Article

Sociological Methods & Research
2023, Vol. 52(1) 43–84

DOI: 10.1177/0049124120926211
journals.sagepub.com/home/smr


Fixed Effects Individual Slopes: Accounting and Testing for Heterogeneous Effects in Panel Data or Other Multilevel Models

Tobias Rüttenauer[1] and Volker Ludwig[2]

Abstract

Fixed effects (FE) panel models have been used extensively in the past, as those models control for all stable heterogeneity between units. Still, the conventional FE estimator relies on the assumption of parallel trends between treated and untreated groups. It returns biased results in the presence of heterogeneous slopes or growth curves that are related to the parameter of interest (e.g., selection into treatment is based on individual growth of the outcome). In this study, we derive the bias in conventional FE models and show that fixed effects individual slope (FEIS) models can overcome this problem. This is a more general version of the conventional FE model, which accounts for heterogeneous slopes or trends, thereby providing a powerful tool for panel data and other multilevel data in general. We propose two versions of the Hausman test that can be used to identify misspecification in FE models. The performance of the FEIS estimator and the specification tests is evaluated in a series of Monte Carlo experiments. Using the examples of the marital wage premium and returns to preschool

[1] Nuffield College, University of Oxford, United Kingdom
[2] Department of Social Sciences, Technische Universität Kaiserslautern, Germany

Corresponding Author:
Tobias Rüttenauer, Nuffield College, University of Oxford, Oxford OX1 1NF, United Kingdom.
Email: tobias.ruttenauer@nuffield.ox.ac.uk

education (Head Start), we demonstrate how taking heterogeneous effects into account can seriously change the conclusions drawn from conventional FE models. Thus, we propose to test for bias in FE models in practical applications and to apply FEIS if indicated by the specification tests.

Keywords

fixed effects, fixed effects individual slopes, Hausman test, Head Start returns, heterogeneous effects, marital wage premium, Monte Carlo simulations

Fixed effects (FE) panel estimators are nowadays extensively used in social sciences (e.g., Best and Rüttenauer 2018; Kühhirt 2012; Lichter, Parisi, and Taquino 2015; Massoglia, Firebaugh, and Warner 2013; Noelke 2016; Qvist and Munk 2018; Reardon and Bischoff 2011; Torche and Villarreal 2014) and economics (e.g., Banzhaf and Walsh 2008; Chetty and Hendren 2018; Figlio et al. 2014; Frijters and Beatton 2012; Henderson et al. 2018; Krueger and Mueller 2012; Rivkin, Hanushek, and Kain 2005). The main advantage of the FE panel estimator is that this type of model controls for time-constant heterogeneity—no matter whether actually observed or unobserved (Allison 2009; Angrist and Pischke 2009; Baltagi 2013; Firebaugh, Warner, and Massoglia 2014; Wooldridge 2010). Consequently, stable unobserved differences between units do not bias the estimates of interest, thereby reducing the likelihood of spurious results due to omitted variable bias.

However, it has been noted earlier that the conventional FE estimator actually constitutes a special case of a more general type of model: the FE model with individual-specific slopes (i.e., fixed effects individual slopes [FEIS]; Brüderl and Ludwig 2015; Frees 2001; Lemieux 1998; Polachek and Kim 1994; Wooldridge 2010). Compared to the conventional FE model, the generalized model has the advantage of controlling for heterogeneous slopes in addition to time-constant heterogeneity. Thus, FEIS models relax the assumption of parallel slopes in "treatment" and "control" observations, an assumption that might be violated in many applications. For instance, using FEIS estimators, V. Ludwig and Brüderl (2018) have recently shown that the "male marital wage premium", that is, the earnings benefit men receive "because" they get married—documented by several studies using FE estimation (e.g., Dougherty 2006; Killewald and Gough 2013)—is actually not the result of a causal effect of marriage on earnings. Rather, men on a steeper

wage trajectory are more likely to marry. Thus, even if the married men had never married, they would have had a stronger increase in earnings over the life course. Because the parallel slopes assumption is not met, conventional FE returns a spurious marriage premium.

Given the advantage of the FEIS model over the more specific FE model, it is surprising that the FEIS has not gained much attention in applied research using individual panel data or otherwise nested data with large numbers of groups (large N, small T). In fact, while controlling for country-specific time trends is standard for FE analyses with macro-level cross-sectional time-series data (e.g., Kneip and Bauer 2009; Noelke 2016), researchers working with other hierarchical data structures do not seem to be fully aware of the potential bias of FE estimates. Furthermore, there is little guidance for practitioners on how to test for a bias due to heterogeneous slopes in concrete applications.

It has been argued that the FEIS model has not received much attention with individual panel data, as it requires relatively long panels (Morgan and Winship 2015), precisely $T > J$, where J is the number of specified slope parameters. Furthermore, the relatively few existing studies for large N individual panel data controlling for heterogeneous slopes are usually based on the random trend model, another special case of FEIS that incorporates unit-specific time trends only (DiPrete and McManus 2000; Heckman and Hotz 1989; Jacobson, LaLonde, and Sullivan 2005; Loughran and Zissimopoulos 2009; Pischke 2001). This random trend model is usually estimated by second differencing (SD) or by applying FE to first differences of the data (FE-FD). Thereby, the estimation sample is often reduced tremendously due to item nonresponse or partial unit nonresponse, and the model becomes even more "data hungry" than FEIS. Moreover, SD and FE-FD account only for heterogeneous time trends, while FEIS allows to control for individual slopes based on any observed variable, making it a more general model (Wooldridge 2010). Given the progress in software, data collection and the availability of long-run panels, FEIS can be usefully applied in various settings.

In this article, we clarify analytically which terms go into the bias of standard FE models. In addition, we propose two versions of the well-known Hausman test (Hausman 1978) that can be used for model selection. By simulation, we pinpoint situations in which FE suffers from a bias and provide evidence on the performance of the statistical tests. Recent software packages *feisr* and *xtfeis* implement the FEIS model and specification tests with standard statistical software (`R` and `Stata`). We illustrate the application of the model and tests with two empirical examples. By applying FEIS to the analysis of the Head Start preschool program, we also demonstrate the

usefulness of the methods for multilevel data in general. Finally, we discuss potential limitations and conclude with some guidelines for practical research.

FEIS Estimator

The conventional FE estimator (Allison 2009; Angrist and Pischke 2009; Baltagi 2013; Firebaugh et al. 2014; Wooldridge 2010) is well-known and often used in economic and social science research. We assume a data generating process (DGP) of the following form:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}, \tag{1}$$

or, in matrix notation:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{i}_T\alpha_i + \boldsymbol{\varepsilon}_i, \tag{2}$$

where $\boldsymbol{y}_i$ is $T \times 1$, $\boldsymbol{X}_i$ is $T \times K$, $\boldsymbol{i}_T$ is $T \times 1$, α_i is a scalar, $\boldsymbol{\varepsilon}_i$ is $T \times 1$, and $\boldsymbol{\beta}$ a $K \times 1$ parameter vector. α_i denotes a person-specific time-constant unobserved characteristic that influences the outcome variable $\boldsymbol{y}_i$. In this case, consistency of pooled Ordinary Least Squares (OLS) and random effects (RE) estimators hinges on the exogeneity assumption that $\mathrm{Cov}(\boldsymbol{x}_{it}, \alpha_i) = 0$. In other words, unobserved heterogeneity between groups or individuals may not be correlated with $\boldsymbol{x}_{it}$. FE models relax this assumption by explicitly controlling for the unit-specific effects α_i.

Therefore, the system of equations in equation (2) is estimated by OLS on the within-transformed data

$$\boldsymbol{y}_i - \bar{\boldsymbol{y}}_i = (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_i)\boldsymbol{\beta} + \boldsymbol{i}_T\alpha_i - \boldsymbol{i}_T\alpha_i + \boldsymbol{\varepsilon}_i - \bar{\boldsymbol{\varepsilon}}_i, \tag{3}$$

$$\ddot{\boldsymbol{y}}_i = \ddot{\boldsymbol{X}}_i\boldsymbol{\beta} + \ddot{\boldsymbol{\varepsilon}}_i, \tag{4}$$

where $\ddot{\boldsymbol{y}}_i$ is $T \times 1$, $\ddot{\boldsymbol{X}}_i$ is $T \times K$, and $\ddot{\boldsymbol{\varepsilon}}_i$ is $T \times 1$. $\ddot{\boldsymbol{y}}_i, \ddot{\boldsymbol{X}}_i, \ddot{\boldsymbol{\varepsilon}}_i$ are the time-"demeaned" data in which the unit-specific means of each variable over time are subtracted. This eliminates all time-constant heterogeneity from the equation, including the error component α_i. Hence, the FE estimator can be written as

$$\hat{\boldsymbol{\beta}}_{FE} = \left(\sum_{i=1}^{N} \ddot{\boldsymbol{X}}_i^{\top} \ddot{\boldsymbol{X}}_i\right)^{-1} \left(\sum_{i=1}^{N} \ddot{\boldsymbol{X}}_i^{\top} \ddot{\boldsymbol{y}}_i\right). \tag{5}$$

This conventional FE estimator offers the advantage of consistent estimates of $\boldsymbol{\beta}$ even if the exogeneity assumption needed for pooled OLS and RE models does not hold.

However, when using conventional FE estimation, we have to assume strict exogeneity of the idiosyncratic error term (Wooldridge 2010). Suppose a second set of variables $\boldsymbol{W}$ impacts the outcome, so $\text{Cov}(\ddot{\boldsymbol{y}}, \ddot{\boldsymbol{W}}) \neq 0$. Some of these variables may also be correlated with the variable(s) of interest $\text{Cov}(\ddot{\boldsymbol{X}}, \ddot{\boldsymbol{W}}) \neq \mathbf{0}$. For instance, $\boldsymbol{W}$ might include time trends for $\boldsymbol{y}$ and $\boldsymbol{X}$. In this case, strict exogeneity does not hold and FE returns biased point estimates for $\boldsymbol{X}$. The obvious way to avoid biased estimates due to omission of $\boldsymbol{W}$ is to include $\boldsymbol{W}$ as a second set of covariates in equation (4) and estimate the model specification:

$$\ddot{\boldsymbol{y}}_i = \ddot{\boldsymbol{X}}_i\boldsymbol{\beta} + \ddot{\boldsymbol{W}}_i\boldsymbol{\gamma} + \ddot{\boldsymbol{\varepsilon}}_i. \tag{6}$$

Still, controlling for $\boldsymbol{W}_i$ in equation (6) does not warrant consistency of the FE estimator: FE may be biased due to heterogeneous slopes on $\boldsymbol{W}_i$.

Bias in FE due to Omitted Slope Heterogeneity

In this section, we will derive the bias of the conventional FE estimator analytically. For simplicity, we use only a single covariate $\boldsymbol{x}$ and one individual slope variable $\boldsymbol{w}$. Assume the true DGP is

$$\boldsymbol{y}_i = \boldsymbol{x}_i\beta + \boldsymbol{i}_T\alpha_{i1} + \boldsymbol{w}_i\alpha_{i2} + \boldsymbol{\varepsilon}_i, \tag{7}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ are $T \times 1$, and α_{i1} and α_{i2} are individual-specific scalars. In full matrix notation, we can write the DGP as

$$\boldsymbol{y} = \boldsymbol{x}\beta + \boldsymbol{i}_{NT} \circ \boldsymbol{\alpha}_1 + \boldsymbol{w} \circ \boldsymbol{\alpha}_2 + \boldsymbol{\varepsilon}, \tag{8}$$

where $\boldsymbol{x}$ and $\boldsymbol{w}$ are $NT \times 1$ vectors, $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are stacked $NT \times 1$ vectors of subvectors $\boldsymbol{i}_T\alpha_{i1}$ and $\boldsymbol{i}_T\alpha_{i2}$, and $\circ$ denotes the Hadamard (element-wise) product. Now suppose, we would erroneously assume the DGP was

$$\boldsymbol{y}_i = \boldsymbol{x}_i\beta + \boldsymbol{i}_T\alpha_{i1} + \boldsymbol{w}_i\gamma + \boldsymbol{\varepsilon}_i, \tag{9}$$

where the scalar γ (note the missing subscript) denotes a constant effect of $\boldsymbol{w}_i$ on $\boldsymbol{y}_i$ over all units i. In full matrix notation, we can write this as

$$\boldsymbol{y} = \boldsymbol{x}\beta + \boldsymbol{i}_{NT} \circ \boldsymbol{\alpha}_1 + \boldsymbol{w}\gamma + \boldsymbol{\varepsilon}. \tag{10}$$

This represents a DGP of a model with homogeneous effects of $\boldsymbol{w}_i$ on $\boldsymbol{y}_i$ across all units (no individual slopes), which equals the DGP assumed in case of a conventional FE model with additional control variables as given by equation (6).

Why would the FE estimator be biased? As mentioned above, FE requires parallel slopes. Otherwise strict exogeneity does not hold. Define $\alpha_{i2} = \bar{\alpha}_2 + \ddot{\alpha}_{2i}$, where $\bar{\alpha}_2$ constitutes the overall mean of α_{2i}, for example, a common trend in the population, and $\ddot{\alpha}_{2i}$ are individual-specific deviations from the mean. The parallel slopes condition would be $E(\alpha_{2i}|\boldsymbol{x}_i, \boldsymbol{w}_i) = \bar{\alpha}_2$. In the case with $\boldsymbol{w}_i$ being calendar time, this means that individual outcome trajectories may depend on unobservables α_{2i}, as long as the average slope in the population does not differ by values of covariates, notably $\boldsymbol{x}_i$. With the definition of α_{2i} above, we can state the parallel slopes condition also as $E(\ddot{\alpha}_{2i}|\boldsymbol{x}_i, \boldsymbol{w}_i) = 0$, which implies $E(\boldsymbol{x}_i\ddot{\alpha}_{2i}) = 0$. That is, individual deviations from the common trend may not be correlated with the causal variable.

