Sociological Methods & Research
1-35

DOI: 10.1177/0049124115600614
smr.sagepub.com

Causal Inference in Panel Data With Application to Estimating Race-of-Interviewer Effects in the General Social Survey

Weihua An[1,2] and Christopher Winship[3]

Abstract

In this article, we review popular parametric models for analyzing panel data and introduce the latest advances in matching methods for panel data analysis. To the extent that the parametric models and the matching methods offer distinct advantages for drawing causal inference, we suggest using both to cross-validate the evidence. We demonstrate how to use these methods by examining race-of-interviewer effects (ROIE) in the 2006 to 2010 panel data of the General Social Survey. We find that ROIE mostly concentrate on race-related outcomes and may vary by respondent's race for some outcomes. But we find no statistically significant evidence that ROIE vary by the interview mode (i.e., in person vs. by phone). Our study has both methodological and substantive implications for future research.

[1] Department of Statistics, Indiana University, Bloomington, IN, USA
[2] Department of Sociology, Indiana University, Bloomington, IN, USA
[3] Department of Sociology, Harvard University, Cambridge, MA, USA

Corresponding Author:
Weihua An, Departments of Statistics and Sociology, Indiana University, 752 Ballantine Hall, 1020 East Kirkwood Avenue, Bloomington, IN 47405, USA.
Email: weihuaan@indiana.edu

Keywords

causal inference, panel data, matching, difference-in-difference, interviewer effects, General Social Survey

Introduction

Panel data offer both opportunities and challenges for causal inference. One key advantage of panel data over cross-sectional data is that it allows researchers to better handle the effects from unobserved time-invariant factors. At the same time, a key problem in analyzing panel data is to account for possible serial correlations in the error terms for each individual. The goal of this article is to present models and methods for analyzing panel data, with particular attention paid to examine how the various models and methods handle these two issues. We present both popular parametric models and nonparametric matching methods.

The parametric models we review include the random effects (RE) model, the fixed effects (FE) model, the random trend and slope (RTS) model, and dynamic panel models which further include a moving average (MA) model (i.e., the error terms are modeled as following a random walk) and an autoregressive (AR) model (i.e., the lagged dependent variable is included as a predictor). The basic motivation of these models is to specify a function characterizing how an outcome is produced by the treatment and covariates. As such, the success of this approach critically depends on how well the specified function approximates the real data generation process. If the function is misspecified (e.g., with wrong functional forms or critical variables omitted from the model), then the estimated coefficients may not well approximate the causal effects of interest.

To avoid such mistakes, we introduce a nonparametric approach based on matching. The basic idea of matching is to compare observations with identical (in practice, often as similar as possible) covariate values except for their treatment status. In this case, any difference in the outcome between the matched observations can be attributed to treatment effects. Since there are no models required to characterize the outcome generation process, this approach is less subject to model specification errors. Also, by matching observations with similar covariates, this approach balances distributions of covariates across treatment groups and so allows for a more focused inference of the treatment effects. Given these advantages, matching has been extensively used for causal inference. But it has rarely been used in panel

data analysis (but see Smith and Todd [2005] and Nguyen [2012]). In this article, we show how to augment matching with difference-in-difference (DID) for analyzing panel data. We outline the conditions for this strategy to work effectively and provide solutions to combining matching estimates from multiple waves.

We apply the above models and methods to studying race-of-interviewer effects (ROIE) in the 2006 panel data of the General Social Survey (GSS). The causal question is whether a respondent would provide different answers (e.g., more favorable to blacks) if interviewed by a black interviewer versus by a nonblack interviewer. In the data we analyzed, the majority of the interviews were conducted by nonblack (74 percent), and so the nonblack interviewers are treated as the base group. We find that ROIE mostly concentrate on race-related survey items, and for two of the ten selected outcomes we examined, ROIE also vary by respondent's race. We do not find evidence that ROIE vary by interview mode.

This article proceeds as follows. In the second section, we review the parametric models and introduce the matching methods. In the third section, we review past research on ROIE. In the fourth section, we introduce the data and our analytical strategies. In the fifth section, we present the results. Finally, we summarize and discuss the implications of our study.

Models and Methods

RE and FE Models

In the absence of treatment, the potential outcome for unit i at time t can be written as

$$Y_{it}^0 = \lambda_t + X_{it}\gamma + C_i + e_{it},$$

where λ_t denotes a common time trend, γ the effects of covariates X_{it}, C_i an unobserved individual constant effect, and e_{it} a random error term. The inclusion of C_i allows one to account for the effects of unobserved time-invariant factors. Assuming an additive constant treatment effect δ, we can write the potential outcome for subject i at time t under treatment as

$$Y_{it}^1 = Y_{it}^0 + \delta.$$

This implies that, for the observed outcomes,

$$Y_{it} = \lambda_t + \delta D_{it} + X_{it}\gamma + C_i + e_{it}, \quad (1)$$

where D_{it} is the treatment status for subject i at time t. Two basic assumptions are needed in using this model. One is that the model includes all relevant variables, or for the purpose of identifying treatment effects, the model does not omit any variables that are associated with both the treatment and the outcome. The other assumption is that the functional form of the model is correctly specified. In the case shown in equation (1), the outcome is linearly dependent on other variables.

There are two major approaches to modeling the unobserved time-invariant effects C_i. First, C_i may be viewed as being generated from a random distribution with a mean zero and a variance σ_c^2. In other words, this approach assumes that C_i is uncorrelated with the covariates and the error terms and that the intercepts are random across subjects. This is so-called RE model. After fitting the model, we can test for random effects by examining whether

$$\hat{\sigma}_c^2 = 0. \tag{T1}$$

If $\hat{\sigma}_c^2$ significantly differs from zero, there is evidence for RE and the RE model is preferred to pooled ordinary least squares (POLS), as it is more efficient.

To estimate the variance–covariance matrix of the coefficients, we need to impose some structures on the distribution of the random error terms e_{it}. A reasonable assumption is that the error terms are correlated within units (i.e., serially correlated) while not correlated across units, and their variances vary across units (i.e., heterogeneity). Hence, in practice it is recommended to use standard errors clustered at the unit level to account for both serial correlation and heterogeneity in the error terms. We can test whether there is serial correlation, by testing whether $\rho_1 = 0$ in the regression

$$\hat{e}_{it} = \rho_1 \hat{e}_{it-1} + \upsilon_{it}, \text{for } t \geq 3, \upsilon_{it} \sim \text{i.i.d}, \tag{T2}$$

where $\hat{e}_{it}$ is the residual of the fitted model and υ_{it}, a random error with a mean zero.

The second approach to dealing with the unobserved time-invariant effect C_i is to remove it by differencing. One method is mean differencing, that is, subtracting the mean values of all variables in equation (1) from equation (1).

$$\begin{aligned} Y_{it} - \bar{Y}_i &= \delta(D_{it} - \bar{D})_i + (X_{it} - \bar{X}_i)\gamma + (e_{it} - \bar{e}_i), \\ \Lambda Y_{it} &= \delta \Lambda D_{it} + \Lambda X_{it}\gamma + \Lambda e_{it}. \end{aligned} \tag{2}$$

This is the so-called FE model (Wooldridge 2001:267).[1] Another method is first differencing, that is, subtracting the previous values of all variables in

equation (1) from equation (1) for each unit. This is often called the first difference (FD) method or sometimes called the change score analysis (Morgan and Winship 2007:253).[2]

$$\Delta Y_{it} = \Delta\lambda_t + \delta\Delta D_{it} + \Delta X_{it}\gamma + \Delta e_{it}. \quad (3)$$

Both the FE and the FD models are consistent estimators given the strict exogeneity assumption, that is, conditional on the unobserved time-invariant effects, the covariates are uncorrelated with the error terms. In general, the FE model produces more efficient estimates if the error terms are not serially correlated, while the FD model is more efficient if the error terms are serially correlated (Wooldridge 2001:284). In practice, their estimates are often statistically indistinguishable from one another. But, if the FE and the FD estimates indeed differ significantly, it may be a sign that the strict exogeneity assumption does not hold. We can test serial correlation in the error terms of the FE model as follows (Wooldridge 2001:275):

$$\Lambda\hat{e}_{it} = \rho_2\Lambda\hat{e}_{it-1} + \upsilon_{it}, \text{for } t \geq 3, \upsilon_{it} \sim \text{i.i.d}, \quad (T3)$$

where $\Lambda\hat{e}_{it}$ is the residual of the fitted FE model. Under the null hypothesis that the original error terms are uncorrelated, the differenced error terms are correlated with a correlation coefficient $\rho_2 = -1/(T-1)$, which equals to -0.5 when there are three time periods (i.e., $T = 3$).[3] Thus, if the estimated coefficient for ρ_2 is statistically significantly different from its expected value, then there is evidence for serial correlation. We can test serial correlation in the error terms of the FD model similarly (Wooldridge 2001:283):

$$\Delta\hat{e}_{it} = \rho_3\Delta\hat{e}_{it-1} + \upsilon_{it}, \text{for } t \geq 3, \upsilon_{it} \sim \text{i.i.d}. \quad (T4)$$

Under the null hypothesis of no serial correlation, $\rho_3 = -0.5$.

We can also test the strict exogeneity assumption that $\text{Cov}(X_{is}, e_{it}|C_i) = 0$, for any s and t (Wooldridge 2001:285). For the FE model, it suffices to test that no future covariates are predictive of current outcome, that is, $\beta_1 = 0$ and $\beta_2 = 0$ in the regression:

$$Y_{it} = \lambda_t + \delta D_{it} + X_{it}\gamma + \beta_1 D_{it+1} + X_{it+1}\beta_2 + C_i + e_{it}. \quad (T5)$$

For the FD model, it amounts to testing β_3, $\beta_4 = 0$ in the regression:

$$\Delta Y_{it} = \Delta\lambda_t + \delta\Delta D_{it} + \Delta X_{it}\gamma + \beta_3 D_{it} + X_{it}\beta_4 + \Delta e_{it}. \quad (T6)$$

When the strict exogeneity assumption is violated, neither the FE model nor the FD model is consistent.

