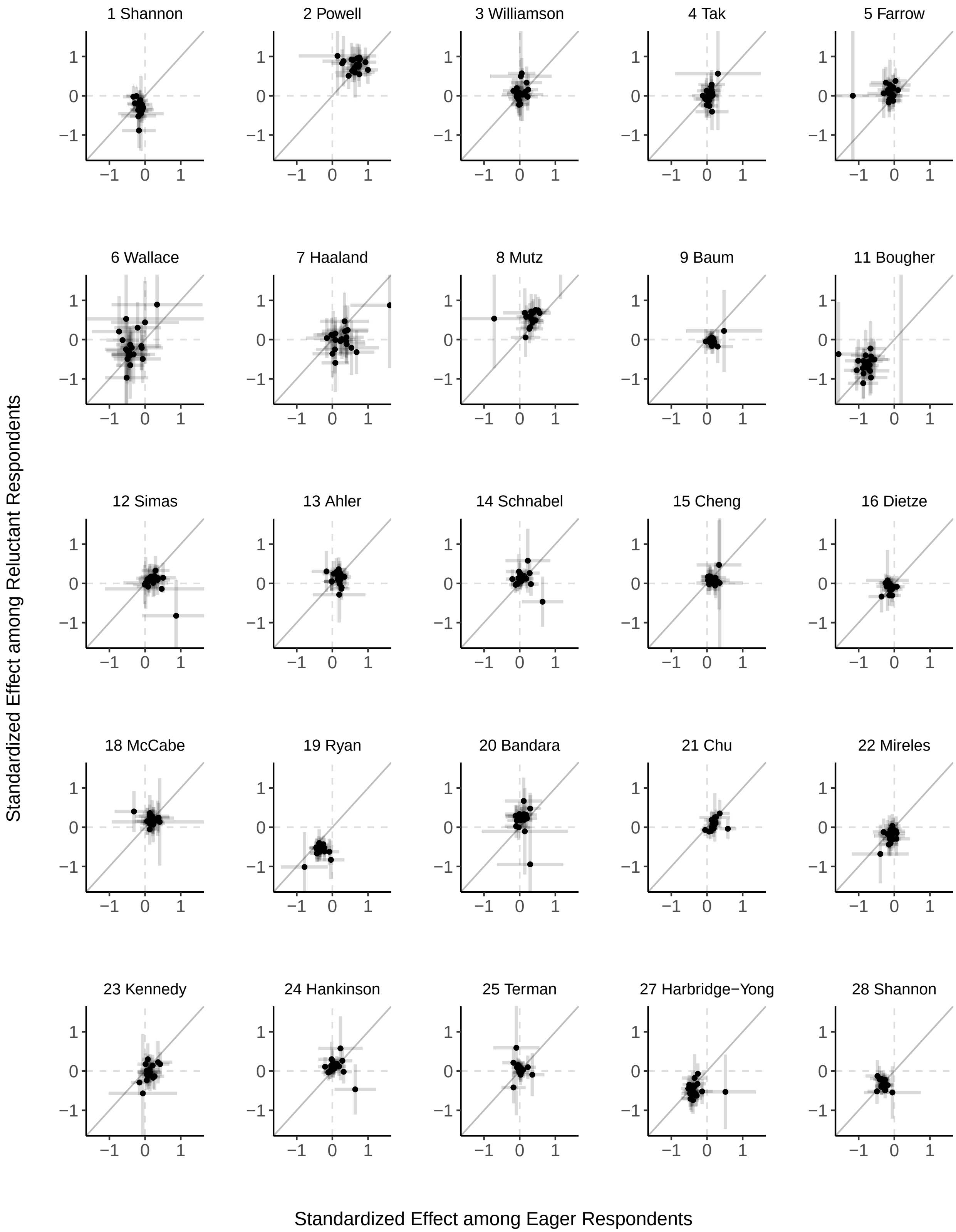

Figure H4: Eager and Reluctant ATEs for Subgroups in Each Study

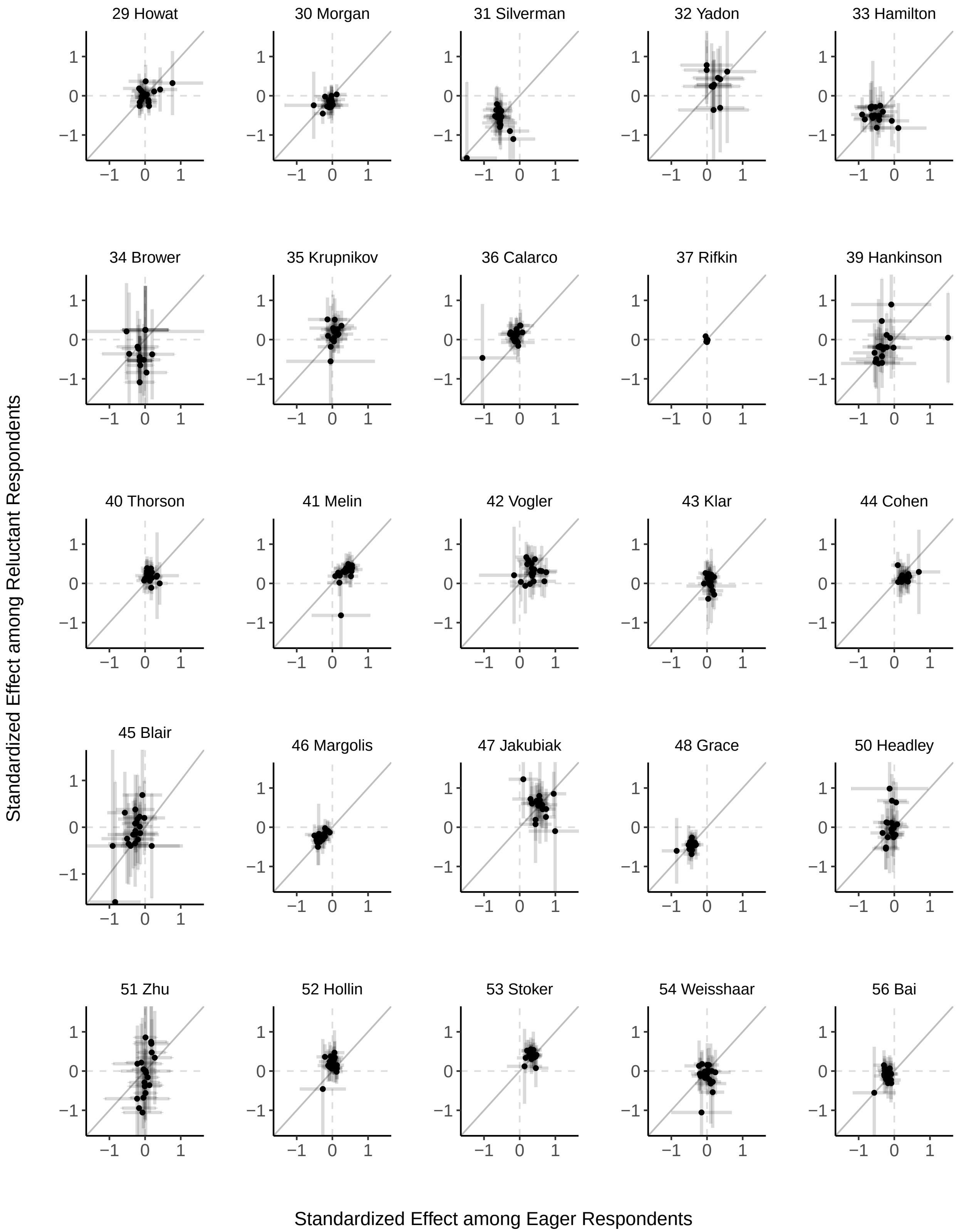

Figure H4: Eager and Reluctant ATEs for Subgroups in Each Study

Predicting NRFU with a Random Forest Model

Figure H5 presents two plots of variable importance derived from the `randomForest` model predicting NRFU (Liaw and Wiener 2002). The left panel shows the variables in descending order according to how much they improve the model's accuracy in classifying respondents as reluctant. The right panel shows how much the variables decrease the "impurity" of the model's predictions. Some demographic variables such as income, age, and education are consistently useful in predicting NRFU. Others, like gender and race, are not. Identity-related variables such as religious attendance and party identification also contribute to the model's effectiveness, whereas household characteristics like telephone and internet service, home type, and size of household do not.