Thus, estimating a DGP like equation (9) by FE as in equation (6), we receive

$$\ddot{\boldsymbol{y}}_i = \ddot{\boldsymbol{x}}_i\beta + \ddot{\boldsymbol{w}}_i\bar{\alpha}_2 + \ddot{\boldsymbol{\xi}}_i, \tag{11}$$

with $\ddot{\boldsymbol{\xi}}_i = \ddot{\boldsymbol{w}}_i\ddot{\alpha}_{2i} + \ddot{\boldsymbol{\varepsilon}}_i$. Since individual deviations from the common trend are contained in the idiosyncratic error term of the model and we estimate the demeaned equation (11) by OLS, a necessary condition for strict exogeneity, that is, for $E(\xi_{it}|\boldsymbol{x}_i, \boldsymbol{w}_i, \alpha_{i1}) = 0$, is $E\left(\ddot{\boldsymbol{x}}_{it}\ddot{\xi}_{it}\right) = E\left[(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)(\xi_{it} - \bar{\xi}_i)\right] = 0$. Hence, even if ε_{it} is strictly exogeneous in the true model (7), strict exogeneity may be violated by FE estimation of a model that specifies a homogeneous effect for $\boldsymbol{w}_i$. It must be violated if the parallel slopes assumption does not hold.

To show the bias, we apply the conventional FE estimator controlling for $\boldsymbol{w}$ to the DGP specified in equation (8). We use the Frisch–Waugh theorem (Frisch and Waugh 1933; Lovell 1963) and define $\mathbf{M} = \mathbf{I} - \ddot{\boldsymbol{w}}(\ddot{\boldsymbol{w}}^\top\ddot{\boldsymbol{w}})^{-1}\ddot{\boldsymbol{w}}^\top$ to receive an estimate of the parameter of interest:

$$\begin{aligned}\hat{\beta}_{FE} &= \left(\ddot{\boldsymbol{x}}^\top\boldsymbol{M}\ddot{\boldsymbol{x}}\right)^{-1}\ddot{\boldsymbol{x}}^\top\boldsymbol{M}\ddot{\boldsymbol{y}} \\ &= \left(\ddot{\boldsymbol{x}}^\top\boldsymbol{M}\ddot{\boldsymbol{x}}\right)^{-1}\ddot{\boldsymbol{x}}^\top\boldsymbol{M}(\ddot{\boldsymbol{x}}\beta + \ddot{\boldsymbol{i}}_{NT} \circ \alpha_1 + \ddot{\boldsymbol{w}} \circ \boldsymbol{\alpha}_2 + \ddot{\boldsymbol{\varepsilon}}).\end{aligned} \tag{12}$$

As $\ddot{\boldsymbol{i}}_{NT} = \boldsymbol{i}_{NT} - \bar{\ddot{\boldsymbol{i}}}_{NT} = 0$, we can eliminate the time-constant part of the unobserved heterogeneity. Moreover, by assuming strict exogeneity $\mathrm{E}(\ddot{\boldsymbol{x}}^\top\ddot{\boldsymbol{\varepsilon}}) = 0$ conditional on $\ddot{\boldsymbol{x}}$, $\ddot{\boldsymbol{w}}$, $\boldsymbol{\alpha}_1$, and $\boldsymbol{\alpha}_2$, equation (12) further simplifies to:

$$\hat{\beta}_{FE} = \left(\ddot{\boldsymbol{x}}^\top\boldsymbol{M}\ddot{\boldsymbol{x}}\right)^{-1}\ddot{\boldsymbol{x}}^\top\boldsymbol{M}(\ddot{\boldsymbol{x}}\beta + \ddot{\boldsymbol{w}} \circ \boldsymbol{\alpha}_2). \tag{13}$$

As above, we split up $\boldsymbol{\alpha}_2 = \bar{\alpha}_2 + \ddot{\boldsymbol{\alpha}}_2$, where $\bar{\alpha}_2$ constitutes the overall mean of $\boldsymbol{\alpha}_2$, and $\ddot{\boldsymbol{\alpha}}_2$ the individual-specific deviation from the mean. Then, we can rewrite equation (13) as

$$\begin{aligned}\hat{\beta}_{FE} &= \left(\ddot{\boldsymbol{x}}^\top \boldsymbol{M} \ddot{\boldsymbol{x}}\right)^{-1} \ddot{\boldsymbol{x}}^\top \boldsymbol{M}(\ddot{\boldsymbol{x}}\beta + \ddot{\boldsymbol{w}}\bar{\alpha}_2 + \ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\alpha}}_2) \\ &= \beta + \left(\ddot{\boldsymbol{x}}^\top \boldsymbol{M} \ddot{\boldsymbol{x}}\right)^{-1} \ddot{\boldsymbol{x}}^\top \boldsymbol{M}(\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\alpha}}_2) \\ &= \beta + \\ &\quad \left[\ddot{\boldsymbol{x}}^\top \ddot{\boldsymbol{x}} - \ddot{\boldsymbol{x}}^\top \ddot{\boldsymbol{w}}\left(\ddot{\boldsymbol{w}}^\top \ddot{\boldsymbol{w}}\right)^{-1} \ddot{\boldsymbol{w}}^\top \ddot{\boldsymbol{x}}\right]^{-1} \\ &\quad \left[\ddot{\boldsymbol{x}}^\top (\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\alpha}}_2) - \ddot{\boldsymbol{x}}^\top \ddot{\boldsymbol{w}}\left(\ddot{\boldsymbol{w}}^\top \ddot{\boldsymbol{w}}\right)^{-1} \ddot{\boldsymbol{w}}^\top (\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\alpha}}_2)\right].\end{aligned} \tag{14}$$

Now suppose, we specify a relationship between $\boldsymbol{x}$ and $\boldsymbol{w}$ of the form:

$$\begin{aligned}\boldsymbol{x} &= \boldsymbol{w} \circ \boldsymbol{\delta} + \boldsymbol{v} \\ &= \boldsymbol{w}\bar{\delta} + \boldsymbol{w} \circ \ddot{\boldsymbol{\delta}} + \boldsymbol{v},\end{aligned} \tag{15}$$

where $\boldsymbol{\delta}$ is a stacked $NT \times 1$ vector of the subvector $\boldsymbol{i}_T \delta_i$, $\bar{\delta}$ is the overall mean, $\ddot{\boldsymbol{\delta}}$ is the deviation from the mean, and $\boldsymbol{v}$ is an independent random vector. Note that δ_i specifies how strongly the slope variable $\boldsymbol{w}$ is associated with our explanatory variable $\boldsymbol{x}$. As for α_{2i}, we may view δ_i as unobserved effect.

If we substitute for $\boldsymbol{x}$ in equation (14), we receive:

$$\begin{aligned}\hat{\beta}_{FE} = \beta \; &+ [(\ddot{\boldsymbol{w}}\bar{\delta} + \ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\delta}} + \ddot{\boldsymbol{v}})^\top (\ddot{\boldsymbol{w}}\bar{\delta} + \ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\delta}} + \ddot{\boldsymbol{v}}) \\ &- (\ddot{\boldsymbol{w}}\bar{\delta} + \ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\delta}} + \ddot{\boldsymbol{v}})^\top \ddot{\boldsymbol{w}}(\ddot{\boldsymbol{w}}^\top \ddot{\boldsymbol{w}})^{-1} \ddot{\boldsymbol{w}}^\top (\ddot{\boldsymbol{w}}\bar{\delta} + \ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\delta}} + \ddot{\boldsymbol{v}})]^{-1} \\ &[(\ddot{\boldsymbol{w}}\bar{\delta} + \ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\delta}} + \ddot{\boldsymbol{v}})^\top (\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\alpha}}_2) \\ &- (\ddot{\boldsymbol{w}}\bar{\delta} + \ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\delta}} + \ddot{\boldsymbol{v}})^\top \ddot{\boldsymbol{w}}(\ddot{\boldsymbol{w}}^\top \ddot{\boldsymbol{w}})^{-1} \ddot{\boldsymbol{w}}^\top (\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\alpha}}_2)] \\ = \beta + &\frac{(\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\delta}})^\top \boldsymbol{M}(\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\alpha}}_2)}{(\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\delta}})^\top \boldsymbol{M}(\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\delta}}) + \ddot{\boldsymbol{v}}^\top \ddot{\boldsymbol{v}}}.\end{aligned} \tag{16}$$

Recall that $\boldsymbol{M}(\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\alpha}}_2)$ equals the residual vector from regressing $(\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\alpha}}_2)$ on $\ddot{\boldsymbol{w}}$, and the expected value from the regression $(\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\alpha}}_2) = \ddot{\boldsymbol{w}}\lambda + \boldsymbol{\varepsilon}$ is given by $\mathrm{E}(\hat{\lambda}) = \bar{\ddot{\boldsymbol{\alpha}}}_2 = 0$, as $\ddot{\boldsymbol{\alpha}}_2$ is the vector of the overall-demeaned $\boldsymbol{\alpha}_2$ values. It follows that the residual vector $\boldsymbol{M}(\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\alpha}}_2) = (\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\alpha}}_2) - \ddot{\boldsymbol{w}}\mathrm{E}(\hat{\lambda}) = (\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\alpha}}_2)$. The same argument applies to the first part of the denominator of equation (16), and we can thus simplify equation (16) to:

$$\begin{aligned} \mathrm{E}(\hat{\beta}_{FE}) &= \beta + \frac{\mathrm{E}[(\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\delta}})^{\top} (\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\alpha}}_2)]}{\mathrm{E}[(\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\delta}})^{\top} (\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\delta}})] + \mathrm{E}[\ddot{\boldsymbol{v}}^{\top} \ddot{\boldsymbol{v}}]} \\ &= \beta + \frac{\mathrm{Var}(\ddot{\boldsymbol{w}})\mathrm{Cov}(\ddot{\boldsymbol{\delta}}, \ddot{\boldsymbol{\alpha}}_2)}{\mathrm{Var}(\ddot{\boldsymbol{w}})\mathrm{Var}(\ddot{\boldsymbol{\delta}}) + \mathrm{Var}(\ddot{\boldsymbol{v}})}. \end{aligned} \tag{17}$$

Equation (17) demonstrates that the conventional FE model produces biased point estimates if $\mathrm{Cov}(\ddot{\boldsymbol{\delta}}, \ddot{\boldsymbol{\alpha}}_2) \neq 0$. Intuitively, this means that FE is biased if the same individuals who experience a stronger influence of $\boldsymbol{w}_i$ on the dependent variable $\boldsymbol{y}_i$ also experience a stronger influence of $\boldsymbol{w}_i$ on the independent variable $\boldsymbol{x}_i$. For instance, imagine that those individuals who have steeper wage earnings over the life course are also those who invest more in occupational training over the life cycle. Using a conventional FE model, we would conclude that additional training increases earnings, though that is due to the fact that more motivated people participate in more training and experience steeper earning trajectories—even without any further training (age is the slope variable). As noted earlier, we may view both $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}_2$ as unobserved variables. Obviously, if some unobservable (e.g., motivation) drives both $\boldsymbol{w}_i$ and $\boldsymbol{x}_i$, the covariance of $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}_2$ will not be zero. Similar conclusions apply with dichotomous treatment variables: V. Ludwig and Brüderl (2018) demonstrate that those individuals who have a higher wage growth tend to marry at higher rates (work experience is the slope variable), which in the conventional FE model shows up in a spurious "marriage wage premium." However, the slope variables are not restricted to timing or life-course variables but can also be some other characteristic in general multilevel scenarios. For instance, some families might have lower abilities to support their children's cognitive development, thereby having a higher correlation between children's starting abilities and later test score outcomes. At the same time, those families might be more selective in choosing children with lower starting abilities to enroll in preschool programs, which means a higher negative correlation between children's starting abilities and the probability of preschool enrolment (e.g., Behrman and Rosenzweig 2004; Griliches 1979). FE models would then underestimate the effect of preschool enrolment on test score outcomes, as the models suffer from selection on starting abilities (starting ability is the slope variable).

FEIS Estimator

As the previous section shows, several quite realistic situations exist in which the conventional FE estimator might yield biased estimates of the true effects. To deal with the problem, several authors (Brüderl and Ludwig 2015; Frees 2001; Lemieux 1998; Polachek and Kim 1994; Wooldridge 2010) have proposed a generalized FE model allowing for unit-specific slopes: the FEIS estimator.

The DGP for the FEIS estimator is given by:

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \alpha_{i1} + w_{it}\alpha_{i2} + \varepsilon_{it}, \tag{18}$$

or, respectively:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{W}_i\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i, \tag{19}$$

where $\boldsymbol{y}_i$ is $T \times 1$, $\boldsymbol{X}_i$ is $T \times K$, and $\boldsymbol{\varepsilon}_i$ is $T \times 1$. $\boldsymbol{W}_i$ is a $T \times J$ matrix of slope variables, and $\boldsymbol{\alpha}_i$ a $J \times 1$ vector of individual-specific slope parameters, for J slope parameters including a constant term. If $\boldsymbol{W}_i$ includes a constant term, $\boldsymbol{\alpha}_i$ contains the α_{i1} known from the conventional FE estimator. Thus, the conventional FE estimator constitutes a specific case of the FEIS estimator where $\boldsymbol{W}_i$ contains only a constant term, or $\boldsymbol{W}_i = \mathbf{i}_T$, as this would reduce equation (19) to (2). Note that estimating equation (19) with $\boldsymbol{W}_i$ containing a constant plus additional slope variables would involve the estimation of N individual regressions of $\boldsymbol{y}_i$ on $\boldsymbol{W}_i$ or a least square dummy variable model with NJ fixed effects. However, as with the conventional FE, the same can be obtained by running OLS on transformed data. We specify the "residual maker" matrix $\boldsymbol{M}_i = \boldsymbol{I}_T - \boldsymbol{W}_i(\boldsymbol{W}_i^\top \boldsymbol{W}_i)^{-1}\boldsymbol{W}_i^\top$ and apply OLS to the transformed data:

$$y_{it} - \hat{y}_{it} = (\boldsymbol{x}_{it} - \hat{\boldsymbol{x}}_{it})\boldsymbol{\beta} + \varepsilon_{it} - \hat{\varepsilon}_{it}, \tag{20}$$

$$\boldsymbol{M}_i\boldsymbol{y}_i = \boldsymbol{M}_i\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{M}_i\boldsymbol{\varepsilon}_i, \tag{21}$$

$$\tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{X}}_i\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}_i, \tag{22}$$

where $\tilde{\boldsymbol{y}}_i$, $\tilde{\boldsymbol{X}}_i$, and $\tilde{\boldsymbol{\varepsilon}}_i$ are the residuals of regressing $\boldsymbol{y}_i$, each column-vector of $\boldsymbol{X}_i$, and $\boldsymbol{\varepsilon}_i$ on $\boldsymbol{W}_i$. In full matrix notation, the model can be written as

$$\tilde{\boldsymbol{M}}\boldsymbol{y} = \tilde{\boldsymbol{M}}\boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{M}}\boldsymbol{\varepsilon}, \tag{23}$$

where we define $\tilde{\boldsymbol{M}} = \boldsymbol{I}_N \otimes \boldsymbol{M}_i$, where $\otimes$ is the Kronecker (block-wise) product and $\boldsymbol{M}_i = \boldsymbol{I}_T - \boldsymbol{W}_i\left(\boldsymbol{W}_i^\top \boldsymbol{W}_i\right)^{-1}\boldsymbol{W}_i^\top$. $\tilde{\boldsymbol{M}}$ constitutes a stacked

block-diagonal matrix with block-diagonal elements equal to $\boldsymbol{M}_i$ and off-block-diagonal elements equal to zero. The resulting estimator is given by:

$$\hat{\boldsymbol{\beta}}_{\text{FEIS}} = \left(\sum_{i=1}^{N} \tilde{\boldsymbol{X}}_i^{\top} \tilde{\boldsymbol{X}}_i \right)^{-1} \left(\sum_{i=1}^{N} \tilde{\boldsymbol{X}}_i^{\top} \tilde{\boldsymbol{y}}_i \right). \tag{24}$$

