Modeling Log Data from an Intelligent Tutor Experiment

Adam Sales[1]

joint work with John Pane & Asa Wilks

College of Education
University of Texas, Austin
RAND Corporation
Pittsburgh, PA & Santa Monica, CA

Educational Data Mining 2016

[1] Supported by IES Grant R305B1000012 and NSF Grant DRL-1420374

The Set-Up

- You just ran an experiment on an Intelligent Tutor
- It works!
  - (on average)
- You have *mounds and mounds* of log data
- Does use predict treatment effect?

The Set-Up

- You just ran an experiment on an Intelligent Tutor
- It works!
  - (on average)
- You have *mounds and mounds* of log data
- Does use predict treatment effect?

The Set-Up

- You just ran an experiment on an Intelligent Tutor
- It works!
  - (on average)
- You have *mounds and mounds* of log data
- Does use predict treatment effect?

The Set-Up

- You just ran an experiment on an Intelligent Tutor
- It works!
  - (on average)
- You have *mounds and mounds* of log data
- Does use predict treatment effect?

The Set-Up

- You just ran an experiment on an Intelligent Tutor
- It works!
  - (on average)
- You have *mounds and mounds* of log data
- Does use predict treatment effect?

Our Actual Dataset

- Cognitive Tutor Algebra I
- Effectiveness trial: ATE $\approx$ 0.2 SDs
- Problem-level usage data:
  - Which Problem/section/unit
  - time-stamp
  - # Hints
  - # Errors
- "Skips": Do students work sections in order?
- "Assistance": # Hints & # Errors
- We used only 2nd-year HS data

Our Actual Dataset

- Cognitive Tutor Algebra I
- Effectiveness trial: ATE $\approx$ 0.2 SDs
- Problem-level usage data:
  - Which Problem/section/unit
  - time-stamp
  - # Hints
  - # Errors
- "Skips": Do students work sections in order?
- "Assistance": # Hints & # Errors
- We used only 2nd-year HS data

Our Actual Dataset

- Cognitive Tutor Algebra I
- Effectiveness trial: ATE ≈ 0.2 SDs
- Problem-level usage data:
  - Which Problem/section/unit
  - time-stamp
  - # Hints
  - # Errors
- "Skips": Do students work sections in order?
- "Assistance": # Hints & # Errors
- We used only 2nd-year HS data

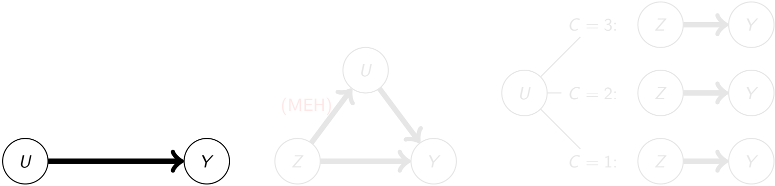
Our Actual Dataset

- Cognitive Tutor Algebra I
- Effectiveness trial: ATE $\approx$ 0.2 SDs
- Problem-level usage data:
  - Which Problem/section/unit
  - time-stamp
  - # Hints
  - # Errors
- "Skips": Do students work sections in order?
- "Assistance": # Hints & # Errors
- We used only 2nd-year HS data

Our Actual Dataset

- Cognitive Tutor Algebra I
- Effectiveness trial: ATE $\approx$ 0.2 SDs
- Problem-level usage data:
  - Which Problem/section/unit
  - time-stamp
  - # Hints
  - # Errors
- "Skips": Do students work sections in order?
- "Assistance": # Hints & # Errors
- We used only 2nd-year HS data

Modeling Usage Data



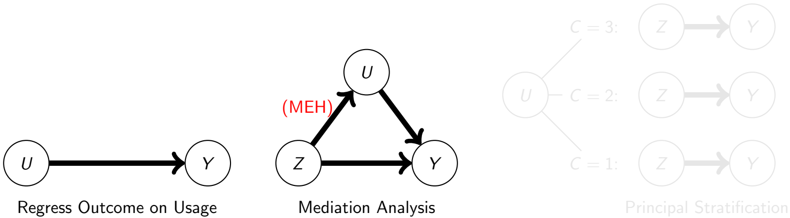
Regress Outcome on Usage

Mediation Analysis

Principal Stratification

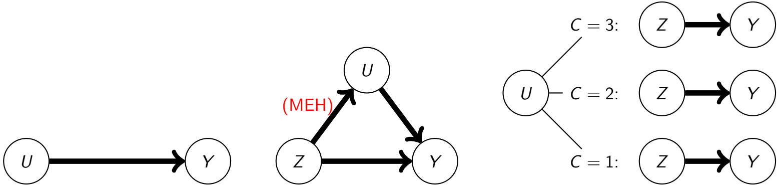
In Principal Stratification, we model the relationship between usage and treatment effect.