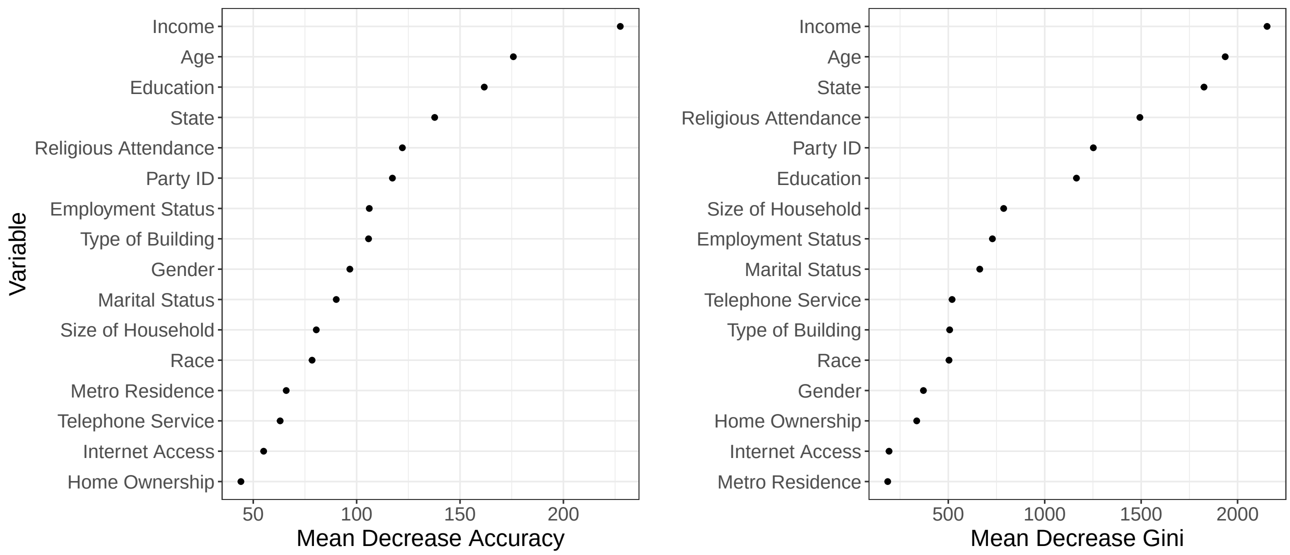

Figure H5: Variable Importance

I Correlations of $\widehat{\tau}_i$ and $\hat{\pi}_i^*$

Table I5 presents the Pearson r correlations of the individual-level treatment effects, $\widehat{\tau}_i$, and the predicted normalized propensities of being a reluctant (NRFU) respondent, $\hat{\pi}_i^*$. P-values associated with the correlations are also presented; 0s indicate the p-value is very close to 0. The $\widehat{\tau}_i$ were derived from causal random forest models estimated using the `grf` R package. The $\hat{\pi}_i^*$ were derived from a random forest model estimated using the `RandomForest` R package.

Table I5: Correlations between $\widehat{\tau}_i$ and $\hat{\pi}_i^*$

Study	Estimate	CI Lower	CI Upper	p-value
1	-0.14	-0.21	-0.08	0.00
2	0.05	-0.02	0.12	0.17
3	0.01	-0.05	0.08	0.65
4	0.05	-0.01	0.11	0.10
5	0.26	0.18	0.33	0.00
6	0.13	0.09	0.17	0.00
7	-0.03	-0.08	0.02	0.19
8	0.06	-0.02	0.13	0.15
9	-0.10	-0.22	0.01	0.08
11	-0.06	-0.15	0.04	0.23
12	-0.04	-0.10	0.02	0.21
13	-0.12	-0.17	-0.07	0.00
14	0.03	-0.01	0.07	0.11
15	-0.07	-0.11	-0.02	0.01
16	-0.06	-0.11	-0.01	0.01
18	0.01	-0.06	0.08	0.86
19	-0.42	-0.46	-0.39	0.00
20	-0.01	-0.07	0.04	0.59
21	-0.06	-0.09	-0.03	0.00
22	-0.07	-0.12	-0.01	0.03
23	-0.01	-0.07	0.04	0.59
24	-0.42	-0.45	-0.38	0.00
25	-0.48	-0.53	-0.42	0.00