To see why premultiplying by $\tilde{\boldsymbol{M}}$ in equation (23) eliminates the bias in equation (16), we will now only consider the last term of equation (16), which is decisive for the bias: $\boldsymbol{M}(\ddot{\boldsymbol{w}} \circ \ddot{\boldsymbol{\alpha}}_2)$. From estimating equation (23) instead of (12), we would receive $\tilde{\boldsymbol{M}}(\boldsymbol{\alpha}_1 + \boldsymbol{w} \circ \boldsymbol{\alpha}_2)$. Now recall that $\boldsymbol{W}$ always contains a constant intercept plus the slope variable(s). So for each group i, $\tilde{\boldsymbol{M}}(\boldsymbol{\alpha}_1 + \boldsymbol{w} \circ \boldsymbol{\alpha}_2)$ contains the residual vector of the following regression:

$$(\alpha_{i1} + \boldsymbol{w}_i \alpha_{i2}) = \boldsymbol{i}_T \lambda_{i1} + \boldsymbol{w}_i \lambda_{i2} + \boldsymbol{\varepsilon}_i, \tag{25}$$

where we get $\mathrm{E}(\hat{\lambda}_{i1}) = \alpha_{i1}$ and $\mathrm{E}(\hat{\lambda}_{i2}) = \alpha_{i2}$, which is the individual-specific mean and the individual-specific slope parameter. Thus, the stacked $NT \times 1$ vector $\hat{\lambda}_1$ of subvectors $\boldsymbol{i}_T \hat{\lambda}_{i1}$ equals the stacked vector $\boldsymbol{\alpha}_1$, and the stacked $NT \times 1$ vector $\hat{\boldsymbol{\lambda}}_2$ equals the stacked vector $\boldsymbol{\alpha}_2$. Consequently, $\tilde{M}(\boldsymbol{\alpha}_1 + \boldsymbol{w} \circ \boldsymbol{\alpha}_2) = (\boldsymbol{\alpha}_1 + \boldsymbol{w} \circ \ddot{\boldsymbol{\alpha}}_2) - \left(\hat{\boldsymbol{\lambda}}_1 + \boldsymbol{w} \circ \hat{\boldsymbol{\lambda}}_2 \right) = 0$, which eliminates the bias in equation (16) and provides a consistent estimator of β.

As shown previously, the conventional FE estimator requires the strict exogeneity assumption $\mathrm{E}(\varepsilon_{it} | \boldsymbol{x}_i, \boldsymbol{w}_i, \alpha_{i1}) = 0$, that is, conditional on $\boldsymbol{w}_i$ and the individual constant α_{i1}, to hold for consistency. Note that α_{i1} is a single scalar parameter for each group. In contrast, the FEIS estimator requires the strict exogeneity assumption of the form $\mathrm{E}(\varepsilon_{it} | \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\alpha}_i) = 0$, where $\boldsymbol{\alpha}_i = \alpha_{i1}, \ldots, \alpha_{iJ}$, that is, conditional on $\boldsymbol{w}_i$ and all J individual slopes of the vector $\boldsymbol{\alpha}_i$ (including the constant α_{i1}), to hold. Thus, consistency of FEIS relies on weaker assumptions than consistency of conventional FE models. Still, for consistent estimation of $\hat{\boldsymbol{\beta}}_{\text{FEIS}}$, we need to impose the rank condition $\operatorname{rank} \mathrm{E}\left(\tilde{\boldsymbol{X}}_i^{\top} \tilde{\boldsymbol{X}}_i \right) = K$, where $\tilde{\boldsymbol{X}}_i = \boldsymbol{M}_i \boldsymbol{X}_i$ and $\operatorname{rank}(\boldsymbol{M}) = T - J$ for K covariates, T time periods, and J slope parameters (including the constant). Thus, we need $T > J$, that is, the number of time periods to be larger than the number of specified slope parameters. While the conventional FE requires $T \geq 2$, the FEIS with one slope parameter (additional to the constant) requires $T \geq 3$, which in unbalanced data may lead to a problem of sample selection (for a discussion, see Limitations section). However, with large N, we should still be able to obtain a consistent estimate of $\hat{\boldsymbol{\beta}}_{\text{FEIS}}$ if some

observations have rank$\left(\tilde{\boldsymbol{X}}_i^\top \tilde{\boldsymbol{X}}_i\right) < K$ because of lacking within-variance on single covariates (Wooldridge 2010:379).

Specification Tests

As the FEIS estimator requires weaker assumptions for consistency, but is less efficient than the conventional FE estimator, it seems desirable to rely on a test statistic that tells us whether we should use the FEIS or the conventional FE estimator. When we need to choose between a RE and a conventional FE specification, it is common to use the Hausman test for a bias due to unobserved heterogeneity (Hausman 1978). In panel data settings, the Hausman test compares two estimators, one of which is known be consistent (unbiased as N approaches infinity) under the alternative (Hypothesis 1) that unobserved heterogeneity is related to the covariate(s). In the standard application of the test, we know that unobserved time-constant heterogeneity α_{i1} would bias RE, but not FE. Similarly, we showed above that omitted slope heterogeneity α_{i2} biases FE, but not FEIS. In this section, we demonstrate how we can extend the logic of the conventional Hausman test to the comparison of FE and FEIS models, and test whether heterogeneous slopes α_{i2} are omitted from the specification and bias FE.

The general Hausman test statistic is given by:

$$H = \left(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0\right)^\top \left(N^{-1}\boldsymbol{V}_{\hat{\boldsymbol{\beta}}_1-\hat{\boldsymbol{\beta}}_0}\right)^{-1}\left(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0\right), \tag{26}$$

where $\hat{\boldsymbol{\beta}}_1$ denotes the consistent parameter estimates and $\hat{\boldsymbol{\beta}}_0$ the efficient estimates, which are only consistent under H_0 but not under H_1, and $N^{-1}\boldsymbol{V}_{\hat{\boldsymbol{\beta}}_1-\hat{\boldsymbol{\beta}}_0} = \mathrm{Var}\left(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0\right)$. Hausman (1978) shows that we can simplify the test statistic by replacing $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}_1-\hat{\boldsymbol{\beta}}_0}$ in equation (26) by $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}_1-\hat{\boldsymbol{\beta}}_0} = \mathrm{Var}(\hat{\boldsymbol{\beta}}_1) - \mathrm{Var}(\hat{\boldsymbol{\beta}}_0)$ if it holds that $\hat{\boldsymbol{\beta}}_0$ is the fully efficient estimator (see also Wooldridge 2010:328-34). Unfortunately, when comparing the FEIS estimator against the FE estimator, we cannot simplify the test statistic as above, because the FE estimator is not fully efficient, and thus we need to take $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_0)$ into account. In this section, we thus suggest two alternative ways of deriving the Hausman test statistic. First, we introduce an artificial regression test (ART) considering the covariance by estimating both coefficient vectors in a common model. Second, we propose a bootstrapped Hausman test (BSHT), which derives the empirical covariance between the

coefficients by simulation. Both of them can be used to compare RE, FE, and FEIS models, and both are implemented in the packages *feisr* and *xtfeis* for estimation in `R` and `Stata`.

Hausman-like ART

As has been shown by Mundlak (1978) and Chamberlain (1982), we can also derive the parameter estimates of the conventional FE model by estimating a correlated random effects (CRE) model. The CRE is given by:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \bar{\boldsymbol{X}}_i\boldsymbol{\gamma} + \boldsymbol{W}_i\boldsymbol{\delta} + \bar{\boldsymbol{W}}_i\boldsymbol{\theta} + \boldsymbol{\varepsilon}_i, \tag{27}$$

where $\bar{\boldsymbol{X}}_i$ and $\bar{\boldsymbol{W}}_i$ contain the individual-specific means of the variables $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$. If we estimate the CRE specification by generalized least squares (GLS), the parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ represent estimates of the within effects—identical to the FE coefficients—and the vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ give us an estimate of the between effect minus the within effect (e.g., Allison 2009; Arellano 1993; Wooldridge 2010). Consistency of the RE estimator is based on the assumption that $\boldsymbol{\gamma} = \boldsymbol{\theta} = \mathbf{0}$. Thus, we can perform a Wald test on the hypothesis $\mathrm{H}_0 : \boldsymbol{\gamma} = \boldsymbol{\theta} = \mathbf{0}$ to obtain a regression-based version of the Hausman test statistic.

An advantage of the ART is that we can easily extend it to test for consistency of the FE estimator or more precisely for comparing FE against FEIS estimates. Therefore, we need to compute the individual-specific predicted values $\hat{\boldsymbol{X}}_i = \boldsymbol{W}_i(\boldsymbol{W}_i^\top \boldsymbol{W}_i)^{-1}\boldsymbol{W}_i^\top \boldsymbol{X}_i$ based on the slope variables in $\boldsymbol{W}_i$ and extend the CRE in equation (27) to:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \bar{\boldsymbol{X}}_i\boldsymbol{\gamma} + \hat{\boldsymbol{X}}_i\boldsymbol{\rho} + \boldsymbol{W}_i\boldsymbol{\delta} + \bar{\boldsymbol{W}}_i\boldsymbol{\theta} + \boldsymbol{\varepsilon}_i. \tag{28}$$

Analogue to the discussion above, the parameter vector $\boldsymbol{\beta}$ of the extended CRE is identical to the parameters estimated by FEIS. Consistency of the FE estimator is based on the assumption that $\boldsymbol{\rho} = \mathbf{0}$. Thus, we can perform a Wald test of H_0: $\boldsymbol{\rho} = \mathbf{0}$ (the FE estimator is consistent). If the Null is not rejected, $\boldsymbol{\rho} = \mathbf{0}$ and we can omit the term $\hat{\boldsymbol{X}}_i\boldsymbol{\rho}$, making the conventional FE estimator more efficient than the FEIS estimator. Note that the terms $\bar{\boldsymbol{X}}_i\boldsymbol{\gamma}$, $\bar{\boldsymbol{W}}_i\boldsymbol{\theta}$, and $\boldsymbol{W}_i\boldsymbol{\delta}$ are actually not necessary in equation (28) to obtain the FEIS coefficients in $\hat{\boldsymbol{\beta}}$ ($\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\mathrm{FEIS}}$ if we include $\hat{\boldsymbol{X}}_i$ as additional covariates). However, they are necessary if we want to compare FEIS against FE, as equation (28) only reduces to the FE estimator conditional on $\boldsymbol{\rho} = \mathbf{0}$ if we include the individual-specific means as in equation (27).

BSHT

The test discussed in the previous section comes with the advantage that we can specify standard errors which are robust to arbitrary forms of heteroscedasticity and serial correlation (e.g., Arellano 1987, 1993). This is not possible with the standard Hausman test, but it is necessary if the errors are correlated within clusters (units) and thus observations are not independent of each other. However, computation of cluster-robust standard errors is based on the assumption that the number of clusters approaches infinity (Stock and Watson 2008). Cameron, Gelbach, and Miller (2008) show by simulation that relying on cluster-robust standard errors can lead to downwardly biased standard errors and an overrejection of H_0 when the number of clusters is small (N cluster < 30).

To overcome this problem, Cameron et al. (2008) and Cameron and Miller (2015) proposed a BSHT. The idea is to perform pairwise-clustered bootstrapping by randomly selecting $N^\star$ clusters with replacement from the original sample of N clusters and to repeat this over R replications. Estimating FEIS and FE models in each replication provides R estimates of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_0$, which can be used to estimate $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}_1-\hat{\boldsymbol{\beta}}_0}$. While the ART takes the variance–covariance of the two estimators into account by assuming an identical error variance for FEIS and FE, the BSHT relies on an estimate of the variance–covariance. The bootstrapped estimate of the variance–covariance matrix is given by

$$\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}_1-\hat{\boldsymbol{\beta}}_0} = \frac{1}{R-1}\sum_{b=1}^{R}\left[\left(\hat{\boldsymbol{\beta}}^{\star}_{b,1}-\hat{\boldsymbol{\beta}}^{\star}_{b,0}\right)-\left(\bar{\boldsymbol{\beta}}^{\star}_{b,1}-\bar{\boldsymbol{\beta}}^{\star}_{b,0}\right)\right]\left[\left(\hat{\boldsymbol{\beta}}^{\star}_{b,1}-\hat{\boldsymbol{\beta}}^{\star}_{b,0}\right)-\left(\bar{\boldsymbol{\beta}}^{\star}_{b,1}-\bar{\boldsymbol{\beta}}^{\star}_{b,0}\right)\right]^{\top}, \tag{29}$$

where $\hat{\boldsymbol{\beta}}^{\star}_{b,1}$ and $\hat{\boldsymbol{\beta}}^{\star}_{b,0}$ are the estimated coefficients for the consistent ($\hat{\boldsymbol{\beta}}_1$) and the efficient ($\hat{\boldsymbol{\beta}}_0$) model returned by single bootstrap replications, and $\bar{\boldsymbol{\beta}}^{\star}_{b,1}$ and $\bar{\boldsymbol{\beta}}^{\star}_{b,0}$ denote the mean values of the coefficients over all R replications. The resulting variance–covariance matrix can then be used to compute the Hausman test statistic of equation (26). This is equal to the method called "pairs cluster bootstrap-se" in the simulations by Cameron et al. (2008). Both methods presented in this section can be used to test for misspecification in conventional FE models.

Monte Carlo Simulations

In the following section, we illustrate the heterogeneity bias in conventional FE and RE models as compared to the more general FEIS model. Moreover, we assess the size and power of the ART and the BSHT as described in

Specification Tests section. Therefore, we employ Monte Carlo experiments using the packages *feisr* and *plm* (Croissant and Millo 2008).