Modeling Usage Data



In Principal Stratification, we model the relationship between usage and treatment effect.

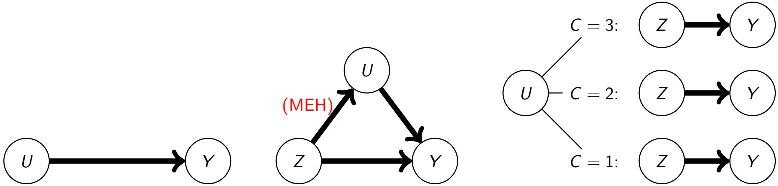
Modeling Usage Data



Regress Outcome on Usage Mediation Analysis Principal Stratification

In Principal Stratification, we model the relationship between usage and treatment effect.

Modeling Usage Data



Regress Outcome on Usage   Mediation Analysis   Principal Stratification

In Principal Stratification, we model the relationship between usage and treatment effect.

Skipping Data

- The Hypothesis: Order Matters
- But sometimes teacher have different priorities than Carnegie Learning
  - Want students to all work on similar things
  - Want students to pass a state standardized test
  - Don't believe student mastery model
- So... they move students to a new section of their choice.

What we have:

Let

$$S_i = \begin{cases} 1 \text{ if student has skipped} \\ 0 \text{ if not} \end{cases} \quad (1)$$

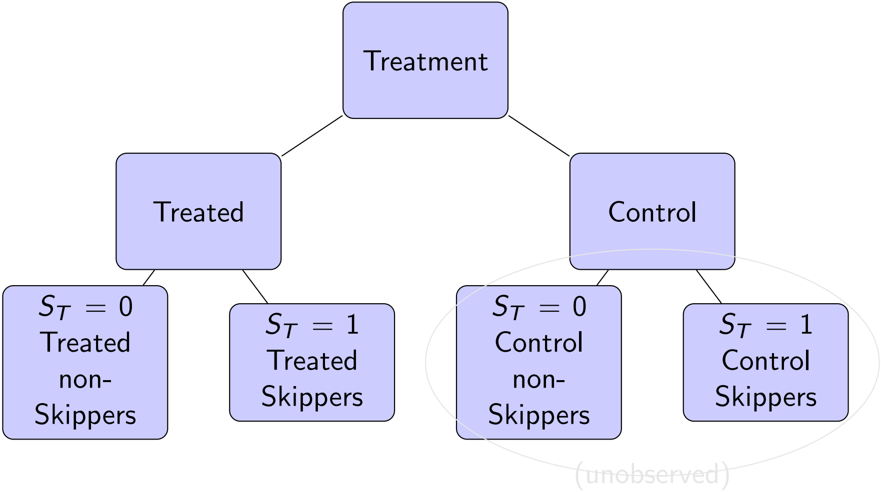
S_i is only defined for the treatment group.

Big Idea: *Potential S*

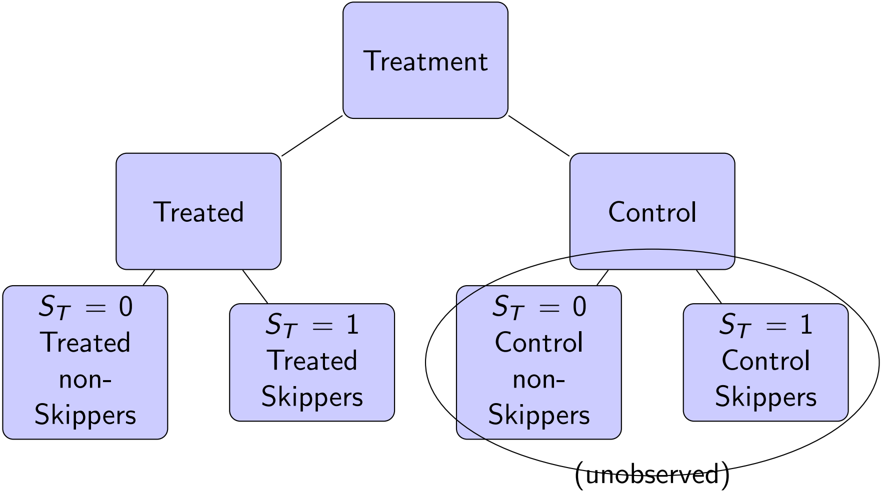
Frangakis and Rubin (2002); Page (2012); Feller et al. (2016)