The FE model is consistent while the RE model is more efficient. A Hausman test may be conducted to compare the estimates from the RE and the FE models. If the interest is in comparing the estimated coefficients for one variable as it is in our case, a single parameter Hausman test is more appropriate. The test statistic asymptotically follows a standard normal distribution (Wooldridge 2001:290):

$$\frac{\hat{\delta}_{FE} - \hat{\delta}_{RE}}{[se(\hat{\delta}_{FE})^2 - se(\hat{\delta}_{RE})^2]^{1/2}}. \tag{T7}$$

The RTS Model

The RTS model allows the effects of time (e.g., survey wave) and/or the treatment effects to vary across respondents. Below is a simple RTS model that includes both a random time trend and a random slope for treatment D:

$$Y_{it} = \lambda_t + g_i t + \delta D_{it} + h_i D_{it} + X_{it}\gamma + C_i + e_{it}. \tag{4}$$

If C_i is assumed to be uncorrelated with the error terms and the covariates, the RTS model is essentially an enhanced version of the RE model.[4] For each subject, besides a random intercept due to C_i, there is a separate time effect assumed to follow a normal distribution $N(\lambda_t, \sigma_g^2)$, where σ_g^2 is the variance of g_i, and a separate treatment effect assumed to follow a normal distribution $N(\delta, \sigma_h^2)$, where σ_h^2 is the variance of h_i. We can test whether the time effect and the treatment effect vary across subjects by testing whether their variances are equal to zero, namely,

$$\sigma_g^2 = 0, \tag{T8}$$

$$\sigma_h^2 = 0. \tag{T9}$$

Dynamic Panel Models

Besides using robust standard errors, another way to address the serial correlation problem in the error terms is to model the error terms as an MA process. The following is a dynamic model with a first order MA process:

$$Y_{it} = \lambda_t + \delta D_{it} + X_{it}\gamma + C_i + e_{it}. \tag{5}$$

We can test whether there is autocorrelation in the error terms by testing whether ρ_4 is different from zero.

$$e_{it} = \rho_4 e_{it-1} + \upsilon_{it}, \upsilon_{it} \sim \text{i.i.d.} \tag{T10}$$

The current outcome may also depend on past outcomes (e.g., subjects tend to be consistent in their responses over time) or past treatment (e.g., due to learning as triggered by the past treatment (Pickery, Loosveldt, and Carton 2001)). We can specify an AR model to approximate such path dependence. The following model includes the immediate past outcome (i.e., an AR(1) process) and the immediate past treatment in the explanatory variables.

$$Y_{it} = \lambda_t + \theta Y_{it-1} + \delta D_{it} + \delta_1 D_{it-1} + X_{it}\gamma + C_i + e_{it}. \quad (6)$$

To estimate model (6), we assume all covariates except Y_{it-1} are strictly independent of the error terms at all times. After first differencing to remove C_i, equation (6) becomes

$$\Delta Y_{it} = \Delta\lambda_t + \theta\Delta Y_{it-1} + \delta\Delta D_{it} + \delta_1\Delta D_{it-1} + \Delta X_{it}\gamma + \Delta e_{it}. \quad (7)$$

A problem is that ΔY_{it-1} may still correlate with Δe_{it} because Y_{it-1} is correlated with e_{it-1}. Thus in general, we need to use instrumental variable (IV) methods in order to consistently estimate the model (Arellano and Bond 1991; Wooldridge 2001:303). For example, for panel data with three time periods, equation (7) at the third time period is given by:

$$\Delta Y_{i3} = \Delta\lambda_3 + \theta\Delta Y_{i2} + \delta\Delta D_{i3} + \delta_1\Delta D_{i2} + \Delta X_{i3}\gamma + \Delta e_{i3}. \quad (8)$$

We can use Y_{i1} as an IV for ΔY_{i2}. We can also use the first difference of the covariates (i.e., $\Delta X_{i2} = X_{i2} - X_{i1}$) as additional IVs for ΔY_{i2}, assuming they are strictly exogenous. Since the IVs are constructed from past time periods, this approach works only when there are at least three time periods of data. When there are only three time periods of data, the estimation is essentially conducted only on the third time period. Including both past outcome and treatment makes model (6) more robust to model misspecification errors than previous models. But the increased robustness comes at a price. The use of IV in general leads to larger standard errors for the estimates.

We can test whether the immediate past outcome and treatment have any effect on predicting the outcome by testing,

$$\theta = 0, \quad (T11)$$
$$\delta_1 = 0. \quad (T12)$$

We may also like to specify a model that can incorporate both MA and AR processes. But for that purpose, we need at least four time periods to find appropriate IVs and to be able to test serial correlation in the MA process. Since we are studying methods that are useful for analyzing three waves of panel data in the GSS, we skip the details of this model.

Matching Methods

Matching has been a popular method for drawing causal inference in cross-sectional data but has rarely been used in analyzing panel data. Below we first introduce the potential outcome framework and then explore ways to extend matching for panel data analysis.

Suppose we are interested in estimating the effects of a binary additive treatment D on an outcome Y. Define Y_i^1 as the potential outcome for unit i if it receives the treatment while Y_i^0 as its potential outcome if it does not receive the treatment. Define the average treatment effects (ATEs) as $\delta = E[Y^1 - Y^0]$, the ATE for the treated (ATT) as $\delta_1 = E[Y^1 - Y^0|D = 1]$, and the ATE for the controls (ATC) as $\delta_0 = E[Y^1 - Y^0|D = 0]$. If we know both of the potential outcomes, the individual treatment effect (ITE) would be $\delta_i = (Y_i^1 - Y_i^0)$. Averaging the ITE across units would provide a natural estimate of ATE. Similarly, averaging the ITE across the units that receive the treatment and those do not receive the treatment would provide natural estimates for the ATT and ATC, respectively. The problem is that we usually can only observe one of the potential outcomes for each unit. If a unit is in the treated group, it cannot be in the control group at the same time and vice versa. The basic idea of matching is to match units with exact (or as similar as possible) covariates in the opposite treatment group to impute the missing potential outcomes.

For matching to work properly it needs two assumptions:

1. Conditional ignoribility: $Y^1, Y^0 \perp D|X$.
2. Common support: $0 < \Pr(D = 1|X = x) < 1$.

The first condition states that units with the same covariates are equally likely to receive the treatment. This condition ensures that units with the same covariates can serve as the counterfactuals for one another even if they end up in different treatment groups. The second condition says units with the same covariates cannot be exclusively in just one of the treatment groups: Some must be in the treated group while others must be in the control group. This condition ensures that at any part of the covariate distribution, there are units in the opposite treatment group that can be used as counterfactuals. Exact matching is often impossible, especially when there are multiple covariates and covariates are continuous. One alternative is to match units with the most similar covariate values, that is, the so-called nearest-neighbor matching.

The key of matching is to measure similarity between units. The distance between two units is often measured according to

$$d(X_i, X_j) = \sqrt{(X_i - X_j)^T W^{-1} (X_i - X_j)},$$

where X_i and X_j represent the covariates for unit i and j, respectively, and W, a weight matrix. There are two popular options for the weight matrix: (1) the variance–covariance matrix of X and (2) the sample variances of X. One potential problem with the first choice is that the variance–covariance matrix of X may not be invertible in cases where there is multicollinearity in the covariates. In contrast, the second choice is more stable, but it ignores the covariances between the covariates.[5] Once the distance between units is measured, matching amounts to finding a number of closest neighbors for a unit from the opposite treatment group and use the neighbors' average outcome to impute the unit's missing potential outcome. Once matching is completed, ATE, ATT, and ATC can be estimated as follows (Abadie et al. 2004; Abadie and Imbens 2006).[6]

- ATE: $\hat{\delta} = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i^1 - \hat{Y}_i^0)$, where $\hat{Y}_i^1 = Y_i^1, \hat{Y}_i^0 = \bar{Y}_{mi}^0$ if $D_i = 1$ and $\hat{Y}_i^1 = \bar{Y}_{mi}^1, \hat{Y}_i^0 = Y_i^0$ if $D_i = 0$, $\bar{Y}_{mi}^0$ is the average outcome of m nearest neighbors in the control group for treated unit i, and $\bar{Y}_{mi}^1$ is the average outcome of m nearest neighbors in the treated group for control unit i. Writing ATE as a function of observed outcomes, $\hat{\delta} = \frac{1}{N}\sum_{i=1}^{N}(2D_i - 1)(1 + K_i)Y_i$, where K_i is the number of times unit i serves as a match to other units.
- ATT: $\hat{\delta}_1 = \frac{1}{N_1}\sum_{i\in\{D=1\}}(Y_i^1 - \hat{Y}_i^0)$, where N_1 is the number of treated units, and $\{\mathrm{D} = 1\}$ is the set of treated units.
- ATC: $\hat{\delta}_0 = \frac{1}{N_0}\sum_{i\in\{D=0\}}(\hat{Y}_i^1 - Y_i^0)$, where N_0 is the number of control units, and $\{\mathrm{D} = 0\}$ is the set of control units.

Abadie et al. (2004) and Abadie and Imbens (2006) show the variances of the sample ATE, ATT, and ATC are

$$\sigma^2_{\hat{\delta}} = \frac{1}{N^2}\sum_{i=1}^{N}\{1 + K_i\}^2\sigma^2(Y_i|X_i, D_i), \tag{9}$$

$$\sigma^2_{\hat{\delta}_1} = \frac{1}{N_1^2}\sum_{i\in\{D=1\}}\sigma^2(Y_i|X_i, D_i), \tag{10}$$

$$\sigma^2_{\hat{\delta}_0} = \frac{1}{N_0^2}\sum_{i\in\{D=0\}}\sigma^2(Y_i|X_i, D_i). \tag{11}$$

The matching methods are available in ''nnmatch'' (Abadie et al. 2004) in Stata or ''Matching'' (Sekhon 2011) in R.