DGP and Simulation Settings

In the simulations, we set up balanced panel data sets with $N = 300$ units and $T = 10$ time points, leading to a total of 3,000 observations. The data are generated by the following DGP:

$$\boldsymbol{y} = \boldsymbol{x}\beta + \boldsymbol{i}_{NT} \circ \boldsymbol{\alpha}_1 + \boldsymbol{w} \circ \boldsymbol{\alpha}_2 + \boldsymbol{\varepsilon}, \tag{30}$$

$$\boldsymbol{x} = \theta \boldsymbol{i}_{NT} \circ \boldsymbol{\alpha}_1 + \boldsymbol{w} \circ \boldsymbol{\delta} + \boldsymbol{v}, \tag{31}$$

where $\boldsymbol{\varepsilon}$ and $\boldsymbol{v}$ are independent Gaussian random vectors $\boldsymbol{\varepsilon}, \mathbf{v} \sim N(0, 1)$. The scalar parameter θ specifies the correlation between $\boldsymbol{x}$ and the time-constant "unobserved" heterogeneity $\boldsymbol{\alpha}_1$, and the $NT \times 1$ parameter vector $\boldsymbol{\delta}$ the correlation between $\boldsymbol{x}$ and the slope variable $\boldsymbol{w}$.

As described earlier, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, and $\boldsymbol{\delta}$ are stacked vectors of the individual-specific subvectors $\boldsymbol{i}_T\alpha_{i1}$, $\boldsymbol{i}_T\alpha_{i2}$, and $\boldsymbol{i}_T\delta_i$. Further, we define $\boldsymbol{\alpha}_1$ as a normally distributed variable with mean μ_{α_1} and standard deviation σ_{α_1}. The values of random variable $\boldsymbol{w}$ are drawn from a normal distribution within each unit i, with mean μ_w and standard deviation σ_w. Finally, we draw $\boldsymbol{\alpha}_2$ and $\boldsymbol{\delta}$ from a bivariate normal distribution,

$$\begin{pmatrix} \boldsymbol{\alpha}_2 \\ \boldsymbol{\delta} \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_{\alpha_2} \\ \mu_\delta \end{pmatrix}, \quad \begin{pmatrix} \sigma_{\alpha_2} & \phi \\ \phi & \sigma_\delta \end{pmatrix} \right],$$

with N draws, which are then replicated T times for each unit i to get time-constant vectors $\boldsymbol{\alpha}_2$ and $\boldsymbol{\delta}$.[1]

As shown earlier (see equation [17]), the theoretical bias of the standard FE estimator has four components: the covariance of $\ddot{\boldsymbol{\delta}}$ and $\ddot{\boldsymbol{\alpha}}$, and the variances of $\ddot{\boldsymbol{\delta}}$, $\ddot{\boldsymbol{w}}$ and $\ddot{\mathbf{v}}$. In the simulations, we experiment with these components in that we vary parameters $\phi := \mathrm{Cov}(\boldsymbol{\delta}, \boldsymbol{\alpha}_2)$, $\sigma^2_\delta := \mathrm{Var}(\boldsymbol{\delta})$, $\sigma^2_w := \mathrm{Var}(\mathbf{w})$, and $\sigma^2_{\mathbf{v}} := \mathrm{Var}(\mathbf{v})$ systematically. For all our simulations, we set $\beta = 1$. In the following, we present two sets of simulation results.

In setting (1), we vary ϕ between -0.8 and 0.8 and set $\sigma^2_{\boldsymbol{\delta}}$, $\sigma^2_{\boldsymbol{w}}$, and $\sigma^2_{\mathbf{v}}$ to fixed values. We are mainly interested in the performance of the FE and FEIS estimators, but we also include the conventional RE estimator because it is the main alternative to the FE model in applied research using panel data. The RE estimator is known to be biased if $\boldsymbol{\alpha}_1$ is related to $\boldsymbol{x}$. In the initial setup, we therefore vary also $\theta \in \{0, 1\}$.

For each value of ϕ, we estimate FEIS, FE, and RE regression models of $\boldsymbol{y}$ on $\boldsymbol{x}$ and $\boldsymbol{w}$ with cluster-robust estimation of the covariance matrix. Similarly, the ART is performed using the robust error variance proposed by Wallace and Hussain (1969). We compute the mean bias of $\hat{\beta}$ for each of the three estimators using $R = 1,000$ replications. Since true β remains fixed at 1 throughout, the mean of the absolute bias is equal to the mean relative bias. So, we will interpret the average bias as a percentage. We use $R_b = 100$ replications to estimate the bootstrapped covariance matrix for the BSHT. To evaluate the size and power of the ART and BSHT in detecting biased estimates, we report rejection rates for the 95 percent confidence level, that is, the proportion of test statistics that yield a p value $< .05$.

In setting (2), we focus on the bias of the FE estimator and the performance of the ART and BSHT with respect to the bias. We set $\theta = 0.5$ and expand the initial setting in that we vary all four parameters that go into the bias of FE. Specifically, we set

- $\phi \in \{0, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.65, 0.8\}$,
- $\sigma^2_\delta \in \{0.2, 0.5, 1, 3, 5\}$,
- $\sigma^2_w \in \{0.2, 1, 5\}$,
- $\sigma^2_v \in \{0.2, 1, 5\}$.

For each of the 360 resulting parameter combinations, we report the mean bias of the FE estimate and the rejection rates of the ART and BSHT comparing FE and FEIS coefficients over $R = 1,000$ replications. This allows us to asses (1) the severity of the bias in FE estimates and (2) the sensitivity of our proposed test statistics across different situations.

Simulation Results

Setting (1). We first look at simulations for varying ϕ with the other three parameters fixed ($\sigma_\delta = 1$, $\sigma_w = 2$, and $\sigma_v = 1$). Figure 1 shows results depending on whether or not unobservables $\boldsymbol{\alpha}_1$ are related to the causal variable $\boldsymbol{x}$. In both cases, we set $\mu_{\boldsymbol{\alpha}_1} = 1$ and $\sigma_{\alpha_1} = 2$, but we vary $\theta \in \{0, 1\}$.

In the first case ($\theta = 0$), $\boldsymbol{\alpha}_1$ does not bias the effect of $\boldsymbol{x}$. So, we focus only on the bias of $\hat{\beta}$ due to $\boldsymbol{\alpha}_2$. As shown in Figure 1A, the conventional FE and RE models return biased estimates of β whenever $\phi \neq 0$. Since there is no bias due to $\boldsymbol{\alpha}_1$, RE and FE estimates are almost identical on average. We can see that the strength of the FE bias increases with ϕ and that ϕ determines the sign of the bias. All this follows directly from equation (17). Also note that the magnitude of the mean bias is very close to the theoretical bias (as calculated with equation 17). In Figure 1B and C, we show rejection rates

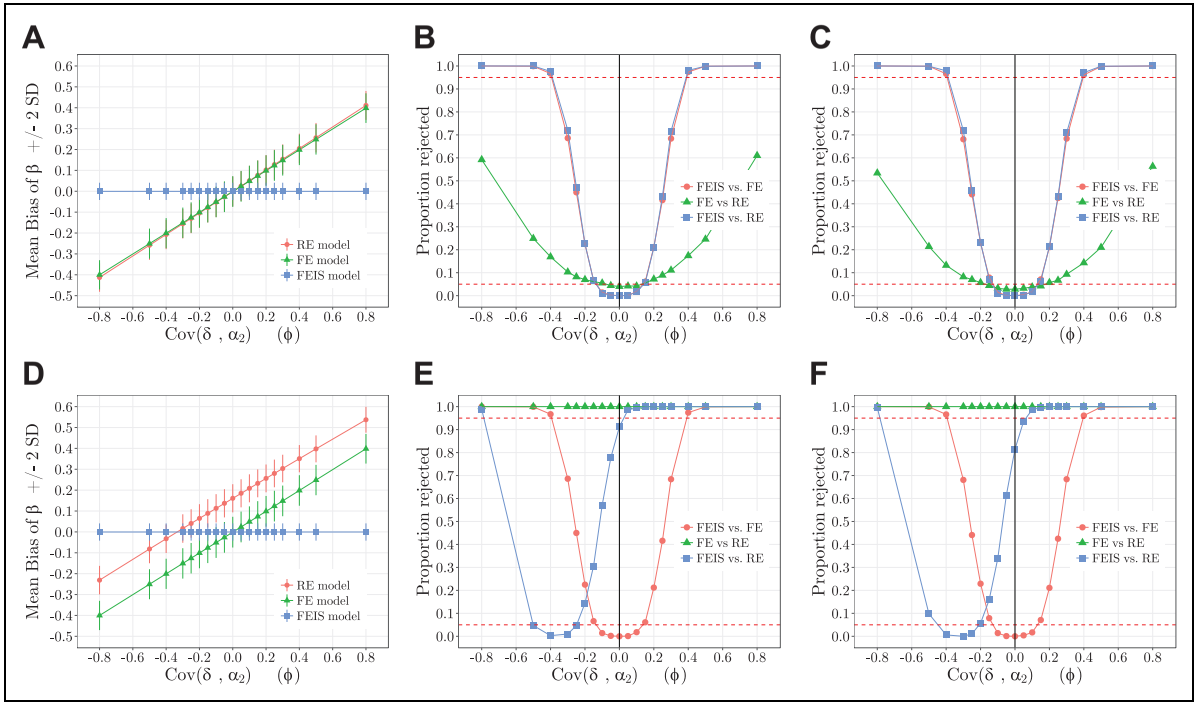

Figure 1. Simulated bias of random effects (RE), fixed effects (FE), and fixed effects individual slope (FEIS) estimators and rejection rates of Artificial Regression Test (B, E) and Bootstrapped Hausman Test (C, F) tests for $\theta = 0$ (A–C) and $\theta = 1$ (D–F). θ = effect of time-constant unobserved heterogeneity α_1 on covariate $\boldsymbol{x}$; $\phi = \mathrm{Cov}(\delta, \alpha_2)$; $\boldsymbol{\delta}$ = effect of slope variable $\boldsymbol{w}$ on covariate $\boldsymbol{x}$; α_2 = effect of slope variable $\boldsymbol{w}$ on outcome $\boldsymbol{y}$; and $N = 300$, $T = 10$, $R = 1,000$, and $R_b = 100$.

of ARTs and BHSTs for the simulated coefficients. For $\phi = 0$, FE is unbiased and each test procedure should seldomly reject the H_0 hypothesis that FEIS and FE return identical estimates. The empirical rejection rate indeed is smaller than 0.01, meaning that we falsely reject the FE model for less than 10 of 1,000 simulated samples. Thus, both tests for the FEIS versus the FE model have excellent size. Moreover, the tests of the FEIS versus the FE model (ART and BSHT) exhibit better size than the test of the FE versus the RE model (size of about 5 percent). Note that the ART without robust standard errors exhibits a much lower size (not shown), leading to an increase in Type I error. Thus, we recommend to use robust standard errors along with the ART test.

The tests also exhibit reasonable power to detect a bias due to $\boldsymbol{\alpha}_2$: The rejection rate of the ART and the BSHT is at about 97 percent for $\phi = 0.4$, where the bias is at 20 percent, and the rate is at 100 percent for $\phi \geq 0.45$ (bias at 22.5 percent). It is important to note that the tests of the FE versus the RE model do not have power to detect a bias due to heterogeneous slopes. Even for $\phi = \pm 0.8$, implying a bias of 40 percent, rejection rates are low (between 50 and 60 percent). For applied research, this means that relying on the standard Hausman test for model selection is a tedious strategy. If the parallel slopes assumption is violated, we run into a serious problem. If the bias due to $\boldsymbol{\alpha}_2$ is strong, we very likely reject the RE model and base our conclusions on the biased FE model. If the bias is less strong, but still substantial, we may report biased RE results. These results give a strong motivation for using the ART or BSHT of FEIS versus FE (or RE coefficients), which provides a more reliable approach of model selection. Note, however, that—similar to the conventional Hausman test not having power to detect slope heterogeneity—our proposed tests do not have power to detect other sources of bias, like functional misspecification or sample selection (see Limitations section).

Figure 1D–F show simulation results for varying ϕ when $\boldsymbol{\alpha}_1$ is related to $\boldsymbol{x}$ ($\theta = 1$). In this situation, the RE estimator suffers from an additional source of bias. Whereas the FE estimator is still consistent for $\phi = 0$, the RE estimator is biased even in this case (see Figure 1D). Whenever $\phi \neq 0$, we get biased results for the FE estimator. Since the FE bias is independent of $\boldsymbol{\alpha}_1$, we simply get the same FE results as in Figure 1A. Interestingly, for the RE estimator, there are situations in which the two biases due to $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ have opposite sign and are of the same (absolute) size, thereby offsetting each other. In such a setting, the FE estimator is biased, but the RE estimator returns the correct estimate on average. In Figure 1D, this is the case for $\phi = -0.35$. Again, relying on a test of the FE versus the RE coefficients

would in many cases lead to erroneous choice of statistical models. The ART and the BSHT for FE versus RE always reject H_0 without systematic difference in coefficients (see Figure 1E and F), even in those situations in which the RE estimator is (by coincidence) close to the true β. The tests for FEIS against FE or RE are more reliable: They have excellent size and good power to detect biased estimates due to $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$.

Setting (2). In the second set of experiments, we focus on the bias of the FE estimator. We expand the setting in that we vary all four parameters that go into the bias of FE. Specifically, we set $\phi \in \{0, ..., 0.8\}$, $\sigma^2_\delta \in \{0.2, 0.5, 1, 3, 5\}$, $\sigma^2_w \in \{0.2, 1, 5\}$, and $\sigma^2_v \in \{0.2, 1, 5\}$. We report the mean of the bias of the FE estimate over 1,000 replications for each of the 360 resulting scenarios. Secondly, we compare the performance of the ART and the BSHT. Figure 2 shows how the bias of the FE estimator depends on ϕ, σ^2_δ, σ^2_w, and σ^2_v. Each of the nine graphs shows a plot of the bias against ϕ for five different values of σ^2_δ for one particular combination of values for σ^2_w and σ^2_v. Note that the *y*-axes of the single graphs are scaled differently to improve visibility.