- S is only defined for the treatment group
- But: *counterfactual* S is defined for everyone
- What would your S be if you were assigned to treatment?
- Call it: S_T
- S_T defines types of students

Principal Strata



Principal Strata



What We Want

μ: Average Posttest

Treatment Status (Z)

	$Z=1$	$Z=0$	
$S_T=0$	$\mu_{Z=1S=0}$	$\mu_{Z=0S=0}$	$\rightarrow \mu_{Z=1S=0}-\mu_{Z=0S=0}=\tau_0$
$S_T=1$	$\mu_{Z=1S=1}$	$\mu_{Z=0S=1}$	$\rightarrow \mu_{Z=1S=1}-\mu_{Z=0S=1}=\tau_1$

Two "Principal Treatment Effects":

$$\tau_0=\mu_{Z=1S=0}-\mu_{Z=0S=0}$$
$$\tau_1=\mu_{Z=1S=1}-\mu_{Z=0S=1}$$

What is the quantity $\tau_1-\tau_0$?

What We Want

μ: Average Posttest

Treatment Status (Z)

	$Z=1$	$Z=0$	
$S_T=0$	$\mu_{Z=1S=0}$	$\mu_{Z=0S=0}$	$\rightarrow \mu_{Z=1S=0} - \mu_{Z=0S=0} = \tau_0$
$S_T=1$	$\mu_{Z=1S=1}$	$\mu_{Z=0S=1}$	$\rightarrow \mu_{Z=1S=1} - \mu_{Z=0S=1} = \tau_1$

Two "Principal Treatment Effects":

$$\tau_0 = \mu_{Z=1S=0} - \mu_{Z=0S=0}$$
$$\tau_1 = \mu_{Z=1S=1} - \mu_{Z=0S=1}$$

What is the quantity $\tau_1 - \tau_0$?

What We Have

Treatment Status (Z)

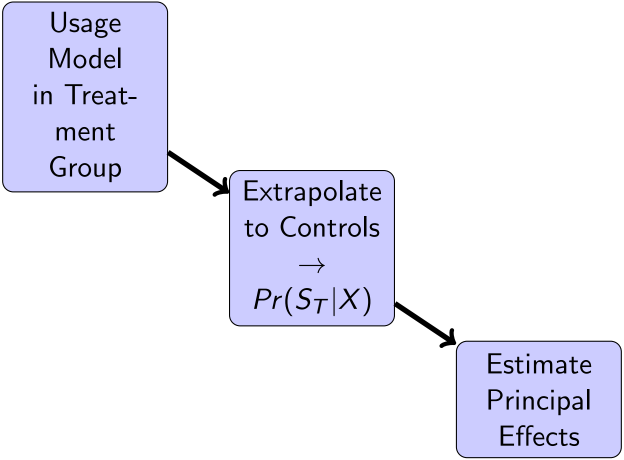
	$Z=1$	$Z=0$
$S_T=0$	$\mu_{Z=1S=0}$	$\mu_{Z=0}$
$S_T=1$	$\mu_{Z=1S=1}$	

Problem:
Decompose $\mu_{Z=0} \rightarrow \mu_{Z=0S=0} \& \mu_{Z=0S=1}$
Decompose Control group into Skippers, non-skippers

But.. But...

