Measurement Error in Latent Predictors: The Role of Identification Restrictions

Stephen Jessee & Connor T. Jerzak

Department of Government, UT Austin

MPSA – Spring 2024

Latent Traits in Political Science

Latent Traits in Political Science

* Many concepts are latent, i.e., not directly observable
  $\rightsquigarrow$ Ideology, political knowledge, respect for human rights

Latent Traits in Political Science

- Many concepts are latent, i.e., not directly observable
  - ⇝ Ideology, political knowledge, respect for human rights
- Latent traits typically estimated using multiple indicators
  - ⇝ Policy views, knowledge questions, expert ratings

Latent Traits in Political Science

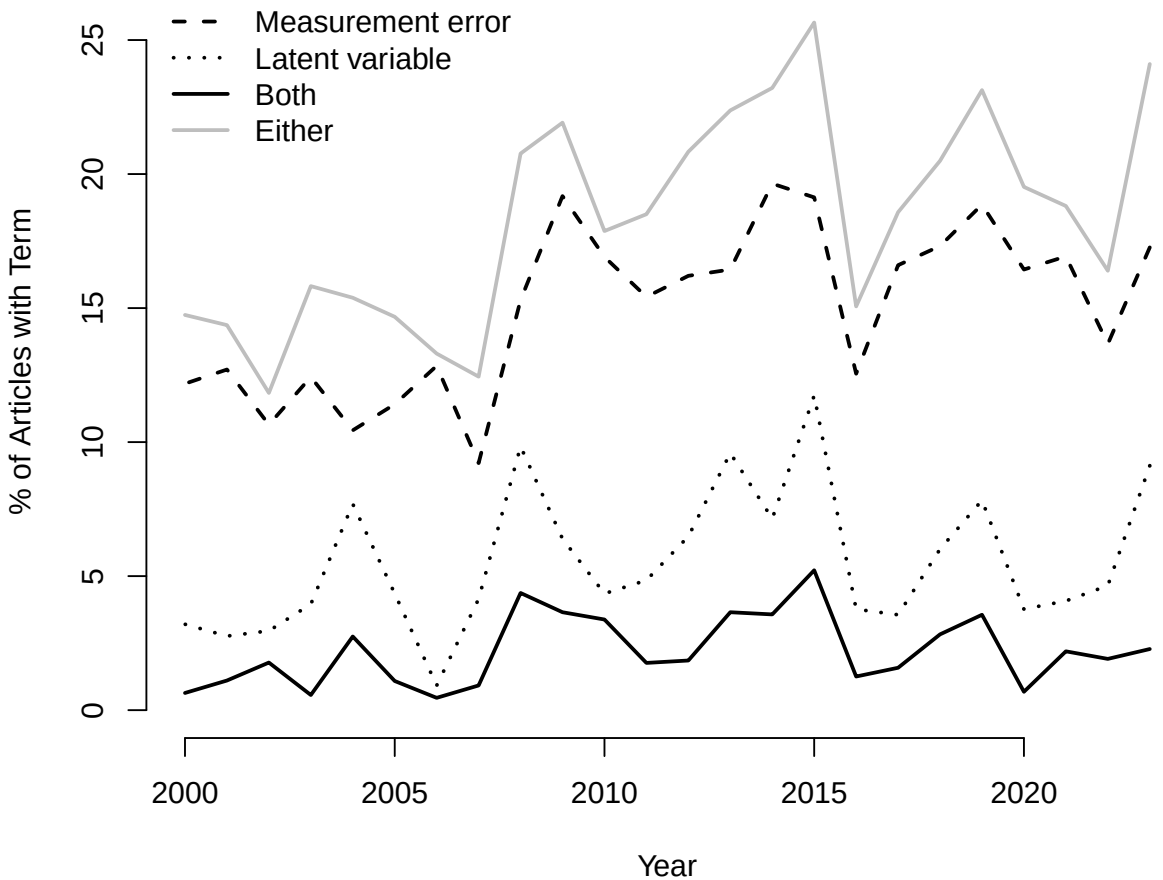
- Many concepts are latent, i.e., not directly observable
  - ⇝ Ideology, political knowledge, respect for human rights
- Latent traits typically estimated using multiple indicators
  - ⇝ Policy views, knowledge questions, expert ratings
- Latent variables have no objective scale

Latent Traits in Political Science

* Many concepts are latent, i.e., not directly observable
  ⇝ Ideology, political knowledge, respect for human rights
* Latent traits typically estimated using multiple indicators
  ⇝ Policy views, knowledge questions, expert ratings
* Latent variables have no objective scale
  - Investigator must use identifying restrictions
    ⇝ Normalization constraints (mean 0, sd 1)

Trends, Top 3+PA: “Measurement Error”/“Latent Var.”

