Special Issue Article

What You Can—and Can't—Do With Three-Wave Panel Data

Sociological Methods & Research
2017, Vol. 46(1) 44-67

DOI: 10.1177/0049124114547769
journals.sagepub.com/home/smr

Stephen Vaisey[1] and Andrew Miles[2]

Abstract

The recent change in the general social survey (GSS) to a rotating panel design is a landmark development for social scientists. Sociological methodologists have argued that fixed-effects (FE) models are generally the best starting point for analyzing panel data because they allow analysts to control for unobserved time-constant heterogeneity. We review these treatments and demonstrate the advantages of FE models in the context of the GSS. We also show, however, that FE models have two rarely tested assumptions that can seriously bias parameter estimates when violated. We provide simple tests for these assumptions. We further demonstrate that FE models are extremely sensitive to the correct specification of temporal lags. We provide a simulation and a proof to show that the use of incorrect lags in FE models can lead to coefficients that are the opposite sign of the true parameter values.

Keywords

panel data, longitudinal, GSS, fixed-effects, surveys

[1] Duke University, Durham, NC, USA
[2] University of Toronto, ON, Canada

Corresponding Author:
Stephen Vaisey, Duke University, Box 90088, 417 Chapel Drive, Durham, NC 27708, USA.
Email: stephen.vaisey@duke.edu

Introduction

The recent change in the general social survey (GSS) to a rotating panel design is a landmark development for social scientists. Panel designs have two key advantages over cross-sectional surveys. First, because each respondent appears in the data multiple times, he or she can in some sense serve as his or her own "control group," allowing for more valid causal inferences (Allison 2009:1). Second, because respondents are measured over time, it is sometimes possible to use temporal ordering to ask more complicated questions about social processes. But despite the widely acknowledged advantages of panel data, many researchers do not know how to make the most of them. Less-than-optimal uses of panel data are common even in work published in high-prestige outlets (e.g., see Halaby 2004).[1] This is perhaps not surprising, given (among other reasons) that very few graduate programs in sociology incorporate training on panel data as part of the required curriculum.

The goal of this article is to provide some practical and theoretical guidance for researchers who have a good grasp of regression but who have limited experience with panel data. In the process, we summarize existing best practices before offering analyses or extensions original to this article. Because our target audience is not the high-end user but rather the "mid-end" user, we cannot jump right to our original contributions but must first set the stage by outlining what has gone before. Readers already familiar with panel data and fixed-effects (FE) models can skip to section on Testing the Limits of FE Models.

We explore the promise and pitfalls of panel data under two major headings. We first consider the ability of panel data to help control for unobserved heterogeneity. Though this is not recognized in sociology as often as it could be, the primary *raison d'être* of panel data is to help control for unmeasured variables (Halaby 2004:508). We demonstrate how to make the most of these possibilities in the context of the GSS panel while avoiding common pitfalls. We then turn to the use of panel data to establish temporal ordering. Sociologists frequently use the ordering of data to attempt to establish causal order, often using lagged values of key predictors (e.g., see Cha 2010; Faris and Felmlee 2011). Though this can be a useful strategy in some circumstances, we show that it can easily lead to misleading results.

We do not attempt to offer a rigorous statistical consideration of the issues involved since there are already many technical treatments (see, e.g., Wooldridge 2010). Instead, we aim to distill the recommendations of specialists in these methods, to demonstrate how various models might be useful (or

hazardous), and to consider the assumptions of different models from the point of view of our knowledge about the social world. We undertake these goals squarely in the context of the GSS panel and thus do not consider models that might be more appropriate with different data structures (e.g., cross-sectional studies or panels with more than three waves of data).

Panel Data and Unobserved Heterogeneity

Review: Common Models for Panel Data

In the GSS panel, respondents are asked (most) questions three times, once at each wave. Consider a model for the response a respondent might give to a particular question:

$$y_{it} = \mu_t + \mathbf{x'_{it}}\beta + \mathbf{z'_i}\gamma + \upsilon_i + \epsilon_{it}, \quad i = 1, \ldots, N, \;\; t = 1, 2, 3. \qquad (1)$$

Equation (1) asserts that the answer respondent i gives to question y at time t is a function of five things: what's going on in the world at that time that affects everyone equally (the intercepts μ_t), the values of any observed time-*varying* variables for the respondent, like age or income (the $\mathbf{x_{it}}$ variables), the values of any observed time-*constant* variables for the respondent, like gender or race ($\mathbf{z_i}$ variables), some *unobserved* time-constant, person-specific "stuff" (like personality) that affects the respondents answers equally at all three waves (υ_i), and some other idiosyncratic "stuff" that varies from wave to wave for each respondent (ϵ_{it}).[2] As always in regression, the main threats to causal inference come from unobserved variables, here υ_i and ϵ_{it}. If the observed $\mathbf{x}$ and $\mathbf{z}$ variables are correlated with these unobserved factors, then estimates of their effects will be biased.

There are three common ways of handling these sorts of data. The first, typical in sociology (especially with two waves of data), is the lagged dependent variable (LDV) model (Halaby 2004:535). The LDV model takes the following form, where $t = 2$ refers to the second of two waves of data:

$$y_{it} = \mu + \rho y_{t-1} + \mathbf{x'_{it}}\beta + \mathbf{z'_i}\gamma + \epsilon_{it}, \quad t = 2. \qquad (2)$$

The idea behind this model (though not always articulated) is that the lagged value of y will serve as a proxy for υ, the unobserved between-person heterogeneity that appears in equation (1).[3] One hopes that controlling for it will allow for less biased estimates of the effects of the measured predictors (see Morgan and Winship 2007:179-81). The primary shortcoming of this strategy is that it does not take full advantage of the panel data structure, relying on unclear assumptions about the relationship between y_{t-1} and υ instead of

attempting to model υ directly. We will show subsequently how this method can yield less-than-optimal results.