- S_T is only observed in the treatment group
- BUT: We know what the skippers look like
- i.e. We can predict S_T with covariates X
- Then extrapolate the model to the control group
  - (this works because of randomization)
- Estimate $Pr(S_{Ti} = 1|X_i)$ for every member of the control group

The Process


Usage Model in Treatment Group
Extrapolate to Controls $\rightarrow$ $Pr(S_T|X)$
Estimate Principal Effects

Outcome Analysis: Normal Mixture Model

- Treated subjects with $S_T = 0$:

$$Y \sim \mathcal{N}(\mu_{Z=1S=0}, \sigma_{Z=1S=0})$$

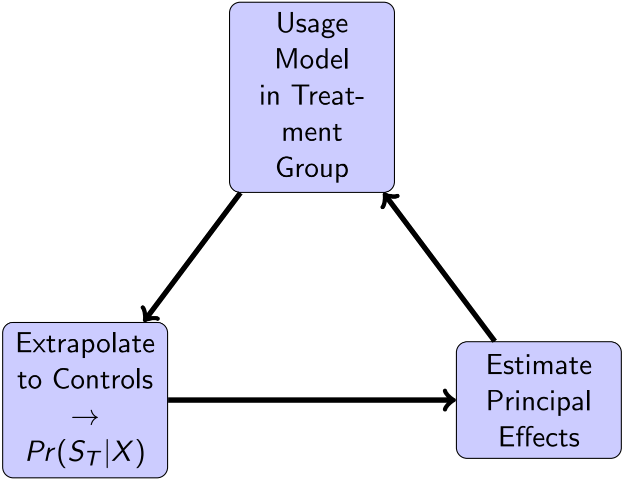
- Treated Subjects with $S_T = 1$:

$$Y \sim \mathcal{N}(\mu_{Z=1S=1}, \sigma_{Z=1S=1})$$

- Control Subjects

$$\begin{aligned} Y \sim & Pr(S_T = 0|X)\mathcal{N}(\mu_{Z=0S=0}, \sigma_{Z=0S=0}) \\ & + Pr(S_T = 1|X)\mathcal{N}(\mu_{Z=0S=1}, \sigma_{Z=0S=1}) \end{aligned}$$

Estimate Everything with MCMC



The Usage Model

Multilevel Logistic Regression

	mean	sd	95%CI
grade 10	0.73	0.53	(-0.37,1.74)
grade 11	1.31	0.78	(-0.25,2.85)
grade 12	-2.49	2.13	(-7.15,1.2)
grade 14	-0.25	2.89	(-5.89,5.58)
raceASIAN / PACIFIC ISLANDER	-0.63	1.38	(-3.29,2.02)
raceBLACK NON-HISPANIC	-1.01	1.2	(-3.33,1.55)
raceHISPANIC	-1.32	1.22	(-3.55,1.31)
raceOTHER RACE / MULTI-RACIAL	-0.83	1.54	(-3.9,2.08)
raceWHITE NON-HISPANIC	0.44	1.13	(-1.63,2.86)
sexM	0.27	0.24	(-0.18,0.73)
spec_speced1	-2.98	0.78	(-4.56,-1.54)*
spec_gifted1	-1.71	0.57	(-2.83,-0.64)*
spec_esl1	1.06	0.75	(-0.42,2.51)
frl1	-0.1	0.31	(-0.69,0.51)
pretest	0.45	0.14	(0.18,0.72)*
x_spec_giftedMIS1	0.55	1.25	(-1.97,2.85)
x_gradeMIS1	-1	0.89	(-2.77,0.7)
x_raceMIS1	-0.81	0.92	(-2.63,0.91)
x_sexMIS1	0.43	0.91	(-1.37,2.15)
x_frlMIS1	0.09	0.52	(-0.99,1.08)

Also used the same covariates in outcome regression.

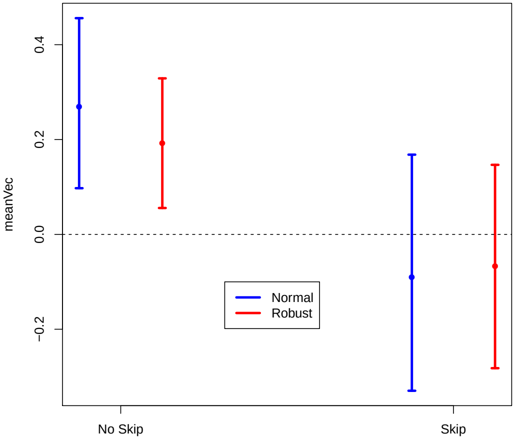
Teacher Level

	mean	sd	95% CI
% ESL	-0.39	2.63	(-5.54,4.96)
avg. pretest	1.47	0.78	(0.03,3.1)*

Robustness Check

- Replace Normal Distribution with t-distribution
- Allows for outliers

Results



But What Does it Mean?