Trends, Top 3+PA: “Measurement Error”/“Latent Var.”



- TAKEAWAY: Growing interest in measurement error; less work on measurement error in latent estimates

Measurement Error in Regression Models: Standard Case

Measurement Error in Regression Models: Standard Case

* Consider a linear model with predictor, X:

$$Y = \beta_0 + X\beta_X + \epsilon$$

Measurement Error in Regression Models: Standard Case

* Consider a linear model with predictor, X:

$$Y = \beta_0 + X\beta_X + \epsilon$$

* Standard (non-latent) measurement error setup assumes we only observe $\tilde{X} = X + U$
  - where $U \perp\!\!\!\perp X$, $\mathbb{E}[U] = 0$ and $\mathbb{V}(U) = \sigma_U^2$
  - implies $\mathbb{V}(\tilde{X}) = \sigma_X^2 + \sigma_U^2$

Measurement Error in Regression Models: Standard Case

* Consider a linear model with predictor, X:

$$Y = \beta_0 + X\beta_X + \epsilon$$

* Standard (non-latent) measurement error setup assumes we only observe $\tilde{X} = X + U$
  - where $U \perp\!\!\!\perp X$, $\mathbb{E}[U] = 0$ and $\mathbb{V}(U) = \sigma_U^2$
  - implies $\mathbb{V}(\tilde{X}) = \sigma_X^2 + \sigma_U^2$
* Regressing Y on $\tilde{X}$ (not true X),

  Standard Attenuation Bias: $\mathbb{E}[\hat{\beta}_{\tilde{X}}] = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2}\ \beta_X$

Measurement Error in Regression Models: Latent Predictor

Measurement Error in Regression Models: Latent Predictor

* With latent X we do not observe $\tilde{X} = X + U$

Measurement Error in Regression Models: Latent Predictor

- With latent X we do not observe $\tilde{X} = X + U$
- Why? We must fix latent scale:
  - Estimates are transformed $\rightsquigarrow$ mean 0, var. 1
  - Thus, we observe $\hat{X} = \frac{\tilde{X} - \text{mean}(\tilde{X})}{\text{sd}(\tilde{X})}$

Measurement Error in Regression Models: Latent Predictor

- With latent X we do not observe $\tilde{X} = X + U$
- Why? We must fix latent scale:
  - Estimates are transformed $\rightsquigarrow$ mean 0, var. 1
  - Thus, we observe $\hat{X} = \frac{\tilde{X} - \text{mean}(\tilde{X})}{\text{sd}(\tilde{X})}$
- Since $\tilde{X}$ is divided by $\text{sd}(\tilde{X}) = \sqrt{\sigma_X^2 + \sigma_U^2} \rightsquigarrow \sqrt{1 + \sigma_U^2}$