The other two most common ways of dealing with panel data are random-effects (RE) models and FE models. These are unfortunate names, however, since they do not convey the real differences between them (see Wooldridge 2010:285-86). Both FE and RE use the repeated measures of the outcome offered by panel data to estimate the υ_i. The only difference between RE and FE lies in the assumption they make about the relationship between υ and the observed predictors: *RE models assume that the observed predictors in the model are not correlated with* υ *while FE models allow them to be correlated.* A moment's reflection on what υ represents—all unmeasured time-constant factors about the respondent—should lead anyone to realize that the RE assumption is heroic in social research, to say the least. The idea that the characteristics we don't (or can't) measure (like personality or genetic influences) are uncorrelated with the things we usually do measure (like income or church attendance) is implausible.[4]

FE models avoid the RE assumption by using only within-respondent variation to estimate the x coefficients. In essence, FE models "subtract off" both observed and unobserved time-constant factors using the panel structure of the data. There are a few ways to do this in practice. For simplicity, assume we have only one time-varying predictor whose effect we want to estimate. The most straightforward approach in an ordinary least squares (OLS) context is just to give each respondent his or her own intercept. This is shown in equation (3), where υ_i stands for a dummy variable added to the model for each respondent.[5]

$$y_{it} = \mu_t + \beta x_{it} + \upsilon_i + \epsilon_{it}. \tag{3}$$

Another strategy is mean differencing, which gives the same estimates for β (equation [4])[6]:

$$y_{it} - \bar{y}_i = (\mu_t - \bar{\mu}) + \beta(x_{it} - \bar{x}_i) + (\epsilon_i - \bar{\epsilon}_i). \tag{4}$$

A final, slightly different, FE model is the change score, or first-difference (FD) model, equation (5), which is like the mean difference model except it subtracts off the respondent's prior value rather than his or her overall mean.[7]

$$y_{it} - y_{it-1} = (\mu_t - \mu_{t-1}) + \beta(x_{it} - x_{it-1}) + (\epsilon_i - \epsilon_{it-1}). \tag{5}$$

Regardless of the exact estimation strategy, however, the strength of the FE approach is that it controls for everything specific about a respondent that

does not vary by time and that has constant effects on the outcome, including rarely measured factors like personality or genetics.

But there are two costs. First, as we noted, FE models use only within-respondent variation to estimate parameters. If a given respondent doesn't, for instance, have any variability in church attendance from wave to wave, then he or she contributes nothing to estimating the effect of church attendance on any outcomes. Because they rely on less variation, parameter estimates from FE models typically have larger standard errors. But many scholars have argued that this is a small price to pay to avoid the RE assumption if it is incorrect (which it almost always is; see Halaby 2004:527).

The second cost of using FE models is that one cannot get estimates of the effects of time-constant predictors. That is, since FE models use only within-respondent variation, it's impossible to estimate the effects of things that don't vary within respondents. For sociologists, this is a seemingly high cost, since many of the things we're interested in (such as race, gender, or family background) don't change over a 4-year period, if ever. This is surely one main reason that sociologists have been slower to adopt FE methods than economists or other social scientists.

Fortunately, it's quite easy to get around the latter objection. There are models that allow combining FE estimates of the effects of time-varying variables with RE-type estimates of the effects of time-constant factors. The most straightforward of these is Allison's hybrid model (Allison 2009:23-25). If we have one time-varying variable (e.g., work hours) and one time-constant variable (e.g., gender), Allison's model is as follows:

$$y_{it} = \mu_t + \beta(x_{it} - \bar{x}_i) + \omega\bar{x}_i + \gamma z_i + \upsilon_i + \epsilon_{it}. \quad (6)$$

This model is essentially an RE model with an FE twist. Just like an RE model, it assumes that υ is uncorrelated with the predictors, but it differs by modeling the time-varying and time-constant parts of x separately. This makes the estimate of β the same as all other FE models while allowing for the inclusion of time-constant variables.

Two Illustrations

Regardless of the exact model used, the consensus among the experts is quite strong that the FE model should be preferred over other approaches (see, e.g., Allison 2009:3; Halaby 2004:517-22). But what is the cost of using suboptimal methods? To make these points concrete, we provide two illustrations. We first present the results of a simulation that demonstrates how coefficient estimates can be biased as the correlation between the observed variables and

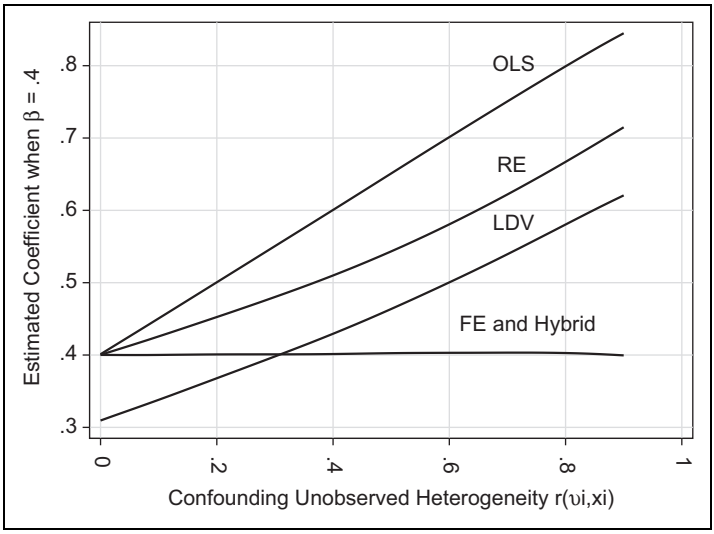

Figure 1. Simulation results for different panel models.

υ increases. Next we conduct a simple analysis of GSS data to illustrate the point in a more realistic research context.

Simulation. We simulated three-wave panel data sets that make y a function of x and υ. We varied the correlation between x and υ from 0 (no confounding time-constant unobserved heterogeneity) up to .9 (a huge degree of confounding time-constant unobserved heterogeneity). The simulation is constructed so that the true $\beta = .4$. For each data set, we estimate β using five different models: OLS, the LDV model from equation (2), RE, FE, and Allison's hybrid model. (See the online appendix for full details.)

Figure 1 shows that the FE and hybrid models are unaffected by the degree of correlation between the observed and unobserved variables. OLS does the worst, followed by RE. The LDV model does better than either OLS or RE but only produces unbiased estimates by coincidence when the amount of unobserved time-constant heterogeneity is just right. If the goal is estimating the effect of a time-varying variable (like church attendance, income, or employment status) the FE (or hybrid) estimators have a clear advantage because they purge out all time-constant unobserved factors.

GSS example. To illustrate these processes further, we consider a (simplified) example from the 2006–2010 GSS panel. Say we are interested in

the effect of church attendance, measured on a 0–8 scale, on opposition to abortion rights, measured from 0 to 6 (see Hout 1999, for more on the scale). We also think the following time-constant variables (measured in 2006) might be relevant: gender, parents' education (an indicator coded 1 if either parent had at least a bachelor's degree), and race (indicators for black and other). As before, we compare the results from a pooled OLS model, an LDV model, an RE model, an FE model, and Allison's hybrid model. Table 1 presents the estimates provided by all of these models.