In the following text, we explore how to combine matching with difference-in-difference for panel data analysis. DID assumes that in the absence of treatment, the original difference between treated units and control units in the outcome will remain constant over time, that is, $E[Y_{t-1}^0(1) - Y_{t-1}^0(0)] = E[Y_t^0(1) - Y_t^0(0)]$, which implies $E[Y_t^0(1) - Y_{t-1}^0(1)] = E[Y_t^0(0) - Y_{t-1}^0(0)]$, or say, $\Delta Y_t^0 \perp D_t$. With this assumption, we can estimate the ATT as $\Delta Y_t(1) - \Delta Y_t(0)$. A short proof is given as follows.

$$\begin{aligned} E[\Delta Y_t(1) - \Delta Y_t(0)] &= E\Big[Y_t(1) - Y_{t-1}(1) - \big(Y_t(0) - Y_{t-1}(0)\big)\Big] \\ &= E\Big[Y_t^1(1) - Y_{t-1}^0(1) - \big(Y_t^0(0) - Y_{t-1}^0(0)\big)\Big] \\ &= E\Big[Y_t^1(1) - Y_{t-1}^0(1) - \big(Y_t^0(1) - Y_{t-1}^0(1)\big)\Big] \\ &= E[Y_t^1(1) - Y_t^0(1)] \\ &= ATT. \end{aligned}$$

Similarly, we can show that with a stronger assumption like

$$\Delta Y_t^1, \Delta Y_t^0 \perp D_t,$$

ATE can be identified. However, DID does not account for the effects of covariates. We propose the below ignorability conditions for this purpose:

1. Strong form of ignorability

$$\Delta Y_t^1, \Delta Y_t^0 \perp D_t | \overrightarrow{X_t}, \overrightarrow{D_{t-1}}. \tag{12}$$

2. Weak form of ignorability:

$$\Delta Y_t^1, \Delta Y_t^0 \perp D_t | \overrightarrow{X_{t,s}}, \overrightarrow{D_{t-1,s}}. \tag{13}$$

The strong form of ignorability states that conditioning on the current and past covariates ($\overrightarrow{X_t}$) and the past treatment history ($\overrightarrow{D_{t-1}}$) at time t, the differenced potential outcomes are interexchangeable between the treated and control units. The weak form of ignorability allows the conditioning to be on more recent covariate and treatment history. For covariates, conditioning on their values in between time t and time s is sufficient while for treatment, conditioning on their values in between time $t-1$ and time s is sufficient. In one of its simplest forms, the ignorability will hold if only current and immediate past covariates and immediate past treatment are conditioned on.

In short, what we propose is to difference the outcome first and then apply the matching method to estimate treatment effects at each wave. Using the differenced outcome helps remove the effects of time-invariant factors while matching helps balance covariates and create a more focused causal inference. For example, at wave 2, we can match units with the same (or as similar as possible) covariates at both wave 2 and wave 1 and that have the same treatment status at wave 1 but different treatment status at wave 2. In other words, units with treatment sequence (0, 0) and (0, 1) are matched while those with treatment sequence (1, 0) and (1, 1) are matched. Similarly, we can conduct matching and estimate treatment effects at later waves. Here is where the weak form of ignorability is a more practical assumption. By conditioning on a shorter span of covariates and treatment history, it reduces the dimensions of matching and so makes it easier to find comparable matches.

We can aggregate the estimated treatment effects from multiple waves into a single set. A simple way to accomplish this is to weight the estimates from multiple waves by their relative sample size. The aggregate ATE is

$$\delta_W = \sum_{t=2}^{T} \frac{N_t}{N} \hat{\delta}_t,$$

where t is the index for panel wave, T is the total number of waves, N_t is the number of subjects at wave t, N is the total number of observations in all waves, and $\hat{\delta}_t$ is the estimate of ATE at wave t. The variance for the weighted ATE may be estimated as

$$\hat{\sigma}^2_{\delta_W} = \sum_{t=2}^{T} \frac{N_t^2}{N^2} \hat{\sigma}^2_{\hat{\delta}_t} + 2 \sum_{2 \le g < h \le T} \frac{N_g N_h}{N^2} \mathrm{Cov}(\hat{\delta}_g, \hat{\delta}_h),$$

where $\hat{\sigma}^2_{\hat{\delta}_t}$ is the estimated variance for treatment effects at wave t and $\mathrm{Cov}(\hat{\delta}_g, \hat{\delta}_h)$ is the covariance of treatment effects across waves g and h.[7] Aggregate measures for ATT and ATC can be derived accordingly.

To date, only a few studies have explored ways to combine matching with DID. Smith and Todd (2005) and Berger et al. (2009) showed how this could be done in two-wave panel data with the treatment conducted in between. Nguyen (2012) studied the issue when baseline data are not available, but two periods of data after the intervention are available. In contrast, our method is more general, allowing repeated treatment and outcomes across multiple ($\geq$2) waves. This generalization is not trivial, not only because panel data with repeated outcomes and treatment are typical, but also because there is little guidance on how to condition on prior covariates and treatment

in this setting and how to combine the estimates from multiple waves. Another difference is that we use covariate matching while the previous studies like Smith and Todd (2005), Berger et al. (2009), and Nguyen (2012) used propensity score matching. Propensity score matching relies on a well specified model for estimating the propensity scores and estimating the propensity scores often creates additional uncertainties for the estimation of treatment effects. In addition, for each different contrast of treatment sequences, it requires a new propensity score model. Therefore, we think covariate matching may be more intuitive and robust to use.

Model Choice

We have presented seven models or methods for estimating causal effects in panel data. Broadly speaking, they can be divided into two groups. One can be called the fixed effects estimators, including the FE model, the FD model, the AR model, and the matching method. These estimators allow for arbitrary correlations between unobserved time-invariant individual effects and treatment and so are generally more robust. The second group can be called the random effects estimators, which include the RE model, the RTS model, and the MA model. These models assume no correlations between the unobserved time-invariant individual effects and treatment. As such, they may be subject to bias due to the existence of such correlations. However, these estimators also tend to be more efficient. So given consistency, the RE estimators are preferred to the FE ones. In practice, it is often a good idea to present results from both types of estimators. If the results differ significantly, then those from the FE estimators should be given more weight in judging the evidence.

Even among the FE estimators, there are great or subtle differences. In general, there is not much difference between the FE model and the FD model if the strict exogeneity assumption holds. But if their results differ significantly, it signals that the strict exogeneity assumption might be violated. The AR model includes the lagged outcome and so is a fuller specified model than others. One problem with it is the estimates tend to have large standard errors because of the use of IVs. To choose between the AR model and the FE/FD models, we can first check whether the coefficient for the lagged outcome is statistically significant or not. If not, we may proceed with the FE and the FD models. If yes, the AR model may be preferred.

Matching is nonparametric and so is more robust to model specification errors. By balancing covariates across treatment groups, matching also provides a more focused estimate of treatment effects. It also allows treatment effects to vary across subpopulations (the treated vs. the untreated). At the

same time, matching may be sensitive to the metric used to measure similarities between units and to the number of matches requested. Also, its external validity may be low. By contrast, the parametric models tend to be more efficient, more flexible, and easier to generalize the results to the population. Because of the distinct advantages offered by the parametric models and the matching method, we suggest using both approaches for causal inference in order to cross-validate the evidence.

Previous Research on ROIE

Below we demonstrate how to use the abovementioned methods to study race of interviewer effects in the General Social Survey. Generally speaking, interviewer effects refer to the extent to which respondents' responses are influenced by interviewers' characteristics, such as race, gender, age, experience, attitude, behavior, workload, and so on. In this study, we are concerned that ROIE may introduce bias in studies of social policy preferences, in comparing social attitudes between racial groups, and so on.

Many prior studies have documented interviewer effects in survey reports in various contexts, from health behaviors (Davis et al. 2010), to network size (Marsden 2003; Paik and Sanchagrin 2013; Tilburg 1998), and to political preference (Finkel, Guterbock, and Borg 1991), from survey participation (Durrant et al. 2010; Pickery et al. 2001; Singer, Frankel, and Glassman 1983) to the effects of interviewer's gender (Kane and Macaulay 1993), and so on. Previous studies have also examined ROIE specifically. Anderson, Silver, and Abramson (1988) showed that the reported racial attitudes varied by the interviewer's race. Schaeffer (1980) examined the GSS data from 1972 to 1977 and found that there were ROIE on both racial and nonracial items. Davis (1997) showed that black respondents were more likely to conceal their true political beliefs when interviewed by a white interviewer. Cotter, Cohen, and Coulter (1982) found that even in phone interviews, there were significant interviewer effects. ROIE have also been shown to vary by race. Schaeffer (1980) showed ROIE appeared to be smaller for white respondents than for black respondents while Cotter et al. (1982) showed the opposite. Past research (Davis 1997; Davis and Silver 2003; Krysan and Couper 2003) has pointed out that ROIE can originate from social ingratiation or social anxiety induced by racial stereotype threat.[8] Social ingratiation is found to be more responsible for ROIE on attitudinal and opinion questions while social anxiety is more important for ROIE on factual questions (Davis and Silver 2003). In this article, the

survey questions we selected to study are mostly attitudinal questions. We conjecture that the primary cause for ROIE will be due to social ingratiation.

Our study of ROIE differs from previous ones in several notable ways. First, using panel data, we are able to better handle the effects of unobserved time-invariant factors and so provide more rigorous evidence on ROIE. In contrast, previous studies mostly rely on multilevel models (e.g., Hox 1994; Marsden 2003; Paik and Sanchagrin 2013; Pickery et al. 2001), which are susceptible to the influence of unobserved time-invariant factors. To date, only Pickery and Loosveldt (2000) and Pickery et al. (2001) have studied interviewer effects in panel data. They used a multilevel cross-classified model to characterize the interviewer effects. But if there are unobserved factors that affect both interviewer assignment and the outcome, then the interviewer effects thus estimated may be biased. This is a standard omitted variable bias problem.[9] Davis (1997) utilized difference in the race of the interviewer in two-wave panel data to identify ROIE. But the evidence was preliminary, as a simple linear regression model was used to fit the data with few covariates, and no means were taken to address potential bias due to unobserved factors. Second, different from previous studies that use only one kind of model to collect evidence, we use a series of different models and methods to triangulate the evidence for ROIE. Finally, we provide a more comprehensive examination of ROIE by studying their variations by interview mode and respondent's race.