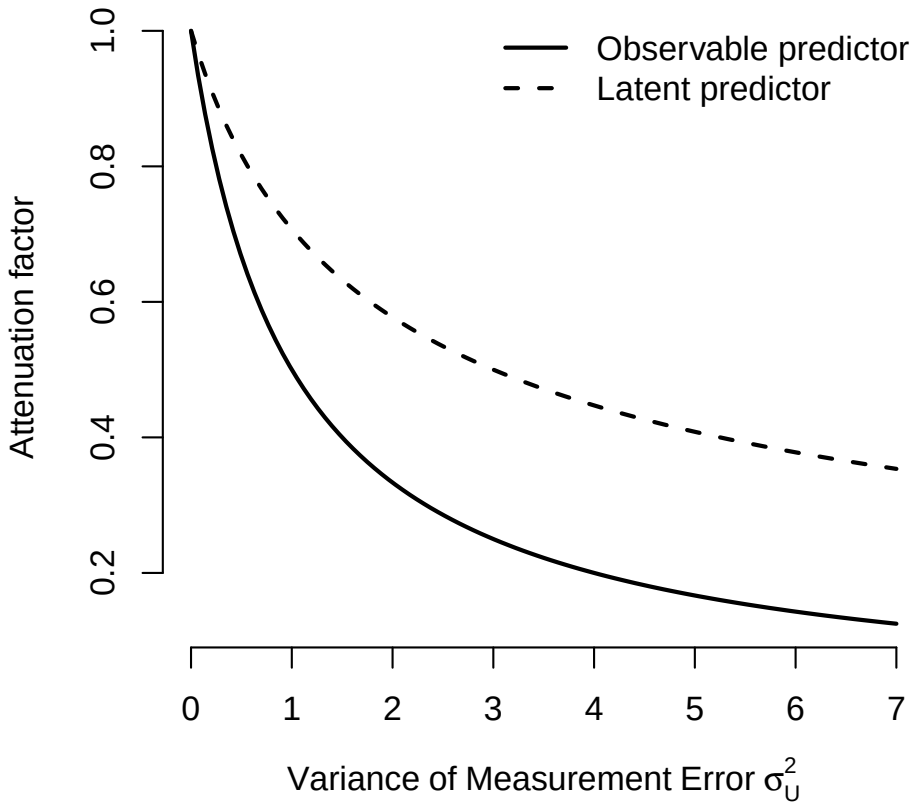
Measurement Error in Regression Models: Latent Predictor

* With latent X we do not observe $\tilde{X} = X + U$
* Why? We must fix latent scale:
  - Estimates are transformed $\rightsquigarrow$ mean 0, var. 1
  - Thus, we observe $\hat{X} = \frac{\tilde{X} - \text{mean}(\tilde{X})}{\text{sd}(\tilde{X})}$
* Since $\tilde{X}$ is divided by $\text{sd}(\tilde{X}) = \sqrt{\sigma_X^2 + \sigma_U^2} \rightsquigarrow \sqrt{1 + \sigma_U^2}$
* Therefore: Attenuation with Latent Variables:

$$\mathbb{E}[\beta_{\hat{X}}] = \mathbb{E}[\beta_{\tilde{X}}] \sqrt{1 + \sigma_U^2} = \frac{1}{1 + \sigma_U^2} \beta_X \sqrt{1 + \sigma_U^2}$$

$$= \frac{1}{\sqrt{1 + \sigma_U^2}} \beta_X$$

Attentuation Factor with Latent Space Predictor



- ATTENUATION FACTOR: $\widehat{\beta}/\beta$ (1 is unbiased)

Instrumental Variables Estimation for Measurement Error

Instrumental Variables Estimation for Measurement Error

* Standard Case with $\tilde{X}$: IV yields consistent estimates of β_X under assumptions:
  - (1) $\mathrm{Cov}(X, Z) \neq 0$; (2) $\mathrm{Cov}(Z, U) = 0$; (3) $\mathrm{Cov}(Z, \varepsilon) = 0$

Instrumental Variables Estimation for Measurement Error

* Standard Case with $\tilde{X}$: IV yields consistent estimates of β_X under assumptions:
  - (1) $\text{Cov}(X, Z) \neq 0$; (2) $\text{Cov}(Z, U) = 0$; (3) $\text{Cov}(Z, \varepsilon) = 0$
  - IV Recipe: Split indicators into 2 sets
    - Estimate X separately based on each set, producing $\hat{X}_1, \hat{X}_2$
    - Define $X \equiv \hat{X}_1, Z \equiv \hat{X}_2$

Instrumental Variables Estimation for Measurement Error

* Standard Case with $\tilde{X}$: IV yields consistent estimates of β_X under assumptions:
  - (1) $\mathrm{Cov}(X, Z) \neq 0$; (2) $\mathrm{Cov}(Z, U) = 0$; (3) $\mathrm{Cov}(Z, \varepsilon) = 0$
  - IV Recipe: Split indicators into 2 sets
    - Estimate X separately based on each set, producing $\hat{X}_1, \hat{X}_2$
    - Define $X \equiv \hat{X}_1, Z \equiv \hat{X}_2$
* **But** in Latent Case with $\widehat{X}$ IV *over*-corrects $\rightsquigarrow$ inflation bias