We begin with estimators that do not model υ directly. The OLS estimate of the effect of attendance is .31. This does nothing to control for unobserved heterogeneity and is therefore biased if there is an association between attendance and unmeasured factors related to the outcome. The LDV model adds the value of *y* from the previous wave in an attempt to model unobserved heterogeneity through its association with υ. The estimate is reduced to .11 and the confidence interval now includes 0.

We now turn to the panel models proper. The RE coefficient is .18, between the OLS and LDV estimates. RE models assume that unobserved heterogeneity is not correlated with the observed predictors, but because this is probably violated, the estimate is likely upwardly biased. The FE coefficient in the next column tends to confirm this suspicion. Using only within-respondent variation across waves, the FE model yields a coefficient only about one-third as large (.055), though the confidence interval still does not include zero.[8] This coefficient is likely a better estimate of the actual *effect* of attendance on attitudes since it is uncontaminated by any time-constant factors with constant effects on abortion attitudes. Note, however, that there are no coefficients for time-constant variables because they do not vary wave to wave.

The final model is Allison's hybrid model. As must be the case, its estimate of the effect of within-respondent change in attendance (Δ attendance) is exactly the same as the FE estimate. Here, however, there are coefficients for all time-constant variables (including mean attendance), though they must be interpreted with caution because they are subject to standard concerns about associations with unmeasured factors.

Testing the Limits of FE Models

The illustrations mentioned previously demonstrate what is already widely known: FE methods can offer good protection against bias due to unobserved time-constant heterogeneity. If one must have a "default" model for panel

Table 1. Unstandardized Coefficients Predicting Opposition to Abortion.

	(1) OLS	(2) LDV	(3) RE	(4) FE	(5) Hybrid
y_{t-1}		.70			
		[.66, .74]			
Attendance	.31	.11	.18	.055	
	[.27, .34]	[.08, .13]	[.14, .21]	[.016, .094]	
ΔAttendance					.055
					[.016, .094]
Mean Attendance					.36
					[.31, .40]
Female	−.0011	−.056	−.13		−.052
	[−.23, .23]	[−.17, .062]	[−.11, .37]		[−.28, .18]
Parent BA	−.38	−.082	−.36		−.39
	[−.65, −.12]	[−.21, .041]	[−.63, −.08]		[−.66, −.13]
Race: black	−.19	−.18	−.054		−.24
	[−.54, .16]	[−.37, .0063]	[−.4, .29]		[−.59, .11]
Race: other	.016	−.01	−.051		.041
	[−.33, .36]	[−.2, .18]	[−.4, .3]		[−.31, .39]
2008	−.094		−.081	−.069	−.069
	[−.2, .013]		[−.18, .022]	[−.17, .033]	[−.17, .033]
2010	−.015	.11	−.013	−.011	−.011
	[−.13, .098]	[−.04, .25]	[−.12, .095]	[−.12, .095]	[−.12, .095]
Constant	1.3	.32	1.7	2.2	1.2
	[1.1, 1.6]	[.17, .46]	[1.5, 2.0]	[2.0, 2.3]	[.91, 1.4]
N	2,550	1,700	2,550	2,550	2,550

Note: FE = fixed-effects; LDV = lagged dependent variable; OLS = ordinary least squares; RE = random-effects. Values within brackets are 95 percent confidence intervals.

data, FE methods are a good place to start and have definite advantages over the more common (in sociology) RE and LDV models.

FE methods, however, rely on two assumptions that are generally not tested. The first is that selection into levels of x is based on unobserved factors (υ) rather than on previous values of y. The second is that the underlying time trajectories of y are the same regardless of the values x takes (see Morgan and Winship 2007:262-71). By construction, our simulations in the previous section were consistent with these assumptions but real data might not be.

Morgan and Winship (2007, chapter 9) explain how to test for (and in some cases, deal with) violations of these assumptions (see also Elwert and Winship, 2014). But they develop these ideas in the context of binary treatments given to some at a single time point between waves 2 and 3 of

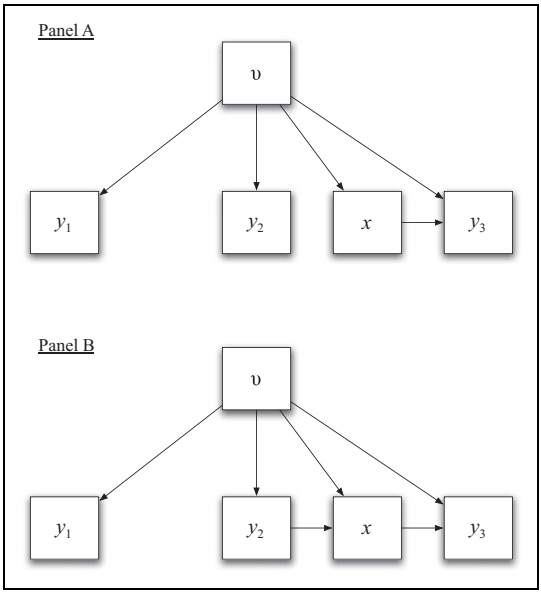

Figure 2. Causal models of a single treatment without (A) and with (B) endogenous selection.

a three-wave panel, a situation that will not correspond to many panel analyses with the GSS, where the predictors of interest might vary from wave to wave along with the outcomes. In this section, we extend Morgan and Winship's ideas to apply to continuous treatment variables whose values can vary from wave to wave.

Treatment Selection Assumption

For clarity, we begin with the case of a binary treatment received (or not) between waves 2 and 3 of a three-wave panel. Figure 2 presents these ideas in graphical form. In panel A, we have the classic FE case: y_t are functions of an unobserved time-constant fixed effect (υ), selection into the treatment (x) is based on υ, and y_3 is affected by both υ and x. In such cases, an FE model will work well because it will appropriately separate the effect of υ from the effect of x on y_3.