Table I5: Correlations between $\widehat{\tau}_i$ and $\frac{1}{\pi_i^*}$

Study	Estimate	CI Lower	CI Upper	p-value
27	-0.04	-0.09	0.02	0.19
28	0.03	-0.01	0.07	0.12
29	0.03	-0.03	0.09	0.32
30	-0.04	-0.10	0.01	0.10
31	-0.04	-0.12	0.03	0.21
32	0.14	0.04	0.24	0.01
33	0.02	-0.12	0.16	0.76
34	-0.23	-0.36	-0.10	0.00
35	-0.06	-0.10	-0.02	0.01
36	-0.17	-0.22	-0.12	0.00
37	0.12	0.06	0.18	0.00
39	0.02	-0.01	0.06	0.19
40	0.04	-0.00	0.09	0.05
41	-0.15	-0.20	-0.11	0.00
42	-0.05	-0.09	-0.01	0.01
43	0.02	-0.04	0.07	0.53
44	-0.07	-0.12	-0.01	0.02
45	0.07	-0.00	0.14	0.06
46	0.01	-0.02	0.05	0.51
47	-0.01	-0.10	0.07	0.73
48	-0.00	-0.04	0.04	0.83
50	0.06	-0.01	0.14	0.11
51	0.10	-0.03	0.22	0.12
52	0.09	0.04	0.14	0.00
53	0.01	-0.03	0.06	0.52
54	-0.03	-0.10	0.04	0.39
56	0.19	0.14	0.24	0.00

J Eager and Reluctant ATE Estimates, SEs, and Ns

Study No.	Eager Estimate	Eager SE	Eager N	Rel. Estimate	Rel. SE	Rel. N
1	-0.166	0.077	714	-0.260	0.080	605
2	0.678	0.094	555	0.771	0.101	464
3	0.044	0.081	588	0.046	0.099	426
4	0.071	0.077	623	0.008	0.091	533
5	-0.102	0.085	527	0.048	0.095	443
6	-0.373	0.173	612	-0.209	0.211	405
7	0.249	0.164	891	0.004	0.179	610
8	0.325	0.097	396	0.538	0.111	273
9	0.131	0.049	1678	-0.019	0.057	1248
11	-0.792	0.121	268	-0.627	0.150	168
12	0.211	0.095	635	0.118	0.104	469
13	0.175	0.073	802	0.114	0.082	641
14	0.029	0.057	1581	0.090	0.068	1161
15	0.140	0.064	1044	0.089	0.076	786
16	-0.115	0.060	1027	-0.077	0.071	768
18	0.188	0.091	423	0.157	0.103	365
19	-0.339	0.062	1010	-0.557	0.077	708
20	0.046	0.089	776	0.231	0.091	620
21	0.172	0.049	1988	0.059	0.056	1430
22	-0.076	0.073	706	-0.197	0.089	481
23	0.121	0.070	789	-0.044	0.094	489
24	-0.209	0.062	1207	-0.402	0.073	789
25	0.022	0.043	484	0.055	0.055	278
27	-0.400	0.071	832	-0.476	0.091	472

Table J6: Eager and Reluctant ATEs, SEs, and Ns

Table continued on next page.

Study No.	Eager Estimate	Eager SE	Eager N	Rel. Estimate	Rel. SE	Rel. N
28	-0.302	0.057	1255	-0.304	0.063	983
29	-0.001	0.086	508	-0.019	0.092	482
30	-0.070	0.073	741	-0.172	0.080	621
31	-0.563	0.102	479	-0.524	0.139	293
32	0.196	0.246	129	0.281	0.320	69
33	-0.538	0.119	109	-0.504	0.136	78
34	-0.148	0.177	127	-0.538	0.316	74
35	0.081	0.118	1157	0.216	0.135	821
36	-0.115	0.072	746	0.112	0.090	543
37	-0.001	0.005	659	-0.021	0.006	487
39	-0.328	0.150	1467	-0.197	0.177	903
40	0.135	0.052	1516	0.208	0.092	564
41	0.397	0.061	989	0.310	0.076	683
42	0.360	0.089	1841	0.304	0.121	975
43	0.062	0.063	1007	0.074	0.102	368
44	0.270	0.058	1005	0.142	0.094	519
45	-0.271	0.163	474	-0.081	0.234	281
46	-0.286	0.045	2097	-0.226	0.073	799
47	0.543	0.093	418	0.551	0.165	147
48	-0.412	0.046	1827	-0.421	0.077	664
50	-0.071	0.091	464	-0.047	0.132	238
51	0.029	0.138	214	-0.047	0.292	36
52	0.029	0.064	1027	0.198	0.105	383
53	0.348	0.057	1324	0.414	0.090	458
54	-0.016	0.080	636	-0.037	0.127	256
56	-0.185	0.060	1068	-0.071	0.097	418

Table J6: Eager and Reluctant ATEs, SEs, and Ns (Full Sample)

K Researcher Survey Questionnaire

The following survey was sent to at least one researcher who proposed a TESS project we reanalyzed. We received 12 responses.