  $\hat{\beta}_{\mathrm{IV},\hat{X}} = \hat{\beta}_{\mathrm{IV},\tilde{X}} \sqrt{1 + \sigma_U^2}$;

  Hence $\operatorname{plim} \hat{\beta}_{\mathrm{IV},\hat{X}} = \beta_X \sqrt{1 + \sigma_U^2}$,

  where $\sqrt{1 + \sigma_U^2} > 0$ whenever $\sigma_U^2 > 0$

Illustrative Simulation Results: Ideology & Vote

Illustrative Simulation Results: Ideology & Vote

* Context: Based on 2016 CCES:
  1,000 Respondents, 30 Indicators
  - $Y_i =$ Presidential Vote (1=Trump, 0=Clinton)
  - $X_i =$ Ideology (mean 0, var 1, higher = conservative)
* Goal: *Estimate slope coefficient in regression:*

$$Y_i = \beta_0 + X_i \beta_X + \epsilon_i$$

* Indicators: P, generated from ideal point model:

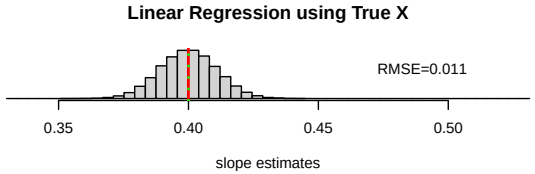
$$\Pr(P_{id} = 1) = \Phi\left(X_i \eta_d + \alpha_d\right)$$

Set true values of all parameters based on estimates from the actual data (100k Monte Carlo sims)

Illustrative Simulation Results: Ideology & Vote

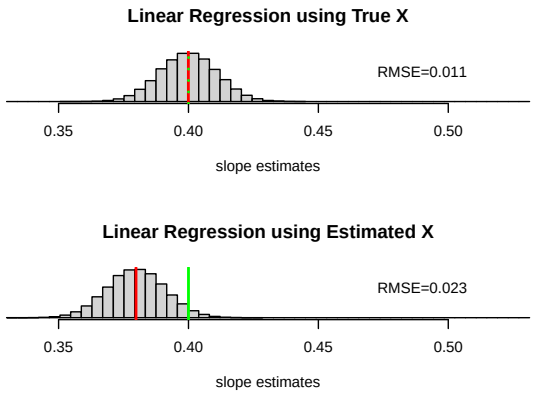
* TRUE X
  assume we knew true values for ideology (X)
  - Unbiased estimation of β_X with high precision (low RMSE)

  ...of course, we don't ever really observe the true X, but have to estimate it



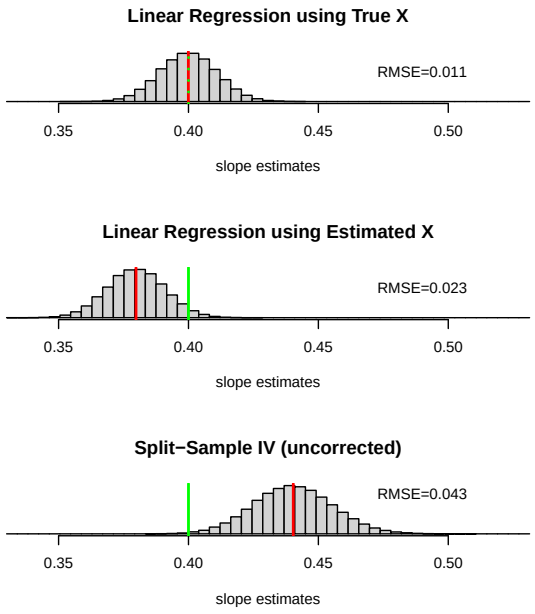
Illustrative Simulation Results: Ideology & Vote

* "Naive" Estimator
  use $\hat{X}$ as our predictor instead of the true X?
  - Attenuation bias – estimates too close to zero on average
  - RMSE more than double what it was when we knew true X