Figure 2 illustrates several points. First, the average FE bias is always (close to) 0 if $\phi = 0$, and it increases linearly with ϕ, holding the other parameters constant. Second, increasing ϕ and decreasing σ^2_δ together magnify the bias: the influence of ϕ increases with lower values of σ^2_δ (and vice versa). Intuitively, correlated slopes (ϕ) produce a stronger bias if the causal variable $\boldsymbol{x}$ is affected more similarly by the slope variable $\boldsymbol{w}$ over all units (low variance in $\boldsymbol{\delta}$). Third, the strength of the interaction of ϕ and σ^2_δ depends on the relative size of σ^2_w (variance in the slope variable) and σ^2_v (independent variance in $\boldsymbol{x}$). If $\sigma^2_w << \sigma^2_v$, the impact of ϕ on the bias hardly depends on the value of σ^2_δ (Figure 2C). If $\sigma^2_w >> \sigma^2_{\boldsymbol{v}}$, the impact of ϕ varies greatly with σ^2_δ (Figure 2G). The more relative variance in $\boldsymbol{x}$ comes from independent random noise $\boldsymbol{v}$, the less important is the heterogeneity in slopes. Moreover, the magnitude of the bias varies enormously between these settings. The maximum average bias is very small (only about 3 percent) in the first case ($\sigma^2_w << \sigma^2_v$), but extremely large in case of $\sigma^2_{\boldsymbol{w}} >> \sigma^2_v$ (more than 300 percent). Obviously, the bias increases with an increasing proportion of the variance in $\boldsymbol{x}$ stemming from the variance in the slope variable. For scenarios with identical values for σ^2_w and σ^2_v (Figure 2A, E, and I on the diagonal, from upper left to lower right), we find a very similar pattern for the bias, with nearly identical maximum bias (around 65 percent for large ϕ and small $\sigma^2_{\boldsymbol{\delta}}$).

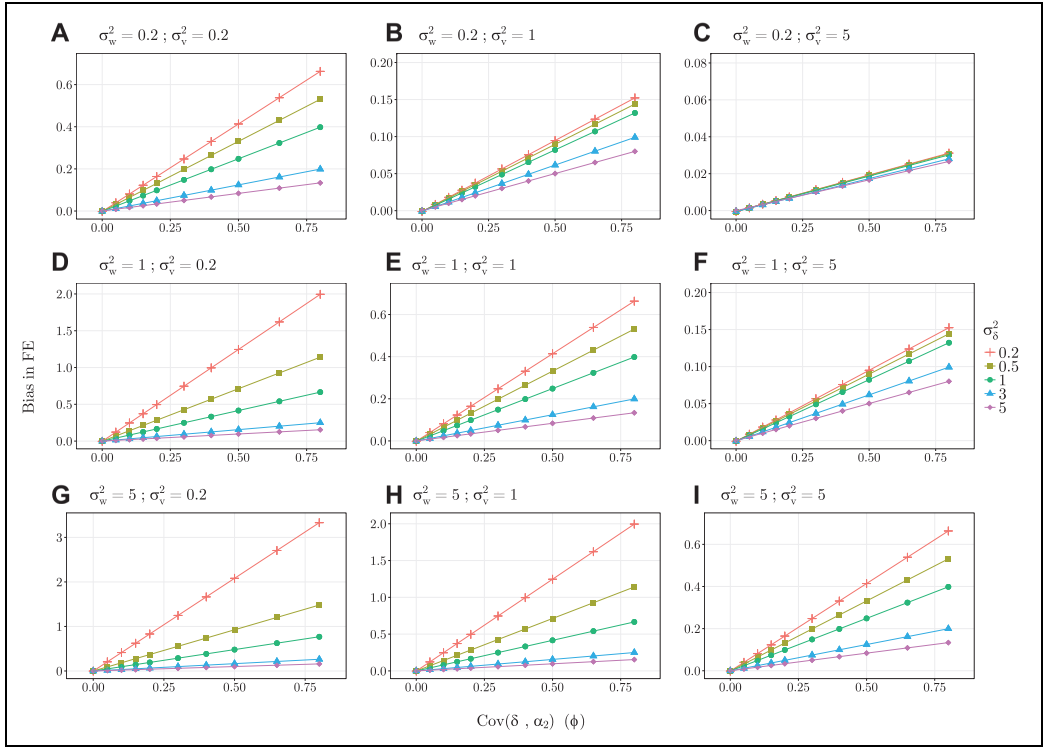

Figure 2. Simulated bias of fixed effects (FE) estimator for varying parameters. $\phi = \text{Cov}(\delta, \alpha_2)$; $\boldsymbol{\delta}$ = effect of slope variable ***w*** on covariate ***x***; $\boldsymbol{\alpha}_2$ = effect of slope variable ***w*** on outcome ***y***; $\boldsymbol{\nu}$ = independent random vector in covariate ***x***; σ^2 = variance of $\boldsymbol{\delta}$, ***w***, and $\boldsymbol{\nu}$, respectively; and $N = 300$, $T = 10$, $R = 1,000$, and $R_b = 100$. (A) $\sigma_w^2 = 0.2$; $\sigma_v^2 = 0.2$. (B) $\sigma_w^2 = 0.2$; $\sigma_v^2 = 1$. (C) $\sigma_w^2 = 0.2$; $\sigma_v^2 = 5$. (D) $\sigma_w^2 = 1$; $\sigma_v^2 = 0.2$. (E) $\sigma_w^2 = 1$; $\sigma_v^2 = 1$. (F) $\sigma_w^2 = 1$; $\sigma_v^2 = 5$; (G) $\sigma_w^2 = 5$; $\sigma_v^2 = 0.2$. (H) $\sigma_w^2 = 5$; $\sigma_v^2 = 1$. (I) $\sigma_w^2 = 5$; $\sigma_v^2 = 5$.

Taken together, these simulations indicate that if there is substantial correlation (ϕ) between individual slopes of variable $\boldsymbol{w}$ (i.e., $\boldsymbol{\alpha}_2$) and the effect of this variable on the causal variable $\boldsymbol{x}$ (i.e., $\boldsymbol{\delta}$), the standard FE model returns substantially biased estimates for the causal variable under almost all scenarios. Furthermore, the bias is magnified in most settings if there is little heterogeneity of $\boldsymbol{\delta}$ in the population. There is one notable exception to these rules: If the variance of the slope variable is very low and the causal variable is very noisy (i.e., $\sigma_w^2 << \sigma_v^2$), the FE bias tends to be small regardless of other parameters. In practice, however, we will often have strong (within) variation of the slope variable as, for instance, in applications with heterogeneous growth over time or age. In such typical applications, correlation of $\boldsymbol{\alpha}_2$ and $\boldsymbol{\delta}$ almost certainly leads to seriously biased FE estimators.

We now turn to the performance of the proposed tests. Figures 3 and 4 show rejection rates at 95 percent significance level for the ART and BSHT, respectively, as applied to the simulated FE coefficients. For each parameter combination of σ_w^2 and σ_v^2 in the simulations, we plot the rates against the mean bias of FE. Hence, we show nine graphs for each test. As explained above, the maximum bias varies strongly between the nine settings. Therefore, the x-axes of the graphs differ. Although ϕ is not shown here, we already know that the bias is a linear function of ϕ, and that the bias is zero for $\phi = 0$. Within each graph, therefore ϕ varies from 0 to 0.8 on the x-axis as the bias increases from 0 to the largest value.

Turning to the results for the ART (Figure 3), we see that the test has very good size. The likelihood to commit a Type I error is never higher than 5 percent. Hence, if the true FE bias is zero, we would seldomly reject the Null of consistent FE estimates. Furthermore, the test provides reasonable levels of statistical power to detect a bias of FE estimates for most scenarios. In most situations, if $\sigma_w^2 \leq \sigma_v^2$ rejection rates are higher than 95 percent for a bias of 20 percent or less. If $\sigma_w^2 >> \sigma_v^2$, however, the power of the ART can be insufficiently low (Figure 3G). Hence, in these situations, we might choose the FE model even though it is heavily biased. The likelihood to commit such a type II error is higher when σ_δ^2 is small. Overall, these results show that the proposed ART for FEIS versus FE estimates has very good size and reasonable power, except in the particular situation just described.

Figure 4 depicts results of the second alternative for testing FEIS against FE models: The BSHT shows very similar performance compared to the ART. In fact, the results hardly differ at all. Thus, in practice, it should make little difference whether we use the ART or the BSHT, at least if N is large (as in our simulations). For a small number of clusters (N cluster < 30), in

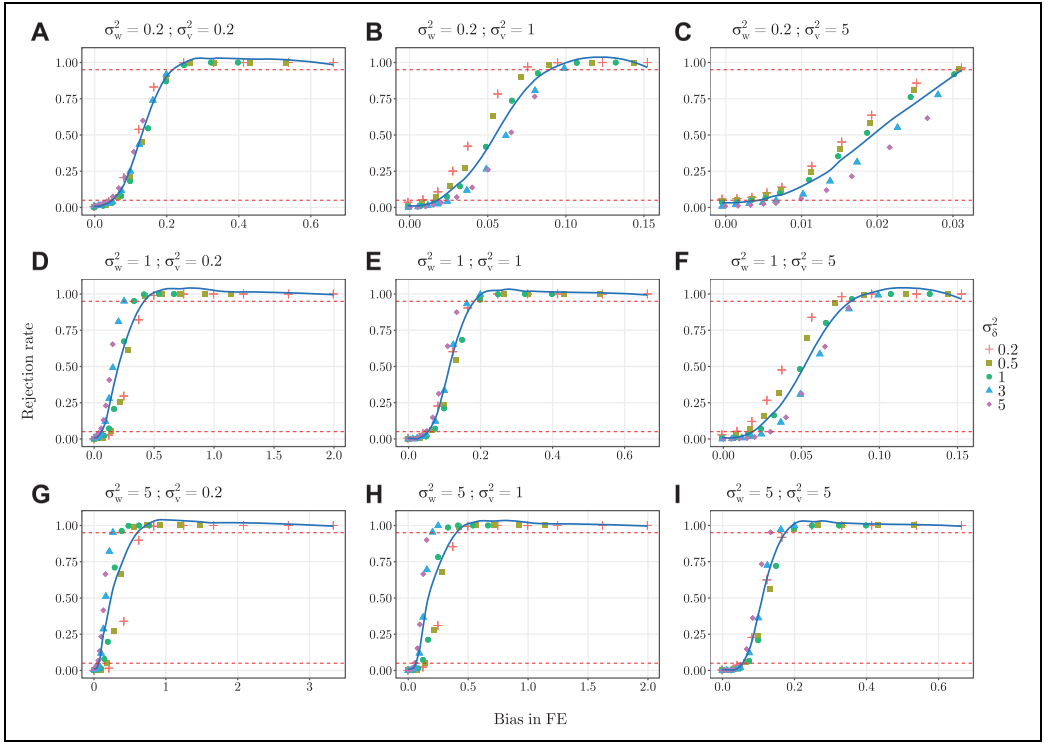

Figure 3. Rejection rates of Artificial Regression Test with robust standard errors. δ = effect of slope variable $\boldsymbol{w}$ on covariate $\boldsymbol{x}$; α_2 = effect of slope variable $\boldsymbol{w}$ on outcome $\boldsymbol{y}$; ν = independent random vector in covariate $\boldsymbol{x}$; σ^2 = variance of δ, $\boldsymbol{w}$, and ν, respectively; and $N = 300$, $T = 10$, $R = 1,000$, and $R_b = 100$. (A) $\sigma_w^2 = 0.2$; $\sigma_v^2 = 0.2$. (B) $\sigma_w^2 = 0.2$; $\sigma_v^2 = 1$. (C) $\sigma_w^2 = 0.2$; $\sigma_v^2 = 5$. (D) $\sigma_w^2 = 1$; $\sigma_v^2 = 0.2$. (E) $\sigma_w^2 = 1$; $\sigma_v^2 = 1$. (F) $\sigma_w^2 = 1$; $\sigma_v^2 = 5$; (G) $\sigma_w^2 = 5$; $\sigma_v^2 = 0.2$. (H) $\sigma_w^2 = 5$; $\sigma_v^2 = 1$. (I) $\sigma_w^2 = 5$; $\sigma_v^2 = 5$.

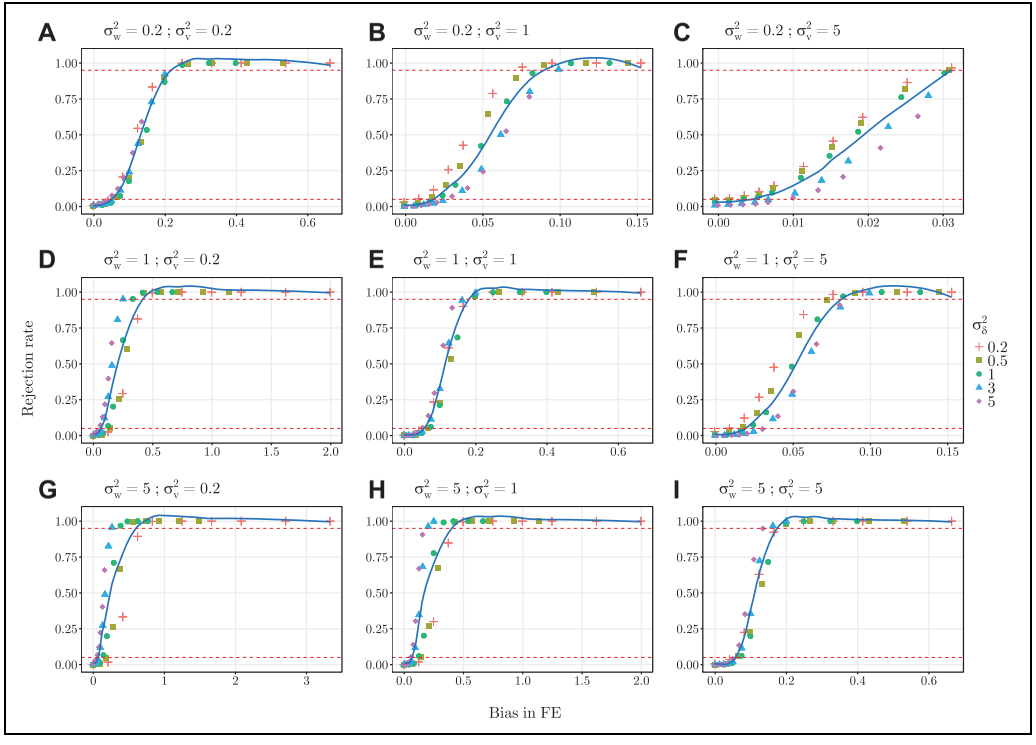

Figure 4. Rejection rates of Bootstrapped Hausman Test. δ = effect of slope variable $\boldsymbol{w}$ on covariate $\boldsymbol{x}$; α_2 = effect of slope variable $\boldsymbol{w}$ on outcome $\boldsymbol{y}$; ν = independent random vector in covariate $\boldsymbol{x}$; σ^2 = variance of δ, $\boldsymbol{w}$, and ν respectively; and $N = 300$, $T = 10$, $R = 1,000$, and $R_b = 100$. (A) $\sigma_w^2 = 0.2$; $\sigma_v^2 = 0.2$. (B) $\sigma_w^2 = 0.2$; $\sigma_v^2 = 1$. (C) $\sigma_w^2 = 0.2$; $\sigma_v^2 = 5$. (D) $\sigma_w^2 = 1$; $\sigma_v^2 = 0.2$. (E) $\sigma_w^2 = 1$; $\sigma_v^2 = 1$. (F) $\sigma_w^2 = 1$; $\sigma_v^2 = 5$; (G) $\sigma_w^2 = 5$; $\sigma_v^2 = 0.2$. (H) $\sigma_w^2 = 5$; $\sigma_v^2 = 1$. (I) $\sigma_w^2 = 5$; $\sigma_v^2 = 5$.

contrast, Cameron and Miller (2015) recommend using the BSHT. Note, however, that Cameron and Miller (2015) consider only the size of the test, and replications of our own simulations with small N and large T ($N = 20$, $T = 30$) yield a poor performance of both tests in terms of power (see Supplementary Material S1, which can be found at http://smr.sagepub.com/supplemental/). Thus, we recommend to take results of the specification tests with caution when relying on a small number of groups. In case of a sufficient sample size, in contrast, both tests do a very good job in our simulations.