Data and Analytical Strategies

Data

We used the 2006 to 2010 panel data of the GSS, which includes three waves of interviews. The 2006 sample includes 2,000 subjects. Of these, 1,536 and 1,276 are reinterviewed in 2008 and in 2010, respectively. The data are the first panel data of the GSS, which since 1972 has been cross-sectional.

We used responses to ten selected attitudinal and opinion questions as outcomes of interest. The outcomes can roughly be divided into four groups according to their content. The first group includes one survey item. Specifically, respondents were asked to report their views on the level of intelligence of blacks and whites based on a scale from 1 (unintelligent) to 7 (intelligent). We use the difference between the level of intelligence for blacks and for whites (i.e., black intelligence − white intelligence) separately reported by each respondent to construct a measure of "perceived intelligence gap." This measure ranges from −6 (blacks are less intelligent than whites)

to 6 (blacks are more intelligent than whites). The second group includes four items, including respondents' reports of their confidence in the executive branch of the federal government, in the Congress, in the Supreme Court, and in banks and financial institutions, respectively. The response can range from 1 (a great deal of confidence) to 3 (hardly any confidence at all). The third group has two items: respondents' views on the level of current social spending on welfare (denoted as "spending on welfare'') and their views on improving the conditions of blacks (denoted as "spending on blacks''). The responses originally range from 1 (too little) to 3 (too much). To be consistent with other measures, we reversely recoded the responses (i.e., too much = 1, too little = 3). The last group includes three items on respondents' views on the appropriate role of the federal government in helping blacks (denoted as "should help blacks''), in helping the poor (denoted as "should help poor''), and in helping the sick (denoted as "should help sick''). The response can range from 1 (strongly agree it is the responsibility of the government to help) to 5 (strongly agree that people should take care of themselves). To be consistent with other outcome measures, we also reversely coded the responses so that high scores correspond to that the government should take more responsibility.

Presumably, the selected survey questions are more likely to induce ROIE than other ones because of respondents' concerns with social desirability. That said, we expect that there will be some variation in their sensitivity to ROIE, as some of the questions (e.g., perceived intelligence gap, "spending on blacks'', and "should help blacks'') appear to be more racially charged than others. Also, not all outcome measures are available in all three ballots of the survey. The questions on social spending are available in all ballots. But perceived intelligence gap is unavailable in ballot C and the questions on public confidence and the role of the government in helping the disadvantaged are unavailable in ballot A.[10]

The main explanatory variable is interviewer's race. It is originally coded in the survey as a categorical variable, including Whites, Blacks, Hispanics, Asians, and others with two or more races. Considering that responses to the survey items we selected are presumably more likely to be affected by whether the interviewer is black or not, we recoded the variable to a binary one (nonblacks = 0, blacks = 1).[11] In other words, the treatment is whether a respondent was interviewed by a black interviewer. We also control for interviewer's other characteristics such as age, sex (binary variable, female = 0; male = 1), and years of working for the GSS.

Other control variables include respondent's age, sex (binary variable, female = 0; male = 1), race (categorical variable, whites = 1; blacks = 2;

others = 3), employment status (binary variable, unemployed = 0; employed = 1), family income (unit, US$1,000), number of children, marital status (binary variable, unmarried = 0; married = 1), years of education, party identification (categorical variable, strong republican = 0; strong democrat = 1; others = 2), religious denomination (categorical variable, no religion = 0; Protestant = 1, Catholics = 2; other religions = 3), type of place living in (categorical variable, large city = 0; suburb = 1; small city or town = 2; others = 3), residential region (categorical variable, South = 0; Northeast = 1; Midwest = 2; West = 3), and interview mode (binary variable, in person = 1; by phone = 0).

Our main hypothesis (Hypothesis 1) is that respondents will provide more favorable responses toward blacks when interviewed by a black interviewer than by a nonblack interviewer. This expectation is based on prior findings (e.g., the social desirability argument by Krysan and Couper 2003).[12] We also expect ROIE to be larger if the interviews are conducted in person than by phone (Hypothesis 2). This is both because face-to-face interviews provide respondents with more cues to infer interviewer's race than phone interviews, and because phone interviews are generally more standardized and supervised more closely (Schaeffer, Dykema, and Maynard 2010). Finally, we expect that nonblack respondents are more subject to ROIE than black respondents (Hypothesis 3). This is because deviant opinions from outside group members may be socially more offensive than those from inside group members, especially given the racial history of the United States.[13]

Analytical Strategies

For simplicity, we assume the outcomes are continuous variables. A few comments on our analytical strategies are in order. First, based on our understanding, the interviewers in the GSS are assigned based on the racial composition and languages spoken in a neighborhood (i.e., the primary sampling unit). Respondents from neighborhoods with a significant portion of minorities are more likely to have corresponding minority interviewers. Hence, the treatment assignment is correlated with some of the neighborhood characteristics while the latter is likely to correlate with the outcomes as well. We have controlled for some of the neighborhood characteristics in our models by including the type of place that respondents live in and the areas that they are from. But the inclusion of these factors may not fully offset the effects of other unobserved neighborhood factors. As a result, the fixed effects estimators may be more appropriate for our study. But for illustrative and

comparative purposes, we also used RE estimators. Second, in the FE and the FD models, we assume the covariates X are strictly exogenous and only carry out the exogeneity tests on the treatment indicator D. Third, in the AR(1) model, we use the outcome at the first wave Y_{i1}, the differenced covariates ΔX_{it-1}, and the differenced treatment ΔD_{it} and ΔD_{it-1} as IVs for ΔY_{it-1}, assuming they are strictly exogenous.

Fourth, we conduct a series of tests to evaluate the model specifications, including a test for the absence of random effects (T1), tests for serial correlations in the RE model (T2), in the FE model (T3), in the FD model (T4), tests for strict exogeneity in the FE model (T5), and in the FD model (T6), a single parameter Hausman test comparing the estimated treatment effects in the RE and the FE models (T7), a test for individual random trend (T8), a test for individual random slope (T9), a test for serial correlations in the MA(1) model (T10), and tests for the effects of the lagged dependent variable (T11) and the lagged treatment variable (T12) in the AR(1) model.

Fifth, we create an interaction term between interview mode and interviewer's race (black vs. nonblack), which equals to 1 if an interview was conducted in person by a black interviewer and equals 0 otherwise. We include this interaction term (along with the corresponding main effects) in one set of models and use the estimated coefficients for the interaction term to whether ROIE vary by interview mode. Sixth, we also create an interaction term between interviewer's race (black vs. nonblack) and respondent's race (black vs. nonblack), which equals to 1 if a nonblack respondent was interviewed by a black and equals to 0 otherwise. We include this interaction term (along with the corresponding main effects) in another set of models and use the estimated coefficients for this interaction term to measure whether nonblack respondents are more subject to ROIE than black respondents. In the Online Appendix, we also show a separate set of regressions in which we only include respondents who are white or black. There we create an interaction term between interviewer's race and respondent's race that equals to 1 if a white respondent was interviewed by a black and otherwise 0. The estimated coefficients for this interaction term measure whether white respondents are more subject to ROIE than black respondents. To note, in all three cases, the estimated coefficients for the interaction terms tend to have large standard errors, because of the small variations in the interaction terms.

Seventh, to adjust for possible serial correlations in the error terms, whenever applicable we use standard errors clustered at individuals. When reporting the results, we use asterisks to show the significance patterns of the two-tailed p values at three levels: .05 (*), .01 (**), and .001 (***).[14]

Finally, we provide matching estimates. Because we used differenced outcomes, matching was effectively conducted only on wave 2 and wave 3, respectively. We match on current and immediate past covariates and immediate past treatment. We use the inverse sample variances of the matching variables to measure distance between units. We also request exact matching on respondent's sex, race, party identification, and immediate past treatment in order to account for their strategic influence on the outcomes. We use robust standard errors that allow the variances of the outcomes to vary by treatment and covariates. We aggregate the results from the two waves by weighting them according to their relative sample sizes. We present the aggregate results in the main text while showing the separate results for each wave in the Online Appendix.[15] The matching is conducted by using the "nnmatch" package in Stata (Abadie et al. 2004).

Results

Summary Statistics

Among the 4,812 interviews in the data, 684 (including 245 in wave 1, 190 in wave 2, and 189 in wave 3) were conducted by black interviewers, which accounts for about 13 percent of the total number of interviews. The proportion of interviews conducted by blacks has increased over time, from 12.25 percent in 2006 to 12.37 percent in 2008, and to 14.81 percent in 2010. Among all the interviews (except one because of missing value in interview mode), 3,541 (about 74 percent) were conducted in person by nonblacks, 551 (about 11 percent) in person by blacks, 646 (about 13 percent) by phone by nonblacks, and 73 (about 2 percent) by phone by blacks. There were 168 unique interviewer IDs in wave 1, 150 in wave 2, and 183 in wave 3. Interviewer ID does not uniquely identify an interviewer across waves. Hence, the actual number of interviewers who conducted the surveys may be smaller than the total number of unique interviewer IDs across waves, that is, 501 (= 168+150+183). Respondents also experienced some degree of change in the race of their assigned interviewers. Among the 2,812 possible changes in interviewer's race across waves, in 223 times it was from a black interviewer to a nonblack interviewer; in 257 times it was the reverse; and a majority of time, there was no change in interviewer's race.