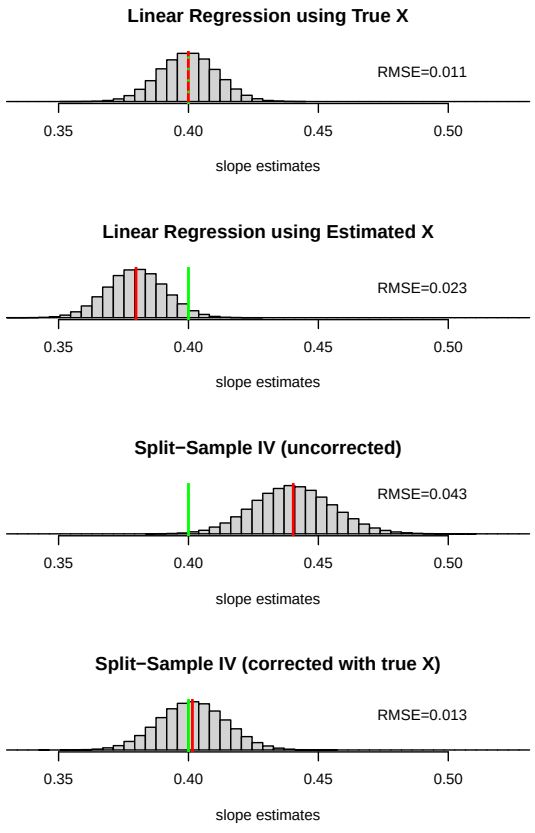
Illustrative Simulation Results: Ideology & Vote

* Split-Half IV, Uncorrected
  - Estimate $\hat{X}_1, \hat{X}_2$ each based on half of indicators
  - Define $X \equiv \hat{X}_1, Z \equiv \hat{X}_2$
- *Over*-adjusts
  - Ignores effects of identification on attenuation
  - RMSE nearly 4$\times$True X estimator, double the naive estimator



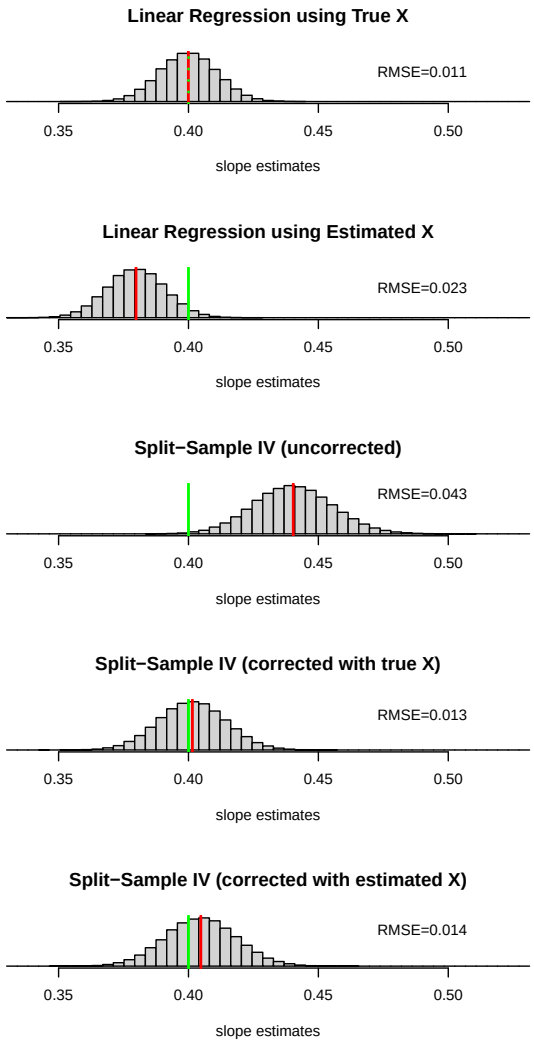
Illustrative Simulation Results: Ideology & Vote

* Split-Half IV, **Corrected** using true X
  - Same split-half IV setup
  - Now $\rightsquigarrow$ Estimate σ_U^2 with variance of $\hat{X}_1 - X$
  - Cheating – uses unknown X!
- *Very low bias, high precision*
  - RMSE barely higher than True X case