Applied Examples

So far, we have shown theoretically and by simulation experiments that heterogeneous slopes can seriously bias the results of conventional FE and RE models. However, a crucial question is whether this actually constitutes a problem in applied research. To demonstrate the usefulness of the FEIS estimator, we replicate and extend two applied examples in which relying on the FE estimator produces misleading conclusions. First, we present results for the already mentioned marital wage premium, starting from the study by V. Ludwig and Brüderl (2018). This constitutes a typical and intuitive example based on individual panel data where person-specific time trends are correlated with selection into treatment. Second, we reproduce the results of a study on Head Start returns by Deming (2009). Using family FE, the article estimates the effect of participation in an early childhood educational program (Head Start) on children's later cognitive ability. In this case, we do not have a panel data setting with time periods nested in persons, but a setting with children who are nested in families.[2] Thus, the example demonstrates how FEIS models can be usefully applied to multilevel (hierarchical) data structures more generally.

Example 1: The Male Marital Wage Premium

In this example (V. Ludwig and Brüderl 2018), we use the data set of the original study, which consists of individual panel data from the National Longitudinal Survey of Youth 1979 (Bureau of Labor Statistics 2014).[3] We investigate the "marital wage premium": We analyze whether marriage leads to an increase in the hourly wage for men. By estimating FEIS models, we rule out the alternative that those men who eventually get married also show steeper wage growth over their career (even before marriage). Therefore, we specify actual work experience (in years) and squared work experience as slope variables.

Table 1. Example 1: Regression results for log hourly wage rates.

	RE	FE	RIRS	FEIS
Married	0.100***	0.080***	0.064***	0.005
	(0.008)	(0.008)	(0.006)	(0.009)
Currently enrolled	−0.192***	−0.199***	−0.148***	−0.123***
	(0.009)	(0.010)	(0.006)	(0.010)
Years education	0.074***	0.068***	0.057***	0.007
	(0.002)	(0.004)	(0.002)	(0.006)
One child	0.015	0.020*	0.005	−0.016
	(0.008)	(0.008)	(0.007)	(0.010)
Two children	0.036***	0.038***	0.011	−0.028
	(0.011)	(0.012)	(0.009)	(0.015)
Three or more children	−0.003	0.004	−0.020	−0.054*
	(0.016)	(0.018)	(0.013)	(0.024)
Tenure (years)	0.013***	0.011***	0.011***	0.008***
	(0.001)	(0.001)	(0.001)	(0.001)
Work experience	0.047***	0.044***	0.055***	
	(0.002)	(0.002)	(0.002)	
Work experience ^{2}	−0.001***	−0.001***	−0.001***	
	(0.000)	(0.000)	(0.000)	
R^2	.380	.336		.019
Adjusted R^2	.380	.274		.019
Number of observations	49,801	49,801	49,801	49,801
Number of groups: id			4,287	4,287

Note: Robust standard errors are given in parentheses. Age group omitted. FE = fixed effects; RIRS = random slopes multilevel model; RE = random effects; FEIS = fixed effects individual slopes.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 1 shows the regression results of the hourly wage rate on marital status (plus additional controls) for different model specifications: RE model, conventional FE model, a random intercept random slopes multilevel model (RIRS), and the FEIS model. Note that we include a random slopes model here, because the RIRS allows for individual-specific trends as well. As in the FEIS, random slopes are specified for work experience and work experience squared. In the RIRS model, however, the individual variation of the intercepts and slopes is assumed exogeneous. This assumption is not needed with the FEIS model.

As can be seen in Table 1, the effect of marriage on the wage rate varies notably across the models. The RE and the FE return a significant effect of

Table 2. Example 1: Specification Tests.

	χ^2	*df*	p (> χ^2)
Artificial regression test			
FEIS versus FE	436.966	13	.000
FE versus RE	201.929	15	.000
FEIS versus RE	550.883	13	.000
Bootstrapped Hausman test			
FEIS versus FE	581.808	13	.000
FE versus RE	168.395	15	.000
FEIS versus RE	748.108	13	.000

Note: FE = fixed effects; RE = random effects; FEIS = fixed effects individual slope.

marriage on the log wage of 0.100 and 0.080, respectively. Similarly, the RIRS returns a significant effect of marriage, but at a slightly lower magnitude of 0.064. In comparison, the FEIS model reports an effect of only 0.005 which is far from significant. Note that the FE estimate exceeds the FEIS estimate by a factor of 16, and also the RIRS is nearly 14 times higher than the FEIS estimate. Consequently, according to the FE (as well as RE and RIRS), we would conclude that there really is a marital wage premium for men. However, the FEIS revises this conclusion by showing that the effect mostly stems from the fact that men with a steeper wage growth tend to marry earlier (those men with a stronger effect of experience on wage also exhibit a stronger effect of experience on marriage). In consequence, our replication underlines the conclusion of the original study by V. Ludwig and Brüderl (2018): The marital wage premium is mainly due to selection on wage growth rather than being a causal effect of marriage on men's productivity.

This example illustrates how heterogeneous growth curves related to the covariates can drastically bias the conclusions drawn from conventional FE and RE models that are typically applied to panel data. Therefore, we propose to test for the presence of heterogeneous slopes in panel models by default.

Table 2 shows the results of the proposed ART and BSHT specification tests. Overall, both tests strongly reject the Null that the conventional FE and RE models provide consistent estimates. Regarding the first test comparing FEIS against FE, a highly significant χ^2 value of 436.966 (581.808 in the bootstrapped version) indicates that estimates of the conventional FE are inconsistent because of heterogeneous slopes. Thus, we should use FEIS instead of FE. The second test of FE against RE is equivalent to the conventional Hausman test under the assumption of equal error variance for both

models, but it does allow for clustered standard errors. The ART returns a highly significant χ^2 and thus favors FE over an RE model.[4] The third test statistic offers a direct comparison of FEIS against RE: Both the ART and the BSHT show a highly significant χ^2, thereby indicating that we should reject the Null hypothesis of consistent estimates in the RE model. Taken together, when ignoring the possibility of heterogeneous slopes, we would lean toward using a conventional FE model, thereby relying on an effect of marriage which equals 16 times the effect obtained in a model accounting for individual-specific slopes.

Example 2: Returns of the Head Start Preschool Program

While the first example constitutes a textbook case for the application of FEIS with heterogenous growth curves, we will now turn to an example in which we apply the FEIS to a more general hierarchical data structure. The example concerns the impact of Head Start, a large-scale U.S. preschool program for disadvantaged families, on children's cognitive performance. In general, there has been a large debate on whether Head Start participation induces cognitive benefits, and especially whether it generates long-term benefits on schooling and other outcomes (e.g., J. Ludwig and Miller 2007). We know from experimental evidence of other preschool programs that preschool interventions improve educational outcomes of children from disadvantaged families in the short and in the long run (Heckman 2006; Schweinhart et al. 2005). In addition, using experimental data of a randomly drawn subsample of Head Start participants, J. Ludwig and Phillips (2008) report short-term treatment effects of 0.15–0.35 standard deviations for a range of cognitive outcomes measured at age 3 or 4. However, findings with nonexperimental data for children participating in Head Start differ markedly in that effects seem to be much smaller, and effects fade-out over time (Duncan and Magnuson 2013). As we usually lack continuous information on children's cognitive outcomes, most observational studies apply a within-family design (siblings estimators) to account for children's individual preprogram conditions.

Below, we replicate and extend the study by Deming (2009). The author used data from the National Longitudinal Mother–Child Supplement, which surveyed the mothers of the NLSY79 (Bureau of Labor Statistics 2014) biannually from 1986 through 2004. When relying on simple OLS models, he does not find substantial short-term effects on cognitive outcomes. Deming (2009) suspects the estimates are biased downward. In the authors' own words,

> [b]ecause children in Head Start come from very disadvantaged families, a simple comparison to children in other preschools or no preschool on outcomes such as test scores or educational attainment will be biased downward. In the absence of a randomized experiment, the challenge is to counteract this downward bias using non-experimental methods. This motivates the use of family fixed effects. (P. 118)

Essentially, using family fixed effects assumes that selection into Head Start among children from the same family is not related to unobserved characteristics of the children, which simultaneously correlate with school performance. This certainly is a strong assumption. Therefore, the author included an extensive set of 11 covariates to control for different pretreatment conditions of children within the same family. He also assembled an index of children's starting conditions from pretreatment covariate values in which high values indicate overall favorable starting conditions. However, controlling for child-specific pretreatment conditions did not alter the results substantively. In the final (preferred) model, Deming (2009) finds that Head Start increases the test scores of children at ages 5–6 by 0.145 standard deviations, which meets the lower bound of results reported in J. Ludwig and Phillips (2008). Moreover, the effects further declines for test score results at older age.

However, the low (long term) returns may be the result of a problem of heterogenous slopes on the family level: If disadvantaged mothers are less able or willing to balance disparate starting conditions, their children's starting conditions would be strongly (positively) linked to later outcomes. At the same time, children's starting conditions may be more strongly (negatively) linked to participation in Head Start in these families (e.g., Behrman and Rosenzweig 2004; Griliches 1979). Perhaps disadvantaged mothers anticipate they will not be able to fully support their offspring given their restricted family resources. In this case, the treatment effect would be underestimated in a conventional family FE model which neglects family-specific heterogeneity in the effect of starting conditions.

Indeed, our reanalysis of Deming's data set[5] supports this argument. Model 1 in Table 3 perfectly replicates Deming's main result. Even after controlling for family FE and an extensive set of covariates to capture children's starting conditions, the effect of participation in Head Start is rather weak, ranging from 0.05 to 0.14 standard deviations of test performance. Model 2 estimates the same family FE model using a slightly smaller sample (families with fewer than three observations excluded) and a different specification of control variables, replacing the covariates for children's starting

Table 3. Example 2: Regression results for cognitive test scores.

	Original Sample			Subsample		
	Replication (1)	FE (2)	FEIS (3)	FE 2 (4)	RIRS 2 (5)	FEIS 2 (6)
Head Start						
Ages 5–6	0.143	0.133	0.350**	0.148	0.176*	0.368**
	(0.085)	(0.087)	(0.115)	(0.103)	(0.086)	(0.124)
Ages 7–10	0.132*	0.117	0.319***	0.105	0.133*	0.330***
	(0.059)	(0.060)	(0.096)	(0.066)	(0.054)	(0.091)
Ages 11–14	0.054	0.029	0.241*	0.041	0.070	0.293**
	(0.061)	(0.061)	(0.102)	(0.066)	(0.053)	(0.100)
Other Preschool						
Ages 5–6	−0.081	−0.105	−0.095	−0.062	0.007	0.037
	(0.084)	(0.083)	(0.132)	(0.166)	(0.113)	(0.270)
Ages 7–10	0.046	0.029	0.009	0.000	0.066	0.092
	(0.064)	(0.061)	(0.120)	(0.118)	(0.078)	(0.254)
Ages 11–14	−0.023	−0.040	−0.060	−0.024	0.037	0.080
	(0.069)	(0.066)	(0.120)	(0.120)	(0.080)	(0.255)
Pretreatment index		0.056		0.109**	0.156***	
		(0.034)		(0.037)	(0.036)	
R^2	.050	.020	.031	.027		.038
Adjusted R^2	−.099	−.115	.027	−.093		.028
Number of observation	4,687	4,646	4,646	2,102	2,102	2,102
Number of groups: MotherID			541		211	211

Note: Robust standard errors are given in parentheses. FE = fixed effects; RIRS = random slopes multilevel model; RE = random effects; FEIS = fixed effects individual slope.
*$p < .05$. **$p < .01$. ***$p < .001$.

conditions by Deming's pretreatment index. The estimates of the treatment effect remain very similar in this model (albeit somewhat smaller compared to Model 1). Model 3 extends the specification by allowing for family-specific slopes on the pretreatment index. This more flexible FEIS model yields substantially larger treatment effects, in the range of 0.24–0.35 standard deviations. Unlike the FE results, this finding is in line with experimental evidence (J. Ludwig and Phillips 2008). Apparently, the effect of children's pretreatment conditions on later test performance differs between

Table 4. Example 2: Specification Tests.

	χ^2	*df*	p (> χ^2)
Artificial regression test			
FEIS versus FE	36.555	21	.019
FE versus RE	32.538	22	.069
FEIS versus RE	35.003	21	.028
Bootstrapped Hausman test			
FEIS versus FE	41.497	21	.005
FE versus RE	20.659	22	.542
FEIS versus RE	32.200	21	.056

Note: FE = fixed effects; RE = random effects; FEIS = fixed effects individual slope.

families, and the heterogeneity in this effect is also linked to program enrolment.