Panel B changes things slightly by making the treatment, x, a function of both υ and y_2, the previous wave's outcome variable. It is not difficult to

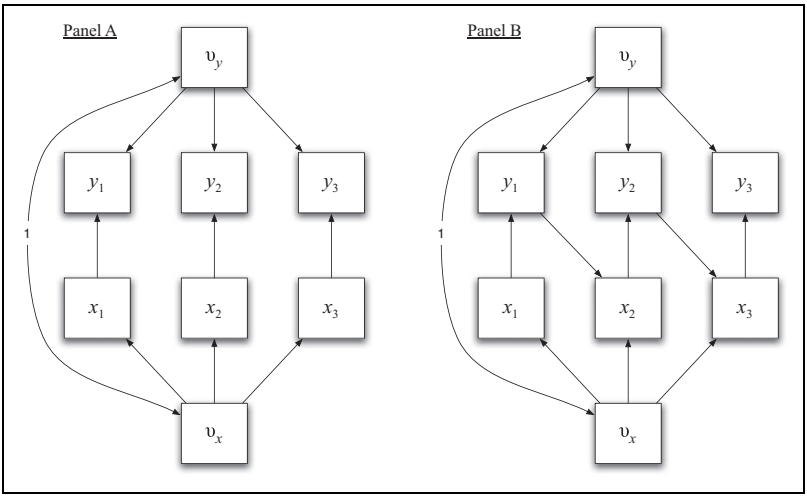

Figure 3. Causal models of an ongoing treatment without (A) and with (B) endogenous selection.

imagine situations like these; unhappiness may be both a cause and a consequence of divorce, for example. When the data are generated in this manner, FE models will not give unbiased estimates of the effect of x on y_3 because controlling for υ alone does not prevent the effect of y_2 on y_3 through x from "leaking through" into the estimate of the effect of x.

We now extend these considerations to multiple continuous "treatments," such as those measured in surveys like the GSS panel. Church attendance, for example, is a sort of treatment that can vary from wave to wave in different amounts rather than being switched on for some as a binary treatment.

Figure 3 illustrates this extension using two diagrams analogous to their counterparts in Figure 2. If the data were generated as in panel A, FE models will work well; if they were generated as in panel B, FE models will give biased estimates because controlling for υ alone does not unconfound the relationship between x_t and y_t.

Simulation. To demonstrate this point, Figure 4 presents simulations like those we presented earlier. This time, instead of varying the amount of unobserved heterogeneity (which we keep constant at $r[\upsilon_i, x_i] = .5$), we vary the extent to which the predictor of interest, x_t, is a function of y_{t-1}, the outcome at the prior wave. We hold the value of β (the effect of x_t on y_t) constant at .4.

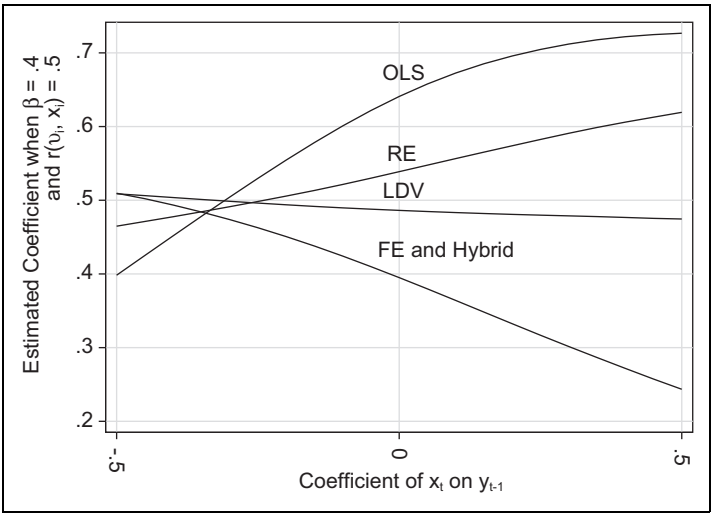

Figure 4. Simulation results under varying levels of endogenous selection.

There are a variety of biases evident in this figure. As in Figure 1, OLS and RE generally overestimate the effects of x and this bias gets worse when high levels of y_{t-1} lead to higher levels of x_t. The LDV results are more robust to endogenous selection, but they are biased because they do not properly account for unobserved factors. Finally, FE and hybrid models quickly become biased in the presence of either positive or negative selection. It is clear at a glance that none of these models is a panacea for both unobserved heterogeneity and endogenous selection into treatment.

There are models that *can* accommodate both time-constant unobserved heterogeneity and endogenous selection, but they need to be estimated using structural equation models (SEMs; see Bollen and Brand 2010). SEMs are getting easier to estimate in standard software packages, but they are still uncommon in sociology. In many cases, what we really want to know is whether we *need* to use something more complicated or whether a simpler model is adequate.

Testing the assumption. Morgan and Winship (2007:267) point out that with two pretreatment waves of data, we can test the no-endogenous-selection assumption simply and directly. Their proposed test begins with estimating the following model:

$$\log\left(\frac{Pr[x=1]}{Pr[x=0]}\right) = a + y_2 b + (y_1 + y_2)c. \quad (7)$$

The test for endogeneity is the test of $b = 0$. The logic is that if selection into the treatment (x) is a function of the unobserved fixed effect (υ) and *not* of the previous wave's outcome, y_2 will have no independent predictive power net of the more informative proxy for the FE ($y_1 + y_2$). If the data are not consistent with $b = 0$, more complex models must be explored.

We extend the Morgan and Winship (2007) approach beyond a single binary treatment. With the GSS panel and similar data sets, we may have three treatments with more continous distributions. In such situations, we do not have two waves of pretreatment data for *all* treatments, but we do have it for the treatment at wave 3. This allows testing whether x_3 can be predicted by the previous wave's outcome (y_2) net of the proxy for the fixed effect ($y_1 + y_2$). Strictly speaking, we are only testing for endogeneity at wave 3, but if we assume the test would be similar if applied to previous waves (a reasonable assumption), we can consider it an overall test of treatment endogeneity.

GSS example. Returning to our GSS example, we test whether the treatment (church attendance) at wave 3 is associated with wave 2 opposition to abortion (*abscale*) more than with the time-constant FE (proxied by the sum of opposition at waves 1 and 2). Since church attendance is not binary but has nine possible responses, it is reasonable to use OLS to estimate an analog to equation (7). Specifically, we estimate:

$$attend_3 = a + abscale_2 b + (abscale_1 + abscale_2)c + e. \quad (8)$$

The estimate of b here is –.12 [–.35, .11] and the estimate of c is .36 [.24, .48]. This means that although opposition to abortion *in general* appears to be associated with church attendance (as indicated by c), there is little evidence of an independent association with *recent* opposition to abortion. This pattern is not consistent with endogenous selection and therefore this assumption of the FE model appears reasonable. Had b been significantly different from 0, we would have to consider SEMs if we wanted to deal with unobserved heterogeneity and treatment endogeneity at the same time. For users of the GSS panel who want to use FE models, this is a quick and easy test of this assumption.