We present the summary statistics of selected variables in Table 1. According to the *t*-test results, the mean responses of those interviewed by blacks and those interviewed by nonblacks are significantly different in 7 of the 10 outcomes at the 5 percent significance level. Specifically, compared

to the respondents interviewed by a nonblack interviewer, the respondents interviewed by a black interviewer on average perceive a wider gap between the intelligence of whites and blacks in favor of blacks, have higher confidence in the Supreme Court, are more likely to view the current social spending on welfare and on blacks are too little, and more likely to advocate for more governmental help for the blacks, the poor, and the sick.[16]

Among the independent variables (including both respondent's and interviewer's characteristics), we also find the majority of the mean values (i.e., 24 of 30 variables) differ significantly across the two groups at the 5 percent significance level. The result of the Hotelling's test clearly rejects the hypothesis that the mean values of these independent variables are equal across the two groups ($p < .001$). As far as the significant differences are concerned, we see that compared to the respondents interviewed by a nonblack interviewer, the respondents interviewed by a black interviewer on average have younger, more female, and less experienced interviewers. The latter respondents are also more likely to be female, black, and born in the United States, have lower family income per capita, have more children, are less likely to be married, are more likely to be strong democrats but less likely to be strong republicans, are more likely to be Protestant but less likely to be Catholic, are more likely to live in big or small cities but less likely to live in suburbs, are more likely from the South and Midwest but less likely from the West, and are more likely to be interviewed in person. These differences suggest, at least, it is important to control for these variables in the statistical models.

Estimates of ROIE From the Parametric Models

Model specification tests. We first present the results of the model specification tests (shown in Table 2). Results for T1 indicate the hypothesis of no random effects are rejected for all the outcomes, which suggests that the RE models are preferred to the POLS estimator. Results for T2 indicate that the error terms in the RE models are possibly serially correlated, which suggests that robust standard errors should be used, as we have done.

Results for T3 indicate that the error terms are not serially correlated in the FE models, except for one outcome, which suggests that the FE models are probably preferred to the FD models. However, results for T4 suggest the opposite, where for 7 of the 10 outcomes, there is evidence that the error terms are serially correlated. As a safe measure, we use robust standard errors in both cases. Results for T5 and T6 indicate that assignment of interviewers by and large is orthogonal to the error terms in the FE and the FD models. We

Table 1. Summary Statistics of Outcomes and Selected Variables.

Variables	*N*	Mean	*SD*	Minimum	Maximum	Black	Nonblack	Diff.	*p*
Outcomes									
Perceived intelligence gap	3,065	−0.34	1.20	−6	6	0.07	−0.40	0.47	.00
Confidence in executive branch of federal government	3,139	2.29	0.67	1	3	2.28	2.29	0.00	.93
Confidence in Congress	3,142	2.30	0.63	1	3	2.33	2.30	0.03	.37
Confidence in Supreme Court	3,121	1.88	0.67	1	3	1.98	1.87	0.11	.00
Confidence in bank and financial institutions	3,163	2.04	0.67	1	3	2.09	2.03	0.06	.07
Spending on welfare	2,337	1.86	0.77	1	3	2.09	1.82	0.27	.00
Spending on blacks	2,179	2.22	0.67	1	3	2.67	2.16	0.51	.00
Should help blacks	3,092	2.49	1.26	1	5	3.21	2.38	0.82	.00
Should help poor	3,116	3.09	1.19	1	5	3.37	3.05	0.32	.00
Should help sick	3,119	3.55	1.24	1	5	3.77	3.52	0.26	.00
Interviewer characteristics									
Age	4,779	53.45	10.34	22	84	51.18	53.79	−2.61	.00
Male	4,654	0.18	0.38	0	1	0.08	0.19	−0.11	.00
Years of being an interviewer	4,779	3.25	4.02	0	26	2.71	3.33	−0.62	.00
Respondent characteristics									
Age	4,761	48.92	17.02	18	89	48.50	48.99	−0.49	.50
Male	4,812	0.42	0.49	0	1	0.35	0.43	−0.08	.00
Race									
White	4,812	0.76	0.42	0	1	0.54	0.80	−0.26	.00
Black	4,812	0.14	0.35	0	1	0.40	0.10	0.30	.00
Others	4,812	0.10	0.29	0	1	0.06	0.10	−0.04	.00
Born in the United States	4,809	0.88	0.32	0	1	0.92	0.88	0.04	.00
Employed	4,812	0.49	0.50	0	1	0.49	0.49	0.00	.99

(continued)

Table 1. (continued)

Variables	*N*	Mean	*SD*	Minimum	Maximum	Black	Nonblack	Diff.	*p*
Family income per capita (US$1,000)	4,272	24.25	24.11	0	178	21.23	24.69	−3.45	.00
Number of children	4,800	1.93	1.68	0	8	2.07	1.91	0.16	.03
Married	4,812	0.49	0.50	0	1	0.39	0.50	−0.11	.00
Years of education	4,802	13.62	3.05	0	20	13.49	13.64	−0.15	.25
Party identification									
Strong republican	4,812	0.12	0.32	0	1	0.08	0.12	−0.04	.00
Strong democrat	4,812	0.17	0.37	0	1	0.27	0.15	0.12	.00
Others in between	4812	0.72	0.45	0	1	0.64	0.73	−0.08	.00
Religion									
None	4,812	0.16	0.37	0	1	0.15	0.16	−0.02	.28
Protestant	4,812	0.51	0.50	0	1	0.62	0.50	0.13	.00
Catholic	4,812	0.24	0.43	0	1	0.14	0.26	−0.12	.00
Others	4,812	0.08	0.28	0	1	0.09	0.08	.01	.39
Types of places living in									
Large city	4,812	0.31	0.46	0	1	0.41	0.29	0.12	.00
Suburb	4,812	0.29	0.45	0	1	0.19	0.30	−0.11	.00
Small city or town	4,812	0.05	0.23	0	1	0.11	0.05	0.06	.00
Others	4,812	0.35	0.48	0	1	0.29	0.36	−0.07	.00
Region									
South	4,812	0.39	0.49	0	1	0.55	0.36	0.19	.00
Northeast	4,812	0.16	0.37	0	1	0.14	0.16	−0.02	.14
Midwest	4,812	0.23	0.42	0	1	0.26	0.22	0.04	.02
West	4,812	0.23	0.42	0	1	0.04	0.25	−0.21	.00
Interviewed in person	4,811	0.85	0.36	0	1	0.88	0.85	0.04	.01
Two-group Hotelling's test	$F(25,4073) = 24.39$								.00

Note: The black and nonblack columns show the mean values for the interviews conducted by black interviewers and nonblack interviewers. The diff. column shows the differences in the means between the two groups (i.e., the former minus the latter). The *p* column shows the *p* values from the *t*-test on such differences. The Hotelling's test examines whether the mean values of the variables are equal across the two groups. *SD* = standard deviation.

Table 2. Model Specification Tests.

Outcomes	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
Perceived intelligence gap	0	0	0.40	0.06	0.96	0.57	0.29	0.04	0.32	0	0.08	0.53
Confidence in executive branch of federal government	0	0	0.00	0.00	0.46	0.00	0.30	0.04	0.20	0	0.00	0.22
Confidence in Congress	0	0	0.09	0.00	0.46	0.10	0.71	0.97	0.29	0	0.18	0.92
Confidence in Supreme Court	0	0	0.59	0.00	0.78	0.01	0.14	0.05	0.00	0	0.05	0.17
Confidence in bank and financial institutions	0	0	0.21	0.03	0.49	0.84	0.00	0.00	0.09	0	0.06	0.14
Spending on welfare	0	0	0.93	0.01	0.21	0.93	0.95	0.70	0.00	0	0.46	0.70
Spending on blacks	0	0	0.72	0.01	0.74	0.93	0.34	NA	NA	0	0.06	0.68
Should help blacks	0	0	0.30	0.20	0.07	0.89	0.45	0.00	0.89	0	0.77	0.69
Should help poor	0	0	0.38	0.00	0.68	0.54	0.10	0.00	0.01	0	0.11	0.86
Should help sick	0	0	0.07	0.15	0.08	0.73	0.39	0.00	0.04	0	0.64	0.75

Note: p value is presented for each test.

T1 = test for the absence of random effects (RE) in the RE model; T2 = test for no serial correlation in the RE model; T3 = test for no serial correlation in the fixed effects (FE) model; T4 = test for no serial correlation in the first difference (FD) model; T5 = test for strict exogeneity in the FE model; T6 = test for strict exogeneity in the FD model; T7 = Hausman test for equality in estimated race-of-interviewer (ROIE) between RE and FE models; T8 = test for no random trend in the random trend and slope (RTS) model; T9 = test for no random slope in the RTS model; T10 = test for no serial correlation in moving average (MA)(1) process; T11 = p value for the effect of the lagged outcome in the autoregressive (AR)(1) model; T12 = p value for the effect of the lagged treatment in the AR(1) model.

report the estimates of both the FE models and the FD models in order to facilitate comparisons.

Results for T7 indicate that the estimated ROIE in the RE models and the FE models are not distinguishable for 9 of the 10 outcomes (i.e., except for confidence in bank and financial institutions). Hence, the bias (if any) in the estimates of the RE models appears not to be substantial.

Results for T8 indicate that there may be significant individual time trends for 6 of the 10 outcomes. Results for T9 suggest that there are probably significant variations in ROIE across respondents for 4 of the 10 outcomes.[17] Results for T10 show that the error terms in the MA(1) model may be serially correlated for all outcomes. Since the RTS models and the MA(1) models are RE estimators, we mention their results mainly for illustrative purpose.

Results for T11 show the lagged outcome is a significant predictor for only one outcome while the race of the last interviewer had no significant effect on current responses for all outcomes. These results suggest that the AR(1) models may not be more advantageous than the FE or FD models.

To summarize, given that the interviewer assignment process is probably determined by unobserved factors at neighborhood and/or individual levels that are correlated with the outcomes, the FE estimators are more robust than their RE counterparts. Among the FE estimators, it appears that the FE models are preferred to others for their relative efficiency. But it is worth noting that the estimates in the RE models are statistically indistinguishable from those in the FE models for most of the outcomes we studied.