- Two types of students:
    1. Those who *would* skip
    2. Those who wouldn't
- Skippers:
    - have higher pretest scores
    - are less likely to be gifted (given pretest)
    - are less likely to need Special Ed
    - Come from classrooms with higher avg pretest
- Is it the skipping that's driving this?
- Or is it something about the students? (low effects for high performers?)
- Or something about the teachers?

Does Order Matter?

A couple hypotheses:

- Students who work sections in the order they were presented learn more from the tutor
- Teachers who let their students skip meddle more with the tutor in general

Causality in PS

- What's causal:
  - The treatment effects within strata are causal
  - Identification is from Randomization
  - No untestable exogeneity assumptions
- What isn't (necessarily) causal:
  - Differences between treatment effects
  - Usage is not randomized

Now for something more complicated...

Causality in PS

- What's causal:
  - The treatment effects within strata are causal
  - Identification is from Randomization
  - No untestable exogeneity assumptions
- What isn't (necessarily) causal:
  - Differences between treatment effects
  - Usage is not randomized

Now for something more complicated...

Causality in PS

- What's causal:
  - The treatment effects within strata are causal
  - Identification is from Randomization
  - No untestable exogeneity assumptions
- What isn't (necessarily) causal:
  - Differences between treatment effects
  - Usage is not randomized

Now for something more complicated...

Assistance

CTAI gives students two forms of assistance on particular problems:

- Hints
- Error feedback

The Data:

$$A_{ip} = \begin{cases} 1 \text{ if student } i \text{ got assistance on problem } p \\ 0 \text{ if not} \end{cases}$$

"an indicator of the extent to which students struggle to complete problems" Ritter et al. (2013)

Do students who need assistance more often have higher or lower effects?

Assistance

CTAI gives students two forms of assistance on particular problems:

- Hints
- Error feedback

The Data:

$$A_{ip} = \begin{cases} 1 \text{ if student } i \text{ got assistance on problem } p \\ 0 \text{ if not} \end{cases}$$

"an indicator of the extent to which students struggle to complete problems" Ritter et al. (2013)

Do students who need assistance more often have higher or lower effects?

Latent Principal Strata

- Model problem-level data:

$$Pr(A_{ip} = 1) = invLogit(\alpha_i + \beta_s)$$

- Extract α_i
  - i's propensity to need assistance
  - "assistance score"
- Model α_i:

$$\alpha_i = X_i^T \beta + \text{(nested error terms)}$$

- Extrapolate to control group

Latent Principal Strata

- Model problem-level data:

$$Pr(A_{ip} = 1) = invLogit(\alpha_i + \beta_s)$$

- Extract α_i
  - i's propensity to need assistance
  - "assistance score"
- Model α_i:

$$\alpha_i = X_i^T \beta + \text{(nested error terms)}$$

- Extrapolate to control group

Latent Principal Strata

- Model problem-level data:

$$Pr(A_{ip} = 1) = invLogit(\alpha_i + \beta_s)$$

- Extract α_i
  - i's propensity to need assistance
  - "assistance score"
- Model α_i:

$$\alpha_i = X_i^T \beta + \text{(nested error terms)}$$

- Extrapolate to control group

Latent Principal Strata

- Model problem-level data:

$$Pr(A_{ip} = 1) = invLogit(\alpha_i + \beta_s)$$

- Extract α_i
  - i's propensity to need assistance
  - "assistance score"
- Model α_i:

$$\alpha_i = X_i^T \beta + \text{(nested error terms)}$$

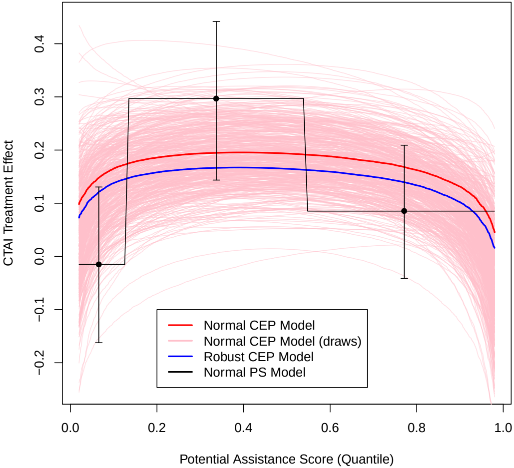
- Extrapolate to control group

Who needs assistance?