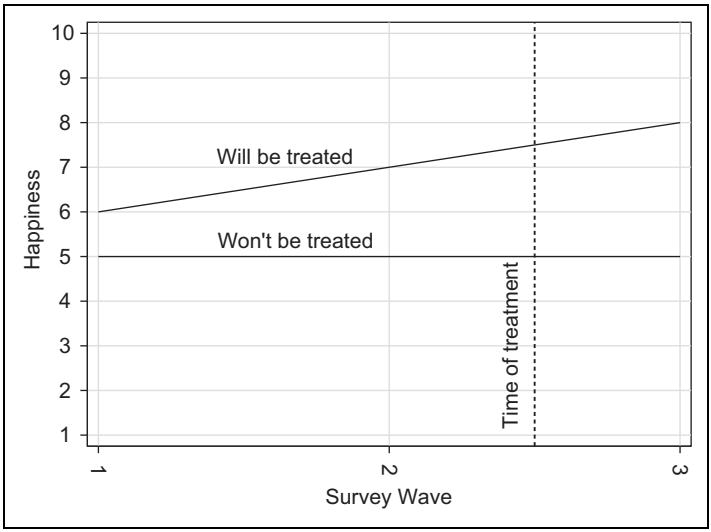

Figure 5. Hypothetical time trajectories for treated and untreated groups.

Equal Trajectories Assumption

A second assumption that is not often tested in FE models is that "treated" and "untreated" cases have the same underlying time trajectory prior to treatment. Morgan and Winship (2007:263-64) point this out, developing their critique once again in the context of three-wave data with a single treatment received (or not) between waves 2 and 3.

To fix ideas, consider two hypothetical populations: one that will get married between waves 2 and 3 of a three-wave panel and one that won't. Let us assume that, for various reasons, those who will get married later are increasing in happiness over time and those who will not are not increasing in happiness over time. Figure 5 illustrates these assumptions visually. Applying an FE model to data like these would lead to the erroneous conclusion that marriage causes increased happiness because the model fails to account for time trends that differ between will-be-treated and won't-be-treated groups prior to the treatment. Morgan and Winship propose a more flexible model that allows different treatment groups to have different time slopes. Specifying time as linear, their model is as follows:

$$y_{it} = \alpha + \beta x_{it} + \omega x_i^* + \gamma T + \gamma'(T \times x_i^*) + \epsilon_{it}, \tag{9}$$

where α is the intercept, β is the effect of the treatment (x), ω is the difference in intercepts at the first wave between will-be-treated ($x^* = 0$) and won't-be-treated ($x^* = 0$) groups, γ is the slope on time (T) for the won't-be-treated group, γ' is difference in the slope for the will-be-treated group, and ϵ is an error term. In a case like that represented in Figure 5, the model in equation (9) will correctly find that $\beta = 0$, that the treatment has no effect once time is properly taken into account. Unfortunately, few analyses using the GSS panel will take the exact form of a single binary treatment, meaning that Morgan and Winship's (2007) model cannot be directly applied.

Extending the model in equation (9) to the typical GSS case is not obvious because treatments (like level of church attendance) are ongoing rather than taking place at a discrete time in the future.[9] There is therefore no simple equivalent to x^*, a time-constant indicator of ever-treated status for each respondent.

Using Allison's hybrid model as a baseline (see equation [6]), we can, however, separate out the time-varying ($x_{it} - \bar{x}_i$) and time-constant ($\bar{x}_i$) components of x. If we allow $\bar{x}$ to interact with time, we have a model that is very close in spirit to Morgan and Winship's (2007).[10] Specfically, by estimating

$$y_{it} = \alpha + \beta(x_{it} - \bar{x}_i) + \omega\bar{x}_i + \gamma T + \gamma'(T \times \bar{x}_i) + \upsilon_i + \epsilon_{it}, \qquad (10)$$

we can allow respondents with different *average* levels of x to have different time trajectories. In this model, β will be an FE estimate of the direct effect of x on y that is not biased by potential differences in time slopes for those with different mean values of x.

As with the case of endogenous selection mentioned previously, we may want to test for these sorts of problems in the hope that using a simpler model is acceptable. Returning to the abortion attitudes and church attendence example, we estimated a hybrid model like that in Table 1, but by adding interaction terms between mean attendance and the dummy variables for 2008 and 2010. This specification did not improve model fit ($\chi^2 = 1.14$, $p = .57$), indicating that the assumption of equal time slopes is reasonable in this particular case.

Determining Causal Order

In the last two sections, we argued that panel data can facilitate causal inference by providing leverage on the problem of unobserved heterogeneity. But unobservables are not the only threats to causal inference. Determining causality is a fraught enterprise; scholars have not even reached consensus on its definition, let alone on how to determine it in particular cases (see Morgan

and Winship 2007). Bollen (1989:41), however, provides a practical definition that incorporates the major elements shared by most discussions of the issue and that can serve as a useful framework here. First, a cause x must have an *association* with an outcome y. Second, x must be *isolated* from all other causes of y to guard against spurious associations. Finally, x must come before y to establish the *direction* of the effect.[11]

All statistical models produce conditional associations. FE models add to this a high degree of isolation by removing the effects of time-constant unobserved variables with time-constant effects. The final condition is the establishment of causal direction. In some cases, determining the direction of causality is obvious (e.g., income does not affect age). But in other cases, determining causal order is both difficult and of great theoretical or substantive importance. In many sociological subfields, some of the most serious debates are about the relative causal priority of two factors that are known to be robustly associated. In the sociology of culture, for example, scholars continue to debate to what extent tastes and worldviews diffuse across existing networks and to what extent cultural similarity leads people to become friends in the first place (see, e.g., Lewis, Gonzalez, and Kaufman 2012; Lizardo 2006; Vaisey and Lizardo 2010).

Returning to our example from Table 1 (and taking it at face value), we see evidence that there is an association between church attendance and abortion attitudes that is not confounded by any time-constant factors with constant effects. But even in the unlikely event that we have also removed time-varying heterogeneity, the attendance coefficient still represents some unknown mix of the effect of attendance on attitudes *and* the effect of attitudes on attendance. Although either direction of influence is plausible, sociologists generally assume that "structural" factors (like institutional participation) cause "cultural" things like attitudes (see Vaisey and Lizardo 2010). But the models in Table 1 cannot test this assumption empirically because both variables are measured at the same time.