Results of the estimated ROIE. Table 3 presents the estimated ROIE in the six parametric models. Focusing on the results of the FE models, we find that when interviewed by a black interviewer in comparison to a nonblack interviewer, with everything else equal, a respondent perceived a larger gap between the intelligence of whites and blacks that is in favor of blacks (coefficient $= 0.36$, $p < .001$).[18] The respondent also tended to report that the current social spending on welfare and on helping blacks was too little (coefficient $= 0.19$, $p < .001$; coefficient $= 0.29$, $p < .001$, respectively) and tended to report that the government should take more responsibility to help blacks (coefficient $= 0.53$, $p < .001$). The estimated ROIE for other outcomes are not statistically significant at the 5 percent significance level. A graphic presentation of the results is shown in Figure 1, which presents the point estimates and their 95 percent confidence intervals in the FE models.

It is also worth noting that the estimated ROIE for the four outcomes mentioned earlier are statistically significant and of similar magnitudes across all

Table 3. Estimated Race of Interviewer Effects on 10 Selected Outcomes.

	(1)	(2)	(3)	(4)	(5)	(6)
Outcomes	RE	FE	FD	RTS	MA(1)	AR(1)
Perceived intelligence gap	0.43***	0.36***	0.32**	0.43***	0.43***	0.27
	0.07	0.09	0.10	0.07	0.07	0.15
	2,616	2,617	1,407	2,616	2,616	658
Confidence in executive branch of federal government	0.03	−0.00	0.02	0.03	0.04	−0.09
	0.04	0.05	0.06	0.04	0.04	0.12
	2,695	2,696	1,449	2,695	2,695	678
Confidence in Congress	0.07*	0.06	0.08	0.07*	0.07*	0.09
	0.04	0.04	0.05	0.04	0.04	0.09
	2,700	2,701	1,451	2,700	2,700	683
Confidence in Supreme Court	0.07	0.03	0.07	0.07	0.08*	−0.10
	0.04	0.05	0.05	0.04	0.04	0.09
	2,684	2,685	1,432	2,684	2,684	668
Confidence in bank and financial institutions	0.05	−0.03	−0.06	0.05	0.05	−0.02
	0.04	0.05	0.05	0.04	0.04	0.11
	2,713	2,714	1,461	2,713	2,713	688
Spending on welfare	0.19***	0.19**	0.16*	0.19***	0.19***	0.26*
	0.05	0.06	0.07	0.05	0.05	0.11
	1,990	1,991	1,059	1,990	1,990	493
Spending on blacks	0.32***	0.29***	0.28***	0.32***	0.32***	0.30*
	0.04	0.05	0.05	0.04	0.04	0.13
	1,858	1,859	938	1,858	1,858	422
Should help blacks	0.57***	0.53***	0.47***	0.56***	0.57***	0.46***
	0.07	0.09	0.09	0.07	0.07	0.16
	2,662	2,663	1,405	2,662	2,662	645
Should help poor	0.08	0.01	0.00	0.08	0.08	−0.02
	0.07	0.08	0.08	0.07	0.06	0.15
	2,676	2,677	1,423	2,676	2,676	653
Should help sick	0.11	0.07	0.02	0.11	0.11	0.03
	0.07	0.09	0.10	0.07	0.07	0.17
	2,683	2,684	1,428	2,683	2,683	659

Note: Models 1 to 6 are random effects (RE), fixed effects (FE), first difference (FD), random trend and slope (RTS), dynamic models with moving average (MA)(1) process, and dynamic models with autoregressive (AR)(1) process. For each outcome, the coefficient is shown in the first line, the standard error clustered at the individual level in the second line, and the sample size in the third line. For conciseness of presentation, results for other variables are not shown. *$p < .05$. **$p < .01$. ***$p < .001$.

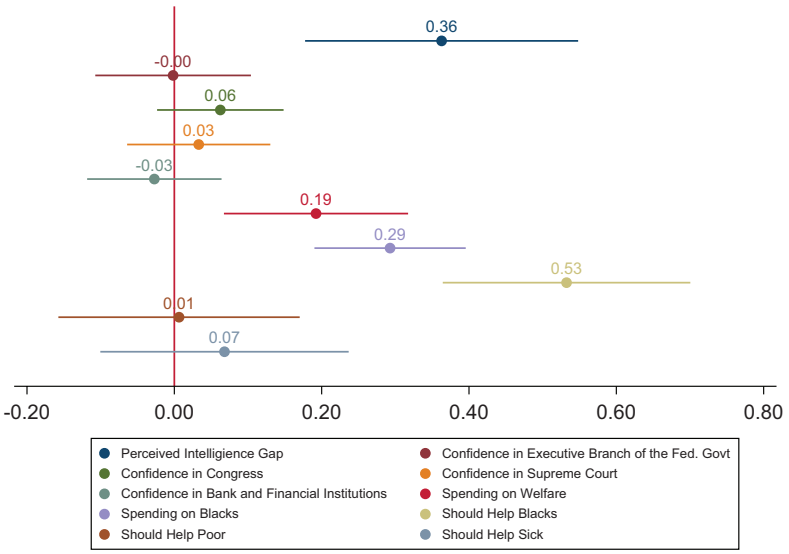

Figure 1. Estimated race-of-interviewer effects (ROIE) in the fixed effects (FE) models as shown in Table 3.

the models.[19] By contrast, the estimated ROIE for other outcomes are either consistently insignificant or inconsistently significant across the models. For example, the estimated ROIE for confidence in the Congress and in the Supreme Court are statistically significant only in some of the RE models but not in any of the FE models. Taken together, these results suggest that ROIE appear to mainly concentrate on race-specific survey questions.

Interviewer effects by interview mode. Table 4 presents the estimated coefficients for the interaction term between interviewer's race and interview mode, which we used to measure whether in-person interviews has led to more ROIE.[20] The RE estimates appear to suggest that ROIE may be larger if the interview was conducted in person than by phone for confidence in the executive branch of the federal government and for confidence in the Supreme Court. But the more credible (and also conservative) results in the FE models show that the estimated coefficients for the interaction term are not significant at the 5 percent significance level for any outcome.

The lack of evidence for ROIE varying by interview mode could be due to two things. First, there may be too few interviews conducted in person by

Table 4. Estimated Race of Interviewer Effects by Interview Mode (in Person vs. by Phone).

	(1)	(2)	(3)	(4)	(5)	(6)
Outcomes	RE	FE	FD	RTS	MA(1)	AR(1)
Perceived intelligence gap	−0.05	−0.19	−0.32	−0.03	−0.07	−0.74
	0.21	0.28	0.35	0.20	0.22	0.41
Confidence in executive branch of federal government	0.29**	0.20	0.20	0.30**	0.28*	0.29
	0.11	0.13	0.14	0.11	0.12	0.26
Confidence in Congress	0.06	−0.09	−0.22	0.06	0.05	−0.26
	0.12	0.15	0.15	0.12	0.11	0.23
Confidence in Supreme Court	0.21*	0.13	0.09	0.22*	0.21	−0.02
	0.11	0.13	0.13	0.11	0.11	0.21
Confidence in bank and financial institutions	0.10	0.18	0.14	0.09	0.10	0.18
	0.12	0.15	0.15	0.12	0.11	0.24
Spending on welfare	−0.16	−0.18	−0.10	−0.15	−0.15	−0.26
	0.16	0.18	0.16	0.15	0.14	0.24
Spending on blacks	−0.14	−0.17	−0.20	−0.13	−0.15	−0.41
	0.11	0.13	0.15	0.11	0.12	0.31
Should help blacks	0.05	0.09	−0.06	0.06	0.05	−0.03
	0.20	0.21	0.23	0.20	0.20	0.31
Should help poor	−0.06	0.03	0.03	−0.04	−0.04	0.16
	0.22	0.28	0.26	0.22	0.20	0.29
Should help sick	−0.14	−0.03	−0.02	−0.12	−0.14	0.22
	0.21	0.25	0.25	0.21	0.21	0.37

Note: Models 1 to 6 are random effects (RE), fixed effects (FE), first difference (FD), random trend and slope (RTS), dynamic models with moving average (MA)(1) process, and dynamic models with autoregressive (AR)(1) process. The coefficient for the interaction term between interview mode and interviewer's race (coded as 1 if an interview was conducted by a black in person and as zero otherwise) is shown in the first line for each outcome. Standard error clustered at the individual level is shown below the coefficient. The sample size is the same as that shown in Table 3 for each model of each outcome, which is omitted from reporting in this table. *$p < .05$. **$p < .01$. ***$p < .001$.

blacks so that the estimates tend to have large standard errors. Second, a more substantive explanation may be that respondents were able to infer the interviewer's race from vocal features (Cotter et al. 1982) and tailor their responses to cater to their interviewer.

Interviewer effects by respondent's race. Table 5 presents the estimated coefficients for the interaction term between interviewer's race and respondent's race, which we used to measure whether nonblack respondents are more

Table 5. Estimated Race of Interviewer Effects by Respondent's Race (Nonblack vs. Black).