Despite the large difference in effects between the FE and FEIS estimates (models 2 and 3), a specification test indicates that the family-FE model is not biased. The ART returns a p value of .219, which seems to indicate that there is no need to specify family-specific slopes for the pretreatment index. However, the nonsignificant test statistic follow from construction of the estimation sample. As Deming (2009) notes, in his study, "the impact of Head Start is identified by comparing siblings in the same family who vary in their participation in preschool programs. Thus, if all three of a mother's children were enrolled in Head Start, I cannot say anything about the effectiveness of the program for them" (p. 115).

Accordingly, we further restrict the sample to mothers with at least two children who differ in Head Start participation and reestimate our FE and FEIS models. Models 4 and 6 still provide similar, but somewhat stronger estimates for the Head Start treatment effect (compared to models 2 and 3). As in the first example, the estimates of the random slope multilevel model (Model 5) lie between FE and FEIS, but it is still far away from treatment effects reported in FEIS. Interestingly, the effect of visiting other Preschools also turns positive in the FEIS model. Based on model 6, the ART returns a p value of .019, which indicates that we should in fact specify heterogeneous slopes for the pretreatment index (Table 4).

Why does the test work with the smaller sample? The reason is that, unless we apply the restriction to families with variation on the treatment variable, the CRE model estimated for the test suffers from collinearity. If children from the same family do not differ with respect to program participation, the family-specific predicted values for Head Start covariates are equal to the

family-specific mean values. Since this holds for 61 percent of families contained in the full original sample, predicted and mean values are highly collinear in the CRE model, resulting in large standard errors and an insignificant test statistic for the ART.[6] In terms of our simulation results, the power of the ART is insufficient when using the full sample because we run into a situation shown in Figure 3G with low variation of both the treatment variable ($\sigma_w^2 >> \sigma_v^2$) and the effect of the slope variable (σ_δ^2 is small).

Table 4 highlights a further important point: In practice, a test for the conventional FE versus RE model may indicate there is no problem with selection into treatment at all. In our case, the test returns a p value of .069 (0.542 for the bootstrapped test). Hence, it seems that family-specific FE do not change the results and are therefore not needed. However, both the FE and the RE model are actually inconsistent because selection is mainly due to heterogeneous slopes across families. Thus, when RE is the preferred model, we highly recommend to test the RE against the FE *and* against the FEIS model.

In summary, our reanalysis of the study by Deming (2009) suggests that the positive impact of the Head Start program on children's cognitive development has been underestimated in previous research because of the heterogeneous ability or willingness of disadvantaged families to support children with particularly bad starting conditions, which is also related to participation of these children in Head Start. When accounting for this heterogeneity, we find higher short- and long-term cognitive returns of Head Start participation, a result resembling previous experimental findings.

Limitations

Though the simulations and previous examples show that testing and accounting for heterogeneous slopes can help to avoid erroneous conclusions, it is important to note that FEIS and our test statistics come with some limitations. FE, and also the extended FEIS model, may suffer from other sources of bias. When applying these methods, researchers should be aware that biases due to measurement error on the causal variable (Freeman 1984; Griliches and Hausman 1986) or true state dependence of the outcome process (Brüderl and Ludwig 2015; Nickell 1981; Phillips and Sul 2007) may lead to erroneous model selection and wrong conclusions. While these limitations have been established years ago, recent methodological studies have pointed out potential shortcomings of FEIS in particular.

First, it has been argued previously that FEIS "absorbs" part of the treatment effect (e.g., Goodman-Bacon 2018; Kneip and Bauer 2009; Meer and

West 2016). On the one hand—as with every control variable—it is important to cautiously select the slope variables (Morgan and Winship 2015; Pearl 2009). The set of possible slope variables should only include confounding variables, which causally influence the response and the treatment variable. As with every control variable, including mediating or colliding variables, in contrast, leads to overcontrol or selection bias and will thus distort the treatment effect. The limitation described by Goodman-Bacon (2018), Kneip and Bauer (2009), and Meer and West (2016), however, relates to a more complicated case. As shown by Meer and West (2016:507), FEIS might absorb part of the treatment effect if the treatment effect works on the long-run trend (or whatever slope) of the outcome. Controlling for individual trends while only specifying a contemporary treatment effect might then lead to an underestimation of total treatment effect size.

To demonstrate this point, we conducted additional simulation runs for the case in which the true DGP follows a lagged treatment structure $y_{in} = x_{in}\beta + x_{in-1}\beta\zeta + x_{in-2}\beta\zeta + \alpha_{1i} + w_{in}\alpha_{2i} + \varepsilon_n$ for data ordered by $\boldsymbol{w}_i$ within each group. ζ specifies the discount rate of the lagged treatment effects and $\zeta \neq 0$, but the estimation model erroneously assumes $\zeta = 0$. The results of the bias and the ART specification test are shown in Figures S4 and S5 of the Supplementary Material, which can be found at http://smr.sagepub.com/supplemental/. In line with the abovementioned argument, FEIS underestimates the contemporary effect in case of erroneously assuming $\zeta = 0$, and this bias increases with the size of the lagged effects (comparing Figure S4 across A–D [which can be found at http://smr.sagepub.com/supplemental/]). If the two lagged treatment effects are of similar size as the contemporary effect, FEIS underestimates the contemporary effect by 23 percent, as changes in the outcome during following periods are erroneously "detrended." However, on the other hand, FE strongly overestimates the treatment effects. When $\phi = 0$ and FE is not affected by heterogeneous slopes, the true effect is overestimated by 59 percent, and obviously the bias increases with ϕ. When treatment might work on the trend, there is a trade-off between overestimating the treatment effect in FE, possibly confounded by heterogeneous trends, and underestimating the treatment effect in FEIS due to controlling for confounded trends or slopes. In this case, cautious theoretical arguing is necessary to guide the model specification. Of course, bias disappears in FEIS if the sequential impacts of treatment are specified correctly in the estimation model (Figure S6, which can be found at http://smr.sagepub.com/supplemental/), a strategy Kneip and Bauer (2009) and Wolfers (2006), for instance, applied to investigate the dynamic effect of unilateral divorce regimes on divorce rates. In our

opinion, this shortcoming is thus rather related to a general problem of misspecification than to a problem of the estimator itself.

Still, the proposed specification tests are of no help in this situation of general misspecification. Though heterogeneous slopes are correlated with the treatment ($\phi \neq 0$), ART does not reject H_0 of FE being consistent if the opposing biases in FE lead to an estimate close to FEIS (e.g., Figure S5B at $\phi = -0.4$, which can be found at http://smr.sagepub.com/supplemental/). Similarly, if both FE and FEIS are biased by misspecified lagged effects, but FE estimates the correct treatment effect because of a countervailing bias due to slope heterogeneity (Figure S5C at $\phi = -0.5$, which can be found at http://smr.sagepub.com/supplemental/), the specification test will be misleading: ART points toward using FEIS, though FE returns the true value. As with the conventional Hausman test, the two proposed specification tests can only detect differences in estimates due to time-constant and slope heterogeneity but are not robust against other forms of misspecification. Still, if the estimation model is specified correctly, the tests perform reasonably well even with dynamic/lagged treatment effects (Figure S7, which can be found at http://smr.sagepub.com/supplemental/).

A second limitation of the FEIS is its need for long panels or other forms of data with relatively many observations per group (e.g., Morgan and Winship 2015). As described above, the FEIS needs $T > J$ for each observation to be included in the estimation of the treatment effect. Thus, for one slope, we need at least three observations per group, while FE requires only two observations. Though this seems not too limiting given the length of many panel studies, often it might be necessary to include more than one slope parameter. For instance, in the marriage wage premium example, V. Ludwig and Brüderl (2018) control for nonlinearity of trends, thereby requiring at least four observations per person. Moreover, often higher polynomials of a time trend may be necessary (e.g., Kneip and Bauer 2009; Kneip, Bauer, and Reinhold 2014), thereby making FEIS increasingly data-demanding. Similarly this problem can arise in other (nonpanel) data structures. In the Head Start example, we used one pretreatment index as slope, summing over 11 indicators of pretreatment conditions. This is, of course, an oversimplification and using all 11 variables as slopes would add more precision. However, this would require at least 13 observations (children) per family, which obviously is rarely available.

This is especially problematic in short unbalanced panels: In case of unbalanced panels and/or a relatively large number of necessary slope variables, we might actually estimate the treatment effect based on a specific subset of groups (for a general discussion on selection, see, e.g., Semykina

and Wooldridge 2010; Wooldridge 2010). For instance, considering sibling FE (like the Head Start example), we can only apply FEIS to families with at least three children. Nevertheless, it might be reasonable to assume that some treatment effects differ between small and large families. Obviously, the problem gets more severe when using multiple slopes, which might be necessary to accurately approximate the true trend. Figures S8 and S9 of the Supplementary Material, which can be found at http://smr.sagepub.com/supplemental/, provide some insights on how different types of sample selection influence bias in the three considered models and the corresponding specification tests. Conventional selection bias, where selection $s_{it} = 1[\ \mathrm{f}\ (\ddot{Y}_{it})]$, affects FE and FEIS in a very similar way (see A and B of Figure S8, which can be found at http://smr.sagepub.com/supplemental/). However, when assuming heterogeneous treatment effects, where the group with a lower treatment effect has a significantly higher probability of contributing only two observations (thereby dropping out of FEIS estimation), the selection bias becomes more severe in FEIS than in FE (Figure S8C and D, which can be found at http://smr.sagepub.com/supplemental/). Again, relying on the proposed tests can be misleading in this case (Figure S9, which can be found at http://smr.sagepub.com/supplemental/). It is thus important to carefully evaluate the trade-off between the number of necessary slope variables and the consequences of increasing sample selectivity in case of unbalanced panels.

Still, we believe that the growing availability of long-run panel studies (like PSID or SOEP) and balanced administrative data allow to apply FEIS for a variety of (but not all) research questions without loosing important cases. Furthermore, FEIS does not only provide a powerful tool for panel data but also for other types of data and nesting structures, like multilevel data. These data often come with a large number of observations clustered within higher units, as for instance, census tracts nested in cities, which drastically reduces the problem due to data requirements and potential selection.

Conclusions

The FE panel estimator has become increasingly popular in economics and social science research because of its property to control for time-constant heterogeneity. In this article, we examined a major drawback of this approach: Conventional FE estimates of a causal effect may suffer from a bias due to heterogeneous unit-specific slopes, which are correlated with (at least) one covariate in the model. We derive the bias of the conventional FE

model analytically and propose to use a more general FE model instead: the FE model with individual slopes (FEIS, Brüderl and Ludwig 2015; Frees 2001; Lemieux 1998; Polachek and Kim 1994; Wooldridge 2010). This model is unbiased in the presence of heterogeneous slopes. In addition, we introduce two versions of the Hausman test that allow to check whether FE estimates are biased in practice: the ART and the BSHT. The FEIS model can be estimated in `R` and `Stata` using the packages *feisr* or *xtfeis*, respectively, also including functions to implement the ART and BSHT. In Monte Carlo experiments, we vary the factors that drive the FE bias and evaluate the performance of the FE estimator and the power of the ART and BSHT to detect biased estimates. Finally, we present two applied examples in which using FEIS instead of conventional FE models leads to important differences in the conclusions and discuss some important limitations of the estimator and the test statistics.

Our study shows that the magnitude of the FE bias due to heterogeneous slopes depends on four components. (a) In particular, the bias grows as the covariance of unit-specific slopes and the effect of the slope variable on the treatment variable increases. This component of the bias is crucial since it determines the sign of the bias. (b) Furthermore, the bias is stronger the lower the variation of the correlation between the slope variable and the causal variable in the population. (c) Finally, the (absolute) size of the bias depends on the variation of both the causal variable and the slope variable: The bias can be large if the variance of the slope variable is large relative to the independent variance of the covariate of interest.

What does that mean in practice? A typical setting for application of FE models is the estimation of some treatment effect while controlling for changes of the outcome over time. For example, we might be interested in the effect of further training on individual earnings, where we would control for the growth of earnings over age. Our analytical results show that (a) the conventional FE model would return a biased training effect if the slopes of individual age-earnings profiles differ between people, and if these individual differences in slopes are related to the effect of age on training. This would be the case if, for example, more motivated individuals are more likely to participate in further training, but due to their higher motivation would experience higher wage growth even without training. Our results further indicate that the FE training effect would be more heavily biased (b) if the effect of age on training participation is homogeneous in the sample, and (c) if variation of age within individuals is large (which we will have "by nature"), but variation in training is small.

In the research on returns to education, it is a well-known problem that conventional FE coefficients are biased because the assumption of parallel slopes does not hold (Dynarski, Jacob, and Kreisman 2018; Heckman and Hotz 1989). However, it is likely that other research areas are led astray by biased FE estimates as well. In our first empirical example motivated by V. Ludwig and Brüderl (2018), we show that the "wage premium" for married men found in previous studies using standard FE estimation is a statistical artifact. In a second example on preschool returns based on work by Deming (2009), we illustrate that accounting for heterogeneous effects is also important in more general nested multilevel scenarios. Those families exhibiting a stronger correlation between children's starting conditions and later outcomes also show a stronger effect of starting conditions on program enrolment. Relying on conventional FE models thus underestimates the effect of Head Start participation on later test score outcomes.

Nevertheless, when applying FEIS, two considerations need to be made. First, it is important that the structure of the treatment effect is correctly specified. If the treatment effect flows dynamically along the slope variable, and this is not accounted for (e.g., due to lags) in the model, FEIS might mask part of the treatment effect. Though FE will be biased in this case too, before applying FEIS, it is crucial to carefully think about the structure of the treatment effect. Kneip and Bauer (2009) and Wolfers (2006) provide two helpful examples in this regard. Second, when using short and/or unbalanced panels, FEIS only uses individuals or groups which contribute at least $J + 1$ observations. Especially with multiple slopes, this might lead to estimation based on a subset of individuals only. It is thus worth considering whether there might be significant differences between the estimation (sub)sample and the population of interest. If there are no theoretical reasons to belief so, or data are rich enough to account for all necessary slopes, we highly recommend testing for heterogeneity of slopes.

Our Monte Carlo simulations show that the ART and BSHT do a good job in detecting this type of bias. If the treatment effect is specified correctly, tests have excellent size, as the FE model is rejected in less than 5 percent of 1,000 samples if the true bias is zero. They both also have reasonable power to detect a bias of the FE estimator. Rejection rates of the FE model are at least 95 percent for a bias of 20 percent in most settings. Overall, the results are qualitatively similar to earlier simulation studies of the Hausman test for detecting misspecified RE models (Ahn and Low 1996; Amini et al. 2016; Baltagi 1981). However, according to our results, following the textbook advice of relying on the standard Hausman test when choosing a model for analyzing panel or multilevel data can be misleading. As our simulations and

examples illustrate, it is crucial to additionally test for a bias due to heterogeneous slopes.