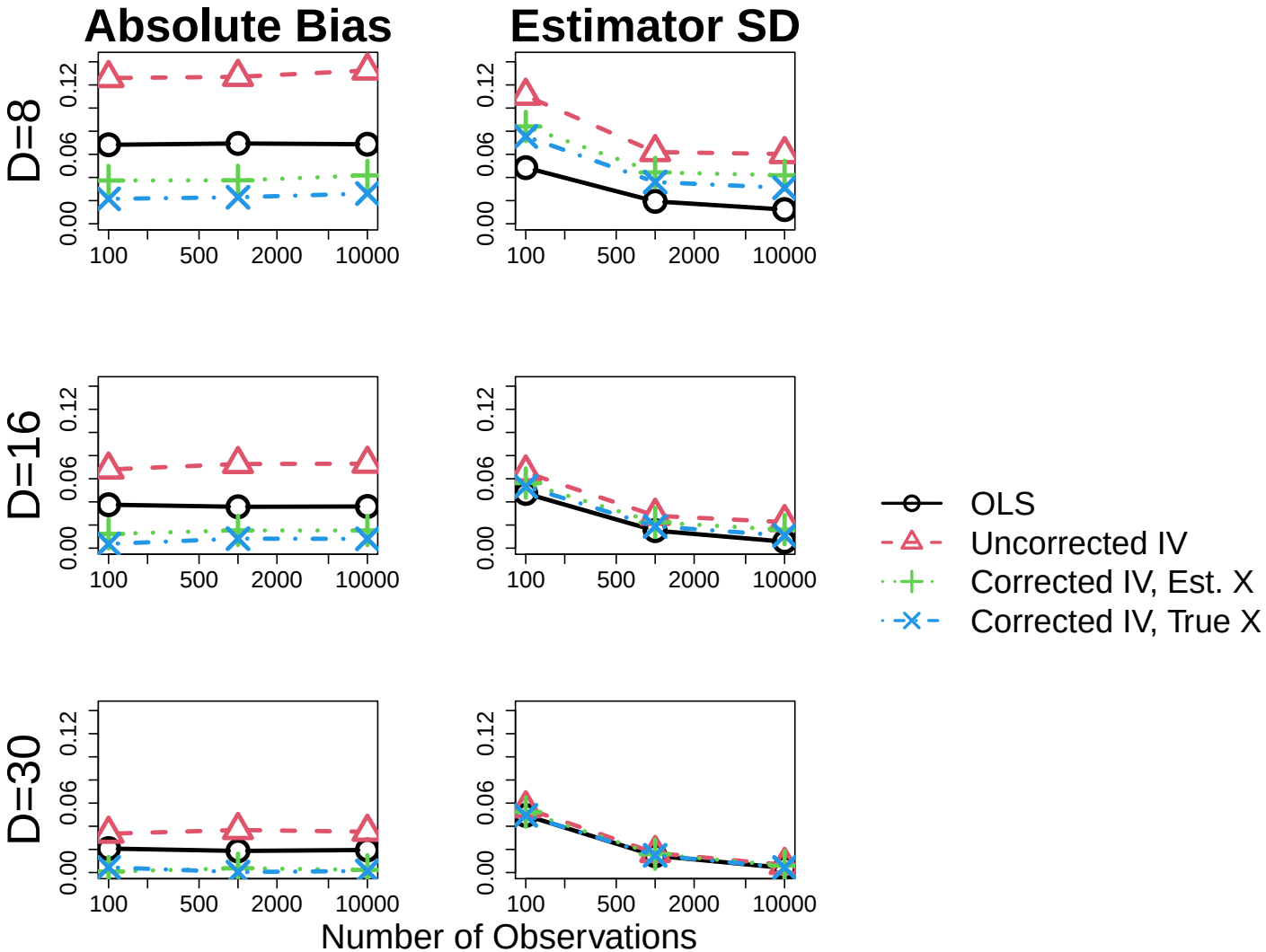
Illustrative Simulation Results: Ideology & Vote

* Split-Half IV, **Corrected** Using $\hat{X}_1, \hat{X}_2$
  - Same split-half IV setup
  - Now, estimate σ_U^2 with $\mathbb{V}(\hat{X}_1 - \hat{X}_2)/2$
  - This is **not** cheating
  - ...but does assume measurement error in $\hat{X}_1, \hat{X}_2$ is $\perp$ and has same var.
- *Very low bias, quite precise*
  - RMSE barely higher than True X case



Large-scale Simulation Results

Large-scale Simulation Results



CCES Empirical Analysis

Latent Error Dynamics in Practice?