NRFU Replications

Start of Block: Demos

Q14 Thank you for participating in this survey about your TESS study. With support of TESS leadership, we are gathering information about all studies conducted recently to evaluate and improve some of the sampling and other processes used. Our analysis does not focus on evaluating or "debunking" any of the TESS studies. Rather, we are interested in assessing sampling procedures used by the vendor.

Your assistance in this short survey will help us to more quickly and more accurately perform this assessment.

Q2 What is your last name?

____

*

Q84 In the email invitation you received, we gave you a 4-digit code. Please enter that 4-digit code below.

____

Q12 Do you have replication code (ideally that uses the raw data from NORC) that you can share with us?

○ No (1)

○ Yes (2)

End of Block: Demos

Page 1 of 5

Start of Block: Code Upload

Display This Question:

If Do you have replication code (ideally that uses the raw data from NORC) that you can share with us? = Yes

Q86 Please upload your replication code here.

End of Block: Code Upload

Start of Block: Treatment Variable

Q5 Our study is focused on replicating one treatment effect per study. We hope to focus on what researchers think of as the main treatment effect of interest, or at least one that is the main treatment effect if there are multiple. Thus, we'd like to ask you some questions about the key variables in your study, beginning now with the **treatment variable**.

In what follows, please use the NORC codebook for variable names if possible. If that is not possible, use as descriptive of a name as possible.

Q7 What **treatment variable** (or combination of treatment variables) would you say is of primary interest in your study?

____

Q11 Using the NORC codebook's variable values, what is the **control** or baseline condition in this experiment? For example, you could write, "VAR123 = 0".

____

Q4 What is the main **treatment** level in this experiment? If you have multiple treatment conditions, please specify the one of primary interest. For example, you could write, "VAR123 = 2".

____

End of Block: Treatment Variable

Start of Block: Outcome Variable

Q79 As we said, our study is focused on what researchers think of as their main analysis. Now we are moving onto the primary **outcome** variable of interest. If your study had multiple outcome variables, please choose only one.

Q9 What is the **outcome** variable of primary interest for your study? If it's an index, please list all constituent items.

____

____

____

____

____

Q10 How do you code your study's primary **outcome** variable? Please copy-paste your code if you have it. If not, please describe how you code it. Code for any standard statistical program (e.g., R or Stata) is fine here.

____

____

____

____

____

End of Block: Outcome Variable

Start of Block: Modeling

Q85 Now onto the modeling and estimation strategy you use.

Q15 How do you estimate your average treatment effect of primary interest?

- ◯ T-test or linear model with no interactions (1)
- ◯ Linear model with interaction terms (2)
- ◯ Other (3) ____

Display This Question:

If How do you estimate your average treatment effect of primary interest? = Linear model with interaction terms

Q16 What moderating variables do you use to create the interaction terms in your model? (Please use names from NORC's raw data.)

____

____

____

____

____

Q19 What is the specific command you use to estimate the primary model/analysis for your study? Please copy-paste your code if you have it. If not, please describe how you code it.

____

____

____

____

____

End of Block: Modeling

Start of Block: Block 5

Submit Those are all our questions. Please click **submit** when you're finished.

Thank you for your time!

End of Block: Block 5

References

Blair, Graeme, Jasper Cooper, Alexander Coppock, Macartan Humphreys, and Luke Sonnet. 2022. *estimatr: Fast Estimators for Design-Based Inference.* R package version 1.0.0.

Huang, Melody, Naoki Egami, Erin Hartman, and Luke Miratrix. 2021. *Leveraging Population Outcomes to Improve the Generalization of Experimental Results.* arXiv: 2111.01357 [stat.ME].

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22.

Miratrix, Luke W., Jasjeet S. Sekhon, Alexander G. Theodoridis, and Luis F. Campos. 2018. "Worth weighting? How to think about and use weights in survey experiments." *Political Analysis* 26 (3): 275–291.

Therneau, Terry. 2018. *deming: Deming, Theil-Sen, Passing-Bablock and Total Least Squares Regression.* R package version 1.4.