As we mentioned in the introduction, our review of the literature suggests that sociologists regard the establishment of causal order as the prime virtue of panel data. From a theoretical standpoint, it is certainly correct to begin from the premise that causality happens in time. Incantations about "mutual constitution" to the contrary, even reciprocal effects happen in time (see Archer 1995:65-92; Emirbayer and Mische 1998:1002-3; Vaisey and Lizardo 2010:1611-12).

By far the most common way for dealing with causal direction in sociology is the use of lagged variables. By predicting the contemporary value of y with the previous waves value of x, the goal is to determine

that x precedes y and is associated with it, coming one step closer to a persuasive causal argument. Given the earlier discussion, we might think that estimating an FE model with lagged predictors would provide the best of all worlds: protection from unobserved heterogeneity *and* the establishment of causal order. Allison (2009:94), who develops a reciprocal effects FE model, extols such models as "enhancing our ability to determine the direction of causality among variables that are associated with one another."

The use of models that combine FE and temporal ordering are relatively rare in sociology. The most straightforward of such FE estimators is the lagged first-difference (LFD) model:

$$y_{it} - y_{it-1} = (\mu_t - \mu_{t-1}) + \beta(x_{it-1} - x_{it-2}) + (\epsilon_{it} - \epsilon_{it-1}), \quad t = 3. \quad (11)$$

In the three-wave case, the LFD estimator models the change in y between waves 2 and 3 with the change in x between waves 1 and 2 (see Martin, Van Gunten, and Zablocki [2012]). Because it uses differences only, any time-constant unobserved heterogeneity is removed. This model is very clear in its assumptions—the first change is the cause of the second change, and there is no contemporaneous effect of x on y.

It is important to consider the reasonableness of this assumption, given the data we are using. In an abstract sense, the t subscript used in all models simply indexes the *ordering* of time. These units of time could be anything from a nanosecond (or less) to a millennium (or more). Compared to data that are widely spaced (as with the GSS panel), short lags will appear practically contemporaneous. This suggests that whether a lagged cause is reasonable, given the available data is a matter for theory and substantive knowledge, not for mathematics or statistics (see Martin et al. 2012:33).

For users of the GSS, there is unfortunately no social law stating that a respondent's time-varying characteristics from one wave will affect his or her answers to survey questions 2 years later. The issue of whether a particular lag is consistent with a theoretically plausible mechanism is one that must be addressed directly. As we show subsequently, getting this wrong can be misleading in the extreme.

Illustration

Reconsider the example from Table 1. As we mentioned earlier, even in the unlikely event that we have succeeded at purging our FE estimate of all time-variable heterogeneity we cannot know what share of the association is due to the effect of attendance on attitudes and what share is due to the effect of

Table 2. Unstandardized Coefficients From LFD Models.

	(1) Attitudes $t3 - t2$	(2) Attendance $t3 - t2$
Attendance $t2 - t1$	–.038 [–.093, .017]	
Attitudes $t2 - t1$		–.09 [–.17, –.013]
N	850	850

Note: LFD = lagged first difference. Values within brackets are 95% confidence intervals.

attitudes on attendance because they were measured at the same time. This problem is often resolved by assuming (usually tacitly) that the effects run in the direction hypothesized by the analyst. Instead of making this assumption, we might make this an empirical question by regressing the wave 2 to wave 3 change in each variable on the wave 1 to wave 2 change in the other, per equation (11). Let us ignore, for the moment, whether this lagged specification is reasonable and estimate the model.

The results are presented in Table 2. The estimated coefficients do nothing to solve the problem that motivated them. Both coefficients are *negative* and the confidence interval of the lagged effect of attitudes on attendance does not include 0. The FE and FD estimates from Table 1 and LFD models here both use the same basic technique for removing unobserved heterogeneity so the difference cannot lie there. The reason for these divergent results must be the different specifications of time.

Simulation

A reasonable question, then, is how the estimates of LFD models are affected by the actual temporal nature of the causal process. By using the lagged difference to predict a subsequent difference, LFD models assume that there is no contemporaneous effect of x on y. Researchers who assume that x and y are not (nearly) contemporaneously related generally do so because they want to use temporal ordering to identify the separate effects of y on x and x on y. Allison (2009:95) reports that LFD models do a good job recovering the correct parameter estimates in his simulations, but these simulations assume that the data were generated by a process with lags matching the spacing of the data collection. In the remainder of this section, we consider how robust LFD models are to violations of this assumption.

To make this question concrete, imagine two worlds defined by equations (12) and (13):

$$y_{it} = \beta x_{it} + \upsilon_i + \epsilon_{it}, \tag{12}$$

$$y_{it} = \beta x_{it-1} + \upsilon_i + \epsilon_{it}. \tag{13}$$

In the first world, y is a function of x at the same time point; the lagged value of x has no effect. In the second world, y is a function of x at the previous time point; the contemporaneous value has no effect. Now we can write an equation that allows for a continuous mixture of these two worlds by adding a new parameter, λ, which can vary from 0 to 1, and where β now represents the total effect of x through both contemporaneous and lagged effects.

$$y_{it} = (1 - \lambda)\beta x_{it} + \lambda\beta x_{it-1} + \upsilon_i + \epsilon_{it}. \tag{14}$$

When $\lambda = 0$, equation (14) is identical to equation (12). When $\lambda = 1$, equation (14) is identical to equation (13). When $\lambda = .5$, the contemporaneous and lagged values of x have the same effect on y. As a shorthand where useful, we refer to $\lambda\beta$ as β_{lag} and $(1 - \lambda)\beta$ as β_{con}.

We use a simulation to investigate how the LFD model performs under different mixtures of these causal worlds. Over 500 iterations, each with a "sample" size of 10,000, we allow λ to vary uniformly and allow β to take on three values, .25, .50, and .75. (See online appendix for full details.) We plot the results in Figure 6.