We therefore recommend to use the ART or the BSHT, and test for heterogeneity bias of FE or RE, whichever is the preferred basic model. Both tests can be made robust to autocorrelation and heteroscedasticity, which is important because panel-robust standard errors are used by default in most applications. Asymptotically, both tests are equivalent. In any case, we recommend to first test the basic specification against the FEIS model. Though—as the conventional Hausman test—the proposed test statistics are not robust against alternative sources of misspecification, they can help preventing faulty conclusions due to a common source of bias. If the test indicates slope heterogeneity, FEIS is the preferred model, while FE and RE results should be taken with caution.

Authors' Note

Supplementary material for reproduction of all analyses in this article is provided at the author's GitHub repository: https://github.com/ruettenauer/Reproduction-Material-Fixed-EffectsIndividual-Slopes. The repository contains R scripts to conduct the Monte Carlo experiments, as well as R and Stata Scripts to reproduce the empirical examples.

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.

ORCID iD

Tobias Rüttenauer https://orcid.org/0000-0001-5747-9735

Supplemental Material

Supplemental material for this article is available online.

Notes

1. To ensure that the bivariate distribution is based on a positive definite covariance matrix, we identify the nearest positive definite matrix based on Higham (2002) algorithm (using the nearPD function in the *Matrix* package by Bates and

Maechler [2018]). Deviations of the parameters chosen by the algorithm from the original parameters are negligible in the simulations and do not distort the results.
2. Actually, most children in the data are observed more than once over time. We, as Deming, will ignore this subtlety in the following.
3. Details on selection of the estimation sample and variable construction can be found in V. Ludwig and Brüderl (2018).
4. However, the test statistic is much smaller than for the test of fixed effects individual slope versus fixed effects (FE)—in line with the fact that random effects (RE) and FE produce relatively similar coefficients for the covariates. We can easily imagine that the test would erroneously direct us to the RE model in samples that are not as large as the one we analyze here (see Example 2).
5. We are thankful to the author for making a comprehensive set of replication materials for the results in Deming (2009) publicly available at www.aeaweb.org/articles?id=10.1257/app.1.3.111.
6. This highlights an important point that is easily overlooked in applied research using FE models: Construction of the estimation sample should always guarantee that the causal variable varies within units to make sure the within effect is identified. Note that in our first empirical example, this was accomplished by restricting the sample to initially never-married men.

References

Ahn, S. C. and S. Low. 1996. "A Reformulation of the Hausman Test for Regression Models with Pooled Cross-section-time-series Data." *Journal of Econometrics* 71(1-2):309-19.

Allison, P. D. 2009. *Fixed Effects Regression Models, Volume 160 of Quantitative Applications in the Social Sciences*. Los Angeles: Sage.

Amini, S., M. S. Delgado, D. J. Henderson, and C. F. Parmeter. 2016. "Fixed vs Random: The Hausman Test Four Decades Later." Pp. 479-513 in *Spatial Econometrics: Qualitative and Limited Dependent Variables*, Volume 29 of *Advances in Econometrics*, edited by B. H. Baltagi, J. P. LeSage, and R. K. Pace. Bingley, England: Emerald Group Publishing Limited.

Angrist, J. D. and J.-S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Arellano, M. 1987. "Computing Robust Standard Errors for Within-groups Estimators." *Oxford Bulletin of Economics and Statistics* 49(4):431-34.

Arellano, M. 1993. "On the Testing of Correlated Effects with Panel Data." *Journal of Econometrics* 59(1-2):87-97.

Baltagi, B. H. 1981. "Pooling: An Experimental Study of Alternative Testing and Estimation Procedures in a Two-way Error Component Model." *Journal of Econometrics* 17(1):21-49.

Baltagi, B. H. 2013. *Econometric Analysis of Panel Data*. 5th ed. Chichester, England: Wiley.

Banzhaf, H. S. and R. P. Walsh. 2008. "Do People Vote with Their Feet? An Empirical Test of Tiebout's Mechanism." *American Economic Review* 98(3):843-63.

Bates, D. and M. Maechler. 2018. "Matrix: Sparse and Dense Matrix Classes and Methods." R package version 1.2-14. https://CRAN.R-project.org/package=Matrix

Behrman, J. R. and M. R. Rosenzweig. 2004. "Returns to Birthweight." *Review of Economics and Statistics* 86(2):586-601.

Best, H. and T. Rüttenauer. 2018. "How Selective Migration Shapes Environmental Inequality in Germany: Evidence from Micro-level Panel Data." *European Sociological Review* 34(1):52-63.

Brüderl, J. and V. Ludwig. 2015. "Fixed-effects Panel Regression." Pp. 327-57 in *The Sage Handbook of Regression Analysis and Causal Inference*, edited by H. Best and C. Wolf. Los Angeles: Sage.

Bureau of Labor Statistics. 2014. *National Longitudinal Survey of Youth 1979 Cohort, 1979-2012* (rounds 1-23). Columbus, OH: Center for Human Resource Research, The Ohio State University.

Cameron, A. C. and D. L. Miller. 2015. "A Practitioner's Guide to Cluster-robust Inference." *Journal of Human Resources* 50(2):317-72.

Cameron, A. C., J. B. Gelbach, and D. L. Miller. 2008. "Bootstrap-based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90(3):414-27.

Chamberlain, G. 1982. "Multivariate Regression Models for Panel Data." *Journal of Econometrics* 18(1):5-46.

Chetty, R. and N. Hendren. 2018. "The Impacts of Neighborhoods on Intergenerational Mobility II: County-level Estimates." *The Quarterly Journal of Economics* 133(3):1163-228.

Croissant, Y. and G. Millo. 2008. "Panel Data Econometrics in R: The plm Package." *Journal of Statistical Software* 27(2):1-43.

Deming, D. 2009. "Early Childhood Intervention and Life-cycle Skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1(3):111-34.

DiPrete, T. A. and P. A. McManus. 2000. "Family Change, Employment Transitions, and the Welfare State: Household Income Dynamics in the United States and Germany." *American Sociological Review* 65(3):343.

Dougherty, C. 2006. "The Marriage Earnings Premium as a Distributed Fixed Effect." *Journal of Human Resources* 41(2):433-43.

Duncan, G. J. and K. Magnuson. 2013. "Investing in Preschool Programs." *Journal of Economic Perspectives* 27(2):109-32.

Dynarski, S., B. Jacob, and D. Kreisman. 2018. "How Important are Fixed Effects and Time Trends in Estimating Returns to Schooling? Evidence from a Replication of Jacobson, Lalonde, and Sullivan, 2005." *Journal of Applied Econometrics* 33(7): 1098-108.

Figlio, D., J. Guryan, K. Karbownik, and J. Roth. 2014. "The Effects of Poor Neonatal Health on Children's Cognitive Development." *American Economic Review* 104(12):3921-55.

Firebaugh, G., C. Warner, and M. Massoglia. 2014. "Fixed Effects, Random Effects, and Hybrid Models for Causal Analysis." Pp. 113-32 in *Handbook of Causal Analysis for Social Research*, *Handbooks of Sociology and Social Research*, edited by S. L. Morgan. Dordrecht, the Netherlands: Springer.

Freeman, R. B. 1984. "Longitudinal Analyses of the Effects of Trade Unions." *Journal of Labor Economics* 2(1):1-26.

Frees, E. W. 2001. "Omitted Variables in Longitudinal Data Models." *Canadian Journal of Statistics* 29(4):573-95.

Frijters, P. and T. Beatton. 2012. "The Mystery of the U-shaped Relationship between Happiness and Age." *Journal of Economic Behavior & Organization* 82(2-3): 525-42.

Frisch, R. and F. V. Waugh. 1933. "Partial Time Regressions as Compared with Individual Trends." *Econometrica* 1(4):387-401.

Goodman-Bacon, A. 2018. "Difference-in-differences with Variation in Treatment Timing." *NBER Working Paper, 25018*. https://www.nber.org/papers/w25018

Griliches, Z. 1979. "Sibling Models and Data in Economics: Beginnings of a Survey." *The Journal of Political Economy* 87(5):S37-64.

Griliches, Z. and J. A. Hausman. 1986. "Errors in Variables in Panel Data." *Journal of Econometrics* 31(1):93-118.

Hausman, J. A. 1978. "Specification Tests in Econometrics." *Econometrica* 46(6): 1251-71.

Heckman, J. J. 2006. "Skill Formation and the Economics of Investing in Disadvantaged Children." *Science* 312(5782):1900-1902.

Heckman, J. J. and V. J. Hotz. 1989. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association* 84(408):862.

Henderson, J. V., T. Squires, A. Storeygard, and D. Weil. 2018. "The Global Distribution of Economic Activity: Nature, History, and the Role of Trade." *The Quarterly Journal of Economics* 133(1):357-406.

Higham, N. J. 2002. "Computing the Nearest Correlation Matrix—A Problem from Finance." *IMA Journal of Numerical Analysis* 22(3):329-43.

Jacobson, L., R. LaLonde, and D. G. Sullivan. 2005. "Estimating the Returns to Community College Schooling for Displaced Workers." *Journal of Econometrics* 125(1-2):271-304.

Killewald, A. and M. Gough. 2013. "Does Specialization Explain Marriage Penalties and Premiums?" *American Sociological Review* 78(3):477-502.

Kneip, T. and G. Bauer. 2009. "Did Unilateral Divorce Laws Raise Divorce Rates in Western Europe?" *Journal of Marriage and the Family* 71(3):592-607.

Kneip, T., G. Bauer, and S. Reinhold. 2014. "Direct and Indirect Effects of Unilateral Divorce Law on Marital Stability." *Demography* 51(6):2103-126.

Krueger, A. B. and A. I. Mueller. 2012. "Time Use, Emotional Well-being, and Unemployment: Evidence from Longitudinal Data." *American Economic Review* 102(3):594-99.

Kühhirt, M. 2012. "Childbirth and the Long-term Division of Labour within Couples: How do Substitution, Bargaining Power, and Norms affect Parents' Time Allocation in West Germany?" *European Sociological Review* 28(5):565-82.

Lemieux, T. 1998. "Estimating the Effects of Unions on Wage Inequality in a Panel Data Model with Comparative Advantage and Nonrandom Selection." *Journal of Labor Economics* 16(2):261-91.

Lichter, D. T., D. Parisi, and M. C. Taquino. 2015. "Toward a New Macro-segregation? Decomposing Segregation within and between Metropolitan Cities and Suburbs." *American Sociological Review* 80(4):843-73.

Loughran, D. S. and J. M. Zissimopoulos. 2009. "Why Wait? The Effect of Marriage and Childbearing on the Wages of Men and Women." *Journal of Human Resources* 44(2):326-49.

Lovell, M. C. 1963. "Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis." *Journal of the American Statistical Association* 58(304): 993-1010.

Ludwig, J. and D. L. Miller. 2007. "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design." *The Quarterly Journal of Economics* 122(1):159-208.

Ludwig, J. and D. A. Phillips. 2008. "Long-term Effects of Head Start on Low-income Children." *Annals of the New York Academy of Sciences* 1136:257-68.

Ludwig, V. and J. Brüderl. 2018. "Is There a Male Marital Wage Premium? New Evidence from the United States." *American Sociological Review* 83(4):744-70.

Massoglia, M., G. Firebaugh, and C. Warner. 2013. "Racial Variation in the Effect of Incarceration on Neighborhood Attainment." *American Sociological Review*, 78(1):142-165.

Meer, J. and J. West. 2016. "Effects of the Minimum Wage on Employment Dynamics." *Journal of Human Resources* 51(2):500-522.

Morgan, S. L. and C. Winship. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.

Mundlak, Y. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica* 46(1):69.

Nickell, S. 1981. "Biases in Dynamic Models with Fixed Effects." *Econometrica* 49(6):1417.

Noelke, C. 2016. "Employment Protection Legislation and the Youth Labour Market." *European Sociological Review* 32(4):471-85.

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge, England: Cambridge University Press.

Phillips, P. C. and D. Sul. 2007. "Bias in Dynamic Panel Estimation with Fixed Effects, Incidental Trends and Cross Section Dependence." *Journal of Econometrics* 137(1):162-88.

Pischke, J.-S. 2001. "Continuous Training in Germany." *Journal of Population Economics* 14(3):523-48.

Polachek, S. W. and M.-K. Kim. 1994. "Panel Estimates of the Gender Earnings Gap." *Journal of Econometrics* 61(1):23-42.

Qvist, H.-P. Y. and M. D. Munk. 2018. "The Individual Economic Returns to Volunteering in Work Life." *European Sociological Review* 34(2):198-210.

Reardon, S. F. and K. Bischoff. 2011. "Income Inequality and Income Segregation." *American Journal of Sociology* 116(4):1092-1153.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73(2):417-58.

Schweinhart, L. J., J. Montie, Z. Xiang, W. S. Barnett, C. R. Belfield, and M. Nores. 2005. *Lifetime Effects: The High/Scope Perry Preschool Study through Age 40*. Ypsilanti, MI: High/Scope Press.

Semykina, A. and J. M. Wooldridge. 2010. "Estimating Panel Data Models in the Presence of Endogeneity and Selection." *Journal of Econometrics* 157(2): 375-80.

Stock, J. H. and M. W. Watson. 2008. "Heteroskedasticity-robust Standard Errors for Fixed Effects Panel Data Regression." *Econometrica* 76(1):155-74.

Torche, F. and A. Villarreal. 2014. "Prenatal Exposure to Violence and Birth Weight in Mexico: Selectivity, Exposure, and Behavioral Responses." *American Sociological Review* 79(5):966-92.

Wallace, T. D. and A. Hussain. 1969. "The Use of Error Components Models in Combining Cross Section with Time Series Data." *Econometrica* 37(1):55.

Wolfers, J. 2006. "Did Unilateral Divorce Laws Raise Divorce Rates? A Reconciliation and New Results." *American Economic Review* 96(5):1802-20.

Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Author Biographies

Tobias Rüttenauer is a Postdoctoral Prize Research Fellow in Sociology at Nuffield College, University of Oxford. He received his doctorate from the TU Kaiserslautern, Germany. His research focuses on environmental sociology, spatial demography, and the application of statistical methods for spatial and panel data.

Volker Ludwig is Assistant Professor of Applied Sociology at TU Kaiserslautern. He works on labor market sociology, family sociology and social research methods, especially on methods for collecting and analyzing longitudinal data.