CCES Empirical Analysis

Latent Error Dynamics in Practice?

- **Re-analyze** 2016 CCES data

CCES Empirical Analysis

Latent Error Dynamics in Practice?

- **Re-analyze** 2016 CCES data
- **Vary** n obs, D (# of indicators for estimating ideology)

CCES Empirical Analysis

Latent Error Dynamics in Practice?

- **Re-analyze** 2016 CCES data
- **Vary** n obs, D (# of indicators for estimating ideology)
- **Estimate** effect of estimated ideology on vote choice

CCES Empirical Analysis

Latent Error Dynamics in Practice?

- **Re-analyze** 2016 CCES data
- **Vary** n obs, D (# of indicators for estimating ideology)
- **Estimate** effect of estimated ideology on vote choice
- **Compare** fractional change:
  (1) corrected IV vs. OLS estimates, (2) vs. uncorrected IV

CCES Empirical Analysis

Latent Error Dynamics in Practice?

- **Re-analyze** 2016 CCES data
- **Vary** n obs, D (# of indicators for estimating ideology)
- **Estimate** effect of estimated ideology on vote choice
- **Compare** fractional change:
  (1) corrected IV vs. OLS estimates, (2) vs. uncorrected IV

Table: Percentage change in effect estimate of ideology on vote choice, corrected IV vs. OLS. Rows correspond to observation number; columns to dimensions used in estimating latent ideology.

	13	21	30
60	0.323	0.188	0.132
519	0.29	0.163	0.103
4491	0.296	0.163	0.101
38851	0.302	0.166	0.104

CCES Empirical Analysis

Latent Error Dynamics in Practice?

- **Re-analyze** 2016 CCES data
- **Vary** n obs, D (# of indicators for estimating ideology)
- **Estimate** effect of estimated ideology on vote choice
- **Compare** fractional change:
  (1) corrected IV vs. OLS estimates, **(2) vs. uncorrected IV**

Table: Percentage change in effect estimate of ideology on vote choice, corrected IV vs. uncorrected IV. Rows correspond to observation number; columns to dimensions used in estimating latent ideology.

	13	21	30
60	-0.158	-0.123	-0.102
519	-0.154	-0.122	-0.099
4491	-0.155	-0.123	-0.101
38851	-0.154	-0.124	-0.1

Conclusion

Conclusion

- Social Science Research:

Conclusion

- Social Science Research:

  Connects observables to unobservables
  via latent variable models
  - This connection relies on identification restrictions

Conclusion

- Social Science Research:

  Connects observables to unobservables via latent variable models
  - This connection relies on identification restrictions
- Our Work:
  - **Probes** how these restrictions affect measurement error dynamics
  - **Provides tools** to de-bias estimates
  - **Gathers evidence** to assess reach of this problem in practice

Conclusion

- Social Science Research:

  Connects observables to unobservables via latent variable models

  - This connection relies on identification restrictions
- Our Work:
  - **Probes** how these restrictions affect measurement error dynamics
  - **Provides tools** to de-bias estimates
  - **Gathers evidence** to assess reach of this problem in practice

For more information:

laits.utexas.edu/sjessee & ConnorJerzak.com

Appendix Slides

Democracy Analysis: Corrected IV vs. OLS

Table: Percentage change in effect estimate of democracy on growth (long difference), corrected IV vs. OLS. Rows correspond to observation number; columns to dimensions used in estimating latent democratic status.

	4	8	13	17
60	0.093	0.032	0.013	0.011
91	0.072	0.032	0.013	0.009
137	0.098	0.032	0.01	0.012
208	0.073	0.028	0.012	0.01

Democracy Analysis: Corrected vs. Uncorrected IV

Table: Percentage change in effect estimate of democracy on growth (long difference), corrected IV vs. uncorrected IV. Rows correspond to observation number; columns to dimensions used in estimating latent democratic status.

	4	8	13	17
60	-0.092	-0.047	-0.029	-0.022
91	-0.091	-0.044	-0.028	-0.022
137	-0.1	-0.044	-0.028	-0.021
208	-0.087	-0.043	-0.027	-0.021