	(1)	(2)	(3)	(4)	(5)	(6)
Outcomes	RE	FE	FD	RTS	MA(1)	AR(1)
Perceived intelligence gap	−0.21	−0.08	−0.11	−0.20	−0.22	0.47
	0.17	0.20	0.23	0.16	0.15	0.26
Confidence in executive branch	0.11	0.07	0.02	0.12	0.09	0.09
of federal government	0.09	0.12	0.13	0.09	0.08	0.27
Confidence in Congress	0.05	−0.01	−0.11	0.04	0.03	0.18
	0.08	0.10	0.11	0.08	0.08	0.18
Confidence in Supreme Court	0.02	0.03	−0.04	0.03	0.00	−0.04
	0.09	0.12	0.13	0.09	0.08	0.19
Confidence in bank and	0.03	0.05	0.08	0.02	0.04	0.16
financial institutions	0.08	0.11	0.12	0.08	0.08	0.23
Spending on welfare	0.17	0.09	0.05	0.18	0.17	0.18
	0.11	0.13	0.14	0.11	0.10	0.19
Spending on blacks	0.32***	0.29**	0.26*	0.32***	0.30***	0.28
	0.08	0.10	0.11	0.08	0.08	0.24
Should help blacks	0.43**	0.45*	0.39	0.44**	0.44**	0.27
	0.16	0.20	0.22	0.16	0.14	0.34
Should help poor	0.23	0.20	0.05	0.25	0.20	−0.19
	0.15	0.19	0.20	0.15	0.14	0.32
Should help sick	−0.21	−0.24	−0.49*	−0.17	−0.24	−0.89*
	0.15	0.19	0.21	0.15	0.14	0.35

Note: Models 1 to 6 are random effects (RE), fixed effects (FE), first difference (FD), random trend and slope (RTS), dynamic models with moving average (MA)(1) process, and dynamic models with autoregressive (AR)(1) process. The coefficient for the interaction term between respondent's race and interviewer's race (coded as 1 if a nonblack respondent was interviewed by a black and as zero otherwise) is shown in the first line for each outcome. Standard error clustered at the individual level is shown below the coefficient. The sample size is the same as that shown in Table 3 for each model of each outcome, which is omitted from reporting in this table.
*$p < .05$. **$p < .01$. ***$p < .001$.

subject to ROIE than black respondents. Looking at the more credible results in the FE models, we find that ROIE is larger for nonblack respondents than for black respondents in current spending on welfare (coefficient $= 0.29$, $p < .001$) and in government's responsibility to help the blacks (coefficient $= 0.45$, $p < .05$). In other words, the black interviewer effect is significantly larger among nonblack respondents than among black respondents for these two outcomes.

Table 6. Combined Matching Estimates of the Race of Interviewer Effects From Waves 2 and 3.

	ATE			ATT		
Outcomes	Estimates	*SE*	*N*	Estimates	*SE*	*N*
Perceived intelligence gap	0.94	0.15***	1,406	0.78	0.19***	173
Confidence in executive branch of federal government	0.02	0.09	1,449	0.09	0.09	196
Confidence in Congress	0.06	0.06	1,451	−0.02	0.08	196
Confidence in Supreme Court	0.04	0.07	1,432	0.09	0.08	192
Confidence in bank and financial institutions	0.12	0.08	1,461	−0.17	0.08*	197
Spending on welfare	0.47	0.13***	1,059	0.24	0.11**	140
Spending on blacks	0.23	0.10**	938	0.51	0.09***	128
Should help blacks	0.44	0.13***	1,405	0.47	0.14***	186
Should help poor	0.05	0.14	1,423	0.00	0.13	189
Should help sick	0.04	0.15	1,428	−0.02	0.15	195

Note: Panel 1 shows matching estimates of the race-of-interviewer effects (ROIE) for all respondents regardless of whether they were interviewed by a black, that is, the average treatment effects (ATE). Panel 2 shows matching estimates of the ROIE for respondents who were interviewed by a black, that is, the ATE for the treated (ATT). For each measure, the first differenced outcome is used. Robust standard errors are reported
*$p < .05$. **$p < .01$. ***$p < .001$.

The fact we have lumped white respondents and other nonblack respondents into a single group may have diluted the estimates of ROIE. To investigate that possibility, we trimmed the sample to include only white and black respondents and specified a new interaction term that equals to one if a white respondent was interviewed by a black interviewer and zero otherwise. The patterns of the results are similar to those shown in Table 5 and so we show them in Table A3 in the Online Appendix.

Matching Estimates of ROIE

Table 6 presents the aggregate matching estimates of ROIE. The separate estimates from waves 2 and 3 can be found in Table A4 in the Online Appendix. We present two sets of estimates: one for ATE and the other for ATT. Looking at the estimates of ATE, we find that the results are quite similar to those produced by the FE models as shown in Table 3. ROIE are found to be statistically significant in the same four outcomes (i.e., perceived intelligence gap, spending on welfare, spending on blacks, and should help

blacks). Matching and the FE models only differ somewhat in their point estimates. For example, the matching estimate of the ROIE for perceived intelligence gap appears to be larger, probably because matching has provided a more focused causal estimate.

The second panel in Table 6 shows the estimates of ATT. The patterns of the results by and large are similar to the estimates of ATE except minor differences in point estimates. But at the same time, we find that the treated group (i.e., the respondents who were indeed interviewed by a black interviewer) reported a lower confidence in bank and financial institutions than when they were (counterfactually) interviewed by a nonblack interviewer (coefficient $= -0.17$, $SE = 0.08$, $p < .05$). If we compare this estimate with the corresponding ATE estimate (coefficient $= 0.12$, $SE = 0.08$), however, there is no statistically significant difference. Overall, these results suggest that there may not be much heterogeneity in ROIE by treatment status.

Matching also allows us to examine how ROIE evolve over time. Table A5 in the Online Appendix shows the estimated ROIE by waves. The estimated ROIE do not change in a universal fashion over time. For some outcomes, they become larger, while for others they become smaller. For example, in terms of ATE, there is evidence of ROIE for perceived intelligence gap, spending on blacks, and should help blacks in wave 2, but in wave 3 there is no evidence of ROIE for spending on blacks any longer whereas there is evidence of ROIE for confidence in bank and financial institutions. These different estimates suggest that ROIE may vary across time for some outcomes.

Summary and Discussion

This article presents methods for drawing causal inference in panel data. The parametric models we examined can be divided into two groups: RE models and FE models. The RE models assume that unobserved time-invariant factors are not correlated with covariates and treatment while the FE models allow for such correlations. Since consistency is of primary concern in many cases, we suggest focusing on the FE models to provide causal evidence. We outlined several tests that can offer guidance on how to choose among the different models.

In this article, we also extended the latest advances in matching methods to panel data analysis. We proposed combining matching with DID in order to utilize the strengths of both methods. Matching helps provide better covariate support and a more focused causal inference while DID helps remove the effects of unobserved time-invariant factors. We outlined the conditions for this strategy to work effectively and provided solutions to aggregating estimates from multiple waves. Overall, we advocated for using several

different models which generally rely on different assumptions to triangulate the evidence for any causal quest.

We applied these methods to analyzing ROIE on ten selected survey items in the 2006 to 2010 panel data of the GSS. We found that respondents interviewed by a black interviewer as opposed to a nonblack interviewer in general provided more favorable responses toward blacks. However, it seems that such interviewer effects, for the most part, pertained to only race-related survey items. Since our unique approach utilizes the strengths of both parametric models and nonparametric matching methods, we provided robust and consistent evidence on race of interviewer effects.

Our study may be improved and extended in both substantive and methodological directions. First, our approaches may be readily applied to studying ROIE on other survey items. Second, we could look into other forms of interviewer effects, like those based on gender, political ideology, religion, and so on. Third, future studies may develop alternative methods for combining matching estimates from multiple waves. For example, meta analysis may be used to combine the results from multiple waves, where it is possible to weight the estimates by the inverse of their estimated variance. Fourth, it is important to develop and study methods for dealing with interviewer effects. For example, whether including interviewer's characteristics in statistical models would be sufficient to address the problems brought by interviewer effects. For another example, whether moving interviews to computers or the Internet will help eliminate the interviewer effect. To be sure, doing so needs to be examined in a larger context, as it may introduce other representation and measurement errors.

Acknowledgment

We thank participants of the General Social Survey Workshop organized by the National Opinion Research Center at the University of Chicago, in particular, Duane Alwin, Kenneth Bollen, Mark Chaves, Michale Hout, Peter Marsden, Cyrus Schleifer, Tom Smith, Steve Vaisey, and John Robert Warren for their helpful comments on an earlier presentation of this article. Our special thanks go to Professor Michale Hout and anonymous reviewers for their valuable suggestions to improve this article and to Genevieve Butler for proofreading this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Another possible way to model the individual fixed effects (FE) is to include a dummy variable for each subject. In linear models, this approach will provide identical estimates to the FE model (Wooldridge 2001:267).
2. Strictly speaking, what equations (2) and (3) show is estimation methods. Differencing is just used as a shortcut to facilitate the estimation of the coefficients. Hence, the estimates should still be interpreted according to the original model as shown in equation (1) (Wooldridge 2001:283).
3. This is because under the assumption that there is no serial correlation in e_{it} and that e_{it} is distributed with a mean zero and the same variance,

$$\begin{aligned}\rho_2 &= \mathrm{Cov}(\Lambda e_{it}, \Lambda e_{it-1})/\mathrm{Var}(\Lambda e_{it-1}) \\ &= \mathrm{Cov}(e_{it} - \bar{e}_i, e_{it-1} - \bar{e}_i)/\mathrm{Var}(e_{it-1} - \bar{e}_i) = -\frac{1}{T}\Big/\left(1 - \frac{1}{T}\right) \\ &= -1/(T-1)\end{aligned}$$

4. A fixed effects variant of the random trend and slope (RTS) model is also possible. See Wooldridge (2001:315-17).
5. The propensity score matching (Rosenbaum and Rubin 1983) avoids the problem of choosing a weighting metric. In brief, it predicts the probability of each unit receiving the treatment and matches units with similar propensity scores. Rosenbaum and Rubin (1983) show that in expectation this is equivalent to matching on covariates. The drawback of this approach is that (1) in general, a parametric model is needed for predicting the propensity scores and (2) the prediction of propensity scores generates additional uncertainties for estimating treatment effects. We skip the details of this approach, but interested readers can consult Rosenbaum and Rubin (1983), Morgan and Winship (2007), An (2010), and Abadie and Imbens (2012).
6. In the following formulas for estimating treatment effects, we have omitted a bias term that may result from inexact matching. But in the ROIE example, the estimates do account for such bias. See Abadie et al. (2004) and Abadie and Imbens (2006) for discussion on the bias and adjustment strategies.
7. Given that $\hat{\delta}_g = \frac{1}{N_g}\sum_{i=1}^{N_g}(2D_{ig} - 1)(1 + K_{ig})\Delta Y_{ig}$, where K_{ig} is the number of times unit i serves as a match to other units at wave g, and a similar formulation for $\hat{\delta}_h$, $\mathrm{Cov}(\hat{\delta}_g, \hat{\delta}_h)$ may be estimated by $\mathrm{Cov}\big(\frac{1}{n}\sum_{i=1}^{n} J_{ig}\Delta Y_{ig},$

$\frac{1}{n}\sum_{i=1}^{n} J_{ih}\Delta Y_{ih}) = \frac{1}{n^2}\sum_{i=1}^{n} J_{\mathrm{ig}}\, J_{\mathrm{ih}}\, \mathrm{Cov}(\Delta Y_{ig}, \Delta Y_{ih})$, where n is the number of common observations in waves g and h, $J_{ig} = (2D_{ig} - 1)(1 + K_{ig})$, and $J_{ih} = (2D_{ih} - 1)(1 + K_{ih})$.