The pattern that appears in Figure 6 is truly astonishing, but it may take some explaining to see why. It is not surprising that when $\lambda = 1$ (i.e., when we are fully in "lag world"), the estimated coefficients are correct. In practice, the LFD estimate of β converges on its true value ($\beta_{\text{lag}} = \beta$) when the lags in our data match the causal lags that exist in the real world (as λ approaches 1). But as λ declines toward 0, the estimated value of β_{lag} does *not* decline toward 0 (which is the true effect of x_{t-1} on y_t when $\lambda = 0$), but rather toward $-\frac{1}{2}\beta$.[12]

Consider the implication: when x has a causal effect on y that is fully contemporaneous (when $\beta_{\text{con}} = \beta$), any lagged-$x$ FE model will yield a coefficient of *opposite sign* and *half the magnitude* of the true causal effect.[13] We discovered this property through simulation and find it useful to present it that way. But it turns out that, under some very general conditions, this property can be derived analytically (see online appendix for a proof).[14]

Because β_{LFD} declines to $-\frac{1}{2}\beta$ when $\lambda = 1$, in order to even hope for a null result, the effect of x_{t-1} on y_t must be at least half as large as the effect of x_t on y_t (i.e., $\lambda = 1/3$). In this case, the estimate of β_{LFD} will (within sampling error) be zero *regardless of the size of the total effect of x on y*. Given the two-year lag between panels in the GSS (and almost all panel data studies

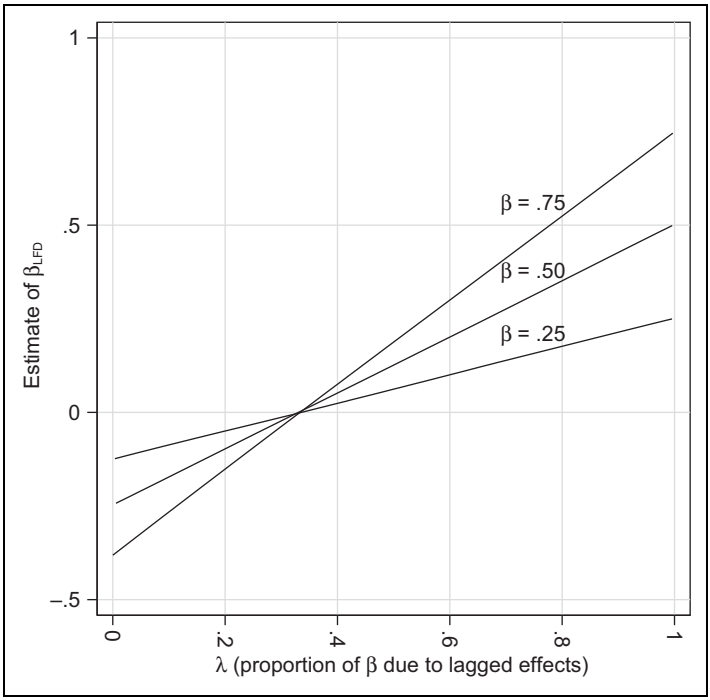

Figure 6. Simulation results for LFD models.

in sociology), hoping for lagged effects of even that magnitude for most processes is probably too sanguine, meaning that artifactual negative "effects" will likely be the rule rather than the exception.

This is a very surprising—even disturbing—finding. Recall from Table 2 that the LFD estimates of the reciprocal effects of attendance and opposition to abortion were negative. In light of the pattern demonstrated in Figure 2 and our substantive knowledge about religion and U.S. politics, the most reasonable conclusion is that these findings are artifactual. Consider which is more likely, that abortion and church attendance actually have negative reciprocal effects on each other or that 2 years is too long of a lag to establish causal ordering for this sort of process. Researchers should ask similar questions of articles that use such models to find null or "counterintuitive" negative effects.[15]

To reiterate, the problem with using lagged variables to establish temporal ordering is that the lags in our data rarely correspond to the lags present in real-world causal processes. Capturing short-term processes with widely

spaced data is best approximated by a cross-sectional approach, but this of course defeats the primary goal that motivated this discussion—the use of temporal ordering to determine causal direction among variables that are known to be associated.

This does not mean that the temporal ordering of the data is useless. For single events that are clearly located in time such as a divorce, job loss, or birth of a child, the ordering of the data can help determine causal effects (with the caveat that one must test the assumptions outlined above). But for continuously varying states (like attitudes and church attendance), relying on the temporal ordering of the data can be much worse than useless.

The Promise and Pitfalls of Panel Data

In recent years, sociological methodologists have encouraged practitioners to rely more on FE models for analyzing panel data because they are a powerful way of controlling for unobserved heterogeneity (e.g., Allison 2009; Halaby 2004). In the section on Panel Data and Unobserved Heterogeneity, we reviewed a variety of common panel models and provided demonstrations of the power of FE models to control for unobserved heterogeneity using both simulated data and GSS panel data.

But as powerful as they are, FE models are not infallible: they rely on two assumptions that are seldom tested and—if violated—can seriously bias estimates. We built on Morgan and Winship's (2007) discussion of these assumptions, extending their ideas to develop straightforward tests for treatment endogeneity and variable time trajectories in situations similar to those that might be encountered in the GSS panel. These tests are easy to estimate and should become standard practice for researchers who use FE models. If their assumptions are met, FE models can provide powerful protection against unmeasured influences. Otherwise, analysts should explore more flexible alternatives using SEMs (see Bollen and Brand 2010).

Finally, in our discussion of causal order, we demonstrated that FE models are extremely sensitive to the correct specification of time and therefore typically *cannot* be used straightforwardly to settle arguments about causal priority. Our simulations revealed that using lagged regressors in FE models can yield incorrect substantive conclusions when causal lags in the real world do not match the lags found in panel data. In extreme circumstances, estimates will be half the magnitude and in the opposite direction of the true parameter values.

At the risk of oversimplifying, we can summarize our article in three recommendations to users of the GSS panel and other three-wave panel data sets:

1. Use fixed-effects models with panel data to control for time-constant unobserved heterogeneity.
2. Test the assumptions of FE models about endogenous selection and temporal trajectories, and use alternative models if these assumptions are violated.
3. Do not rely on the ordering of the data to establish causal priority unless the lags between panels match the real-world causal lags in the processes under study.

As we stated in the introduction, the GSS panel provides new opportunities for social scientists to get more compelling and accurate answers to their research questions. Getting these answers, however, will require users to understand the tools that are at their disposal and apply them appropriately. FE models are a powerful and (we believe) underused method that leverages the power of panel data to provide protection against unobserved heterogeneity, but they are not a panacea. Ultimately, the confidence we place in our estimates must also rely theoretical justifications for the adequacy of our controls for time-varying confounders, and the extent to which our data accurately capture the real-world processes under study.

Acknowledgment

Thanks to Eric Bair for providing this proof.

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.

Supplemental Material

The online appendix is available at http://journals.sagepub.com/doi/suppl/10.1177/0049124114547769.