8. More extensive literature review on interviewer effects and race-of-interviewer effects can be found in Schaeffer et al. (2010), particularly pages 451-53.
9. See subsection 3.3 of Wooldridge (2013) for a brief introduction to this issue.
10. GSS respondents are randomly assigned to ballots. So the missing data here are by design and do not present any selection problem.
11. Some nonblack interviewers may be multi-race of which one is black. They can potentially be viewed by respondents as blacks too. Thus, our estimates are conservative and may have slightly underestimated the ROIE.
12. To clarify, the social desirability argument also applies to white interviewers. The respondents being interviewed by white interviewers are expected to provide more favorable responses toward whites (and perhaps also less favorable responses toward blacks) than if they are interviewed by nonwhite interviewers. Here we focus on black interviewers because we follow the statistical convention of using the larger group (here the nonblack interviewers) as the reference group and also because responses to the survey questions are presumably more sensitive to black interviewers. The intention of this article is not to single out black interviewers, but to show that respondents may adjust their responses according to the interviewer's race.
13. Similarly, we can argue that black respondents may be more likely to adjust their responses than white respondents if interviewed by white interviewers.
14. We are only interested in testing ROIE for each individual outcome. But if we want to draw inference on the ten outcomes altogether and maintain an overall type 1 error at a desired significance level α, we would request a significance level of α/n (n being the number of tests) for each of the n tests by following the Bonferroni correction procedure. In other words, it amounts to multiplying each p value by ten in our case to check for significance.
15. For simplicity, we assume treatment effects are uncorrelated across waves when calculating the standard errors in the aggregate results.
16. We also calculated the response rate for each survey question by interviewer's race. We took into account the fact that not all outcome measures appeared in each ballot. We find no statistically significant difference at the 5 percent significance level between the response rates for the two groups in any of the survey questions. Results are shown in Table A1 in the Online Appendix.
17. In future work, it would be interesting to examine the fixed effects variants of the RTS model and see whether heterogeneity in time trend and treatment effects is still present. See subsection 11.2 of Wooldridge (2001).

18. We also studied ROIE on the ratings of intelligence of whites and blacks separately. It appears that ROIE raises the rating of blacks while at the same time lowers the rating of whites. See Table A2 in the Online Appendix.
19. The only exception is that the estimated ROIE for the perceived intelligence gap in the AR(1) model is not statistically significant despite it is significant in all other models. Considering that the standard error is quite large for the estimate in the AR(1) model, we cannot really make the case that it is statistically different from the estimates in other models.
20. These models also included the main effects: interviewer's race and interview mode. For brevity, we did not report their estimates.

Supplemental Material

The online appendix is available at http://smr.sagepub.com/supplemental.

References

Abadie, Alberto, David Drukker, Jane Leber Herr, and Guido Imbens. 2004. "Implementing Matching Estimators for Average Treatment Effects in Stata." *The Stata Journal* 1:1-18.

Abadie, Alberto and Guido Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74:235-67.

Abadie, Alberto and Guido Imbens. 2012. "Matching on the Estimated Propensity Score." *Harvard University and NBER Working Paper*. Retrieved January 29, 2013, from http://www.hks.harvard.edu/fs/aabadie/pscore.pdf.

An, Weihua. 2010. "Bayesian Propensity Score Estimators: Incorporating Uncertainties in Propensity Scores into Causal Inference." *Sociological Methodology* 40: 151-89.

Anderson, Barbara A., Brian D. Silver, and Paul R. Abramson. 1988. "The Effects of the Race of the Interviewer on Race-related Attitudes of Black Respondents in SRC/CPS National Election Studies." *Public Opinion Quarterly* 52:289-324.

Arellano, M. and S. Bond. 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *Review of Economic Studies* 58:277-97.

Berger, Lawrence M., Sarah K. Bruch, Elizabeth I. Johnson, Sigrid James, and David Rubin. 2009. "Estimating the 'Impact' of Out-of-home Placement on Child Well-being: Approaching the Problem of Selection Bias." *Child Development* 80:1856-76.

Cotter, Patrick R., Jeffery Cohen, and Philip B. Coulter. 1982. "Race-of-interviewer Effects in Telephone Interviews." *Public Opinion Quarterly* 46:278-84.

Davis, Darren W. 1997. "Direction of Race of Interviewer Effects among African Americans: Donning the Black Mask." *American Journal of Political Science* 41:309-22.

Davis, Darren W. and Brian D. Silver. 2003. "Stereotype Threat and Race of Interviewer Effects in a Survey on Political Knowledge." *American Journal of Political Science* 47:33-45.

Davis, R. E., M. P. Couper, N. K. Janz, C. H. Caldwell, and K. Resnicow. 2010. "Interviewer Effects in Public Health Surveys." *Health Education Research* 25: 14-26.

Durrant, Gabriele B., Robert M. Groves, Laura Staetsky, and Fiona Steele. 2010. "Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys." *Public Opinion Quarterly* 74:1-36.

Finkel, Steven E., Thomas M. Guterbock, and Martin J. Borg. 1991. "Race-of-interviewer Effects in a Preelection Poll: Virginia 1989." *Public Opinion Quarterly* 55:313-30.

Hox, J. J. 1994. "Hierarchical Regression Models for Interviewer and Respondent Effects." *Sociological Methods & Research* 22:300-18.

Kane, Emily and Laura I. Macaulay. 1993. "Interviewer Gender and Gender Attitudes." *Public Opinion Quarterly* 57:1-28.

Krysan, Maria and Mick P. Couper. 2003. "Race in the Live and the Virtual Interview: Racial Deference, Social Desirability, and Activation Effects in Attitude Surveys." *Social Psychology Quarterly* 66:364-83.

Marsden, Peter V. 2003. "Interviewer Effects in Measuring Network Size Using a Single Name Generator." *Social Networks* 25:1-16.

Morgan, Stephen and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.

Nguyen, Viet Cuong. 2012. "Program Impact Evaluation Using a Matching Method with Panel Data." *Statistics in Medicine* 31:577-88.

Paik, Anthony and Kenneth Sanchagrin. 2013. "Social Isolation in America: An Artifact." *American Sociological Review* 78:339-60.

Pickery, Jan and Geer T. Loosveldt. 2000. "Modeling Interviewer Effects in Panel Surveys: An Application." *Survey Methodology* 26:189-98.

Pickery, Jan, Geer T. Loosveldt, and Ann Carton. 2001. "The Effects of Interviewer and Respondent Characteristics on Response Behavior in Panel Surveys: A Multilevel Approach." *Sociological Methods & Research* 29: 509-23.

Rosenbaum, Paul and Donald Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41-55.

Schaeffer, Nora Cate. 1980. "Evaluating Race-of-interviewer Effects in a National Survey." *Sociological Methods & Research* 8:400-19.

Schaeffer, Nora Cate, Jeniffer Dykema, and Douglas Maynard. 2010. *Interviewers and Interviewing*. 2nd ed. Bingley, UK: Emerald Group.

Sekhon, Jasjeet. 2011. "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R." *Journal of Statistical Software* 42:1-52.

Singer, Eleanor, Martin R. Frankel, and Marc B. Glassman. 1983. "The Effect of Interviewer Characteristics and Expectations on Response." *Public Opinion Quarterly* 47:68-83.

Smith, Jeffrey and Petra E. Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125: 305-53.

Tilburg, Theo Van. 1998. "Interviewer Effects in the Measurement of Personal Network Size: A Nonexperimental Study." *Sociological Methods & Research* 26: 300-28.

Wooldridge, Jeffrey M. 2001. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: The MIT Press.

Wooldridge, Jeffrey M. 2013. *Introductory Econometrics: A Modern Approach*. 5th ed. Mason, OH: Cengage Learning.

Author Biographies

Weihua An is assistant professor of statistics and sociology at Indiana University. He graduated from Harvard with a PhD in sociology (2011) and a Master's degree in statistics (2009). His main research areas are social networks and causal inference. In social network research, he currently focuses on three major topics: causality in networks, network measurement and surveillance, and big network analysis (e.g., constructing networks from massive text data like biographies, citations, and match records and developing scalable statistical methods for analyzing big networks). In causal inference, he focuses on propensity score methods, matching methods, instrumental variable methods, and causal inference under interference. With a broad academic background, he also has strong interests in studying organizations, health and social policy, development, etc.

Christopher Winship is the Diker-Tishman Professor of Sociology, Harvard University and a member the senior faculty at the Harvard Kennedy School of Government. He is affiliated with the Harvard's Institute for Quantitative Social Science and its Center for Public Leadership. He holds a BA in sociology and mathematics from Dartmouth College and a PhD in sociology from Harvard. He is currently doing research on several topics: statistical models for causal analysis; the effects of education on mental ability; how people act when rationality is not a possibility; the linkage in Pragmatism between action and knowledge; analysis of Age-Period-Cohort models; and inner city youth behavior and violence. With Steve Morgan he is the author of *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, now in its 2nd edition.