Notes

1. The first author acknowledges that he has not always followed the advice given here.
2. Unless otherwise specified, i always refers to respondents and $t = 1, 2, 3$.
3. Sometimes lagged values of y are included because the researcher posits an actual effect of Y_{t-1} on Y_1 (see Halaby 2004:536; Wooldridge 2010:371). Without also modeling unobserved heterogeneity, however, it is impossible to distinguish an autoregressive process like this from the existence of unobserved factors that affect both Y_{t-1} and Y_t.
4. Of course, any time we estimate a cross-sectional regression, we are making the same assumption. The difference with panel data is that we can test this assumption or even avoid it entirely.
5. The dummy variable method does not work properly for most limited dependent variables (see Allison 2009:16-18, 32-33).
6. Manually differencing will not, however, produce the correct standard errors. See Allison 2009:18).
7. The estimate of β can be somewhat different because the FD model allows adjacent errors to be correlated. See Wooldridge (2010:321-26) for more on the differences between FE and FD models.
8. If the RE assumption were justified, the RE and FE coefficients on attendance would be about the same. This can formally be tested with a Hausman test or other similar tests. See Allison (2009:21-23) and Cameron and Trivedi (2010:266-68).
9. Some of the ideas in this section emerged from informal discussions with Mike Hout (1999).
10. Here, following Morgan and Winship (2007:269), we specify time linearly. This assumption could easily be relaxed by using dummy variables for survey waves instead of linear time. Indeed, that is what we do in our empirical test.
11. This definition is similar to Granger (1969) causality. It is not a philosophical definition of causality, but a practical one.
12. For LFD models, the value is $-\frac{1}{2}\beta$, regardless of the number of waves. For FE models, the value is $-\frac{1}{T-1}\beta$, where t is the number of waves of data. We focus the discussion here on the three-wave case, where this distinction is not relevant.
13. The opposite holds as well, of course, if $\lambda = 1$, modeling y as a function of contemporaneous x in an FE or FD model will produce a negative artifact in the same manner. We don't investigate this issue any further here since this world would actually allow us to identify a consistent causal estimate through time ordering (which would be a good thing). Unfortunately, it is hard to imagine many processes in the general social survey (GSS) panel for which $\lambda \approx 1$ is likely to be the case.

14. The basic conditions are that the variance of x and y are the same at each time point and that x is not a state-dependent process over the time period of the panel (see Wooldridge 2010:371 for a clear discussion of state dependence). As x becomes fully state dependent, $\hat{\beta}_{LFD} \rightarrow \lambda\beta$. Our analyses of the GSS panel data (not shown) suggest that very few variables are even weakly state dependent over the term of the panel.
15. As we concluded writing this article, we came across an article by Ousey, Wilcox, and Fisher (2011) that uses a lagged predictor SEM model very close to the LFD model used here to estimate the reciprocal relationship between criminal offending and victimization. Despite a long history of research suggesting that these factors are positively related, their model indicates negative reciprocal effects using data with a 1-year lag. We are by no means willing to assert that their results are artifactual since we know little about the substantive issues involved. This does fit the pattern demonstrated here, however. To be fair, their article does draw on other research to outline a number of alternative theoretical mechanisms that could account for their findings. England, Allison, and Wu (2007) use the same model with lags ranging from 2 to 9 years. But since their data were on occupational aggregates, this assumption was probably more realistic since institutional change happens on a longer time scale than individual change. This is primarily a matter of theory and substantive knowledge.

References

Allison, Paul. 2009. *Fixed Effects Regression Models*. Thousand Oaks, CA: Sage.

Archer, Margaret S. 1995. *Realist Social Theory: The Morphogenetic Approach*. New York: Cambridge University Press.

Bollen, Kenneth A. 1989. *Structual Equations with Latent Variables*. New York: Wiley.

Bollen, Kenneth A. and Jennie E. Brand. 2010. "A General Panel Model with Random and Fixed Effects: A Structural Equations Approach." *Social Forces* 89:1-34.

Cameron, A. Colin and Pravin K. Trivedi. 2010. *Microeconometrics Using Stata*. Revised ed. College Station, TX: Stata Press.

Cha, Youngjoo. 2010. "Reinforcing Separate Spheres: The Effect of Spousal Overwork on Mens and Womens Employment in Dual-earner Households." *American Sociological Review* 75:303-29.

Elwert, Felix and Christopher Winship. 2014. "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable." *Annual Review of Sociology* 40:31-53.

Emirbayer, Mustafa and Ann Mische. 1998. "What Is Agency?" *The American Journal of Sociology* 103:962-1023.

England, Paula, Paul Allison, and Yuxiao Wu. 2007. "Does Bad Pay Cause Occupations to Feminize, Does Feminization Reduce Pay, and How Can We Tell with Longitudinal Data?" *Social Science Research* 36:1237-56.

Faris, Robert and Diane Felmlee. 2011. "Status Struggles, Network Centrality, and Gender Segregation in Same- and Cross-gender Aggression." *American Sociological Review* 76:48-73.

Granger, C. W. J. 1969. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods." *Econometrica* 37:424-38.

Halaby, Charles N. 2004. "Panel Models in Sociological Research: Theory into Practice." *Annual Review of Sociology* 30:507-44.

Hout, Michael. 1999. "Abortion Politics in the United States, 1972-1994: From Single Issue to Ideology." *Gender Issues* 17:3-34.

Lewis, Kevin, Marco Gonzalez, and Jason Kaufman. 2012. "Social Selection and Peer Influence in an Online Social Network." *Proceedings of the National Academy of Sciences* 109:68-72.

Lizardo, Omar. 2006. "How Cultural Tastes Shape Personal Networks." *American Sociological Review* 71:778-807.

Martin, John Levi, Tod Van Gunten, and Benjamin D. Zablocki. 2012. "Charisma, Status, and Gender in Groups With and Without Gurus." *Journal for the Scientific Study of Religion* 51:20-41.

Morgan, Stephen L. and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 1st ed. Cambridge, UK: Cambridge University Press.

Ousey, Graham, Pamela Wilcox, and Bonnie Fisher. 2011. "Something Old, Something New: Revisiting Competing Hypotheses of the Victimization-offending Relationship Among Adolescents." *Journal of Quantitative Criminology* 27:53-84.

Vaisey, Stephen and Omar Lizardo. 2010. "Can Cultural Worldviews Influence Network Composition?" *Social Forces* 88:1595-618.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: The MIT Press.

Author Biographies

Stephen Vaisey is an associate professor of sociology at Duke University. The main goal of his research is to understand the structure, origins, and consequences of different moral and political worldviews.

Andrew Miles is an assistant professor at the University of Toronto. He was a doctoral candidate in sociology at Duke University. His work examines how perspectives from social psychology and cultural sociology can be synthesized to create more accurate models of human action, focusing particularly on values, identities, and dual-process cognition.