(Only included significant predictors)
(This is bad practice)
(Sorry)

	mean	sd	95% CI
pretest	-0.09	0.03	(-0.15,-0.03)*
lag2_math_score	-0.123	0.016	(-0.155,-0.089)*
lag1_math_score	-0.120	0.016	(-0.152,-0.088)*
sexM	-0.15	0.02	(-0.19,-0.1)*
spec_speced1	0.16	0.04	(0.07,0.25)*

Results



0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
CTAI Treatment Effect
0.0
0.2
0.4
0.6
0.8
1.0
Potential Assistance Score (Quantile)
Normal CEP Model
Normal CEP Model (draws)
Robust CEP Model
Normal PS Model

Interpretation

- Is the relationship causal?
    - If you're too quick to ask for a hint, make an error, you're not working hard enough to experience an effect
    - Assistance is part of the CTAI mechanism, but there's a sweet spot.
- Or not
    - Students who are insufficiently prepared for CTAI need more.
    - If you never need assistance, it's too easy for you
    - Another student characteristic?

Interpretation

- Is the relationship causal?
  - If you're too quick to ask for a hint, make an error, you're not working hard enough to experience an effect
  - Assistance is part of the CTAI mechanism, but there's a sweet spot.
- Or not
  - Students who are insufficiently prepared for CTAI need more.
  - If you never need assistance, it's too easy for you
  - Another student characteristic?

Model Misspecification

- Wrong Models beget wrong results
  - Usage Model (logistic? Linear?)
  - Outcome Model (Normal? Linear?)
  - Treatment Model (Quadratic?)
- Some solutions:
  - Try different models
  - Check model fit
  - Nonparametrics
- "All models are false, some models are useful"

Model Misspecification

- Wrong Models beget wrong results
  - Usage Model (logistic? Linear?)
  - Outcome Model (Normal? Linear?)
  - Treatment Model (Quadratic?)
- Some solutions:
  - Try different models
  - Check model fit
  - Nonparametrics
- "All models are false, some models are useful"

Model Misspecification

- Wrong Models beget wrong results
  - Usage Model (logistic? Linear?)
  - Outcome Model (Normal? Linear?)
  - Treatment Model (Quadratic?)
- Some solutions:
  - Try different models
  - Check model fit
  - Nonparametrics
- "All models are false, some models are useful"

Other Stuff

- Use the same covariates in usage and outcome model?
- Fitting algorithm (i.e. MCMC) work properly?
- Do variables mean what you think they do?

Moral of the story:
This is worthwhile, but proceed with caution!

This is worthwhile!

- Treatment Effects driven by randomization!
- Estimate Effects without assuming exogeneity
- Difficult assumptions are testable

Proceed with caution!

- Must tailor analysis to data
- Do lots of specification checks

Moral of the story:
This is worthwhile, but proceed with caution!
This is worthwhile!

- Treatment Effects driven by randomization!
- Estimate Effects without assuming exogeneity
- Difficult assumptions are testable

Proceed with caution!

- Must tailor analysis to data
- Do lots of specification checks

Moral of the story:
This is worthwhile, but proceed with caution!
This is worthwhile!

- Treatment Effects driven by randomization!
- Estimate Effects without assuming exogeneity
- Difficult assumptions are testable

Proceed with caution!

- Must tailor analysis to data
- Do lots of specification checks

Future Work

- Improve existing models
  - Non-parametric options
  - Better IRT model for Assistance
- Fancier EDM
  - Cluster log data?
  - Better motivated effect models?
  - Longitudinal modeling?
  - Give us ideas! Please!

Bibliography

Avi Feller, Todd Grindal, Luke W Miratrix, and Lindsay Page. Compared to what? variation in the impact of early childhood education by alternative care-type settings. *Annals of Applied Statistics*, 2016. URL `http://www.e-publications.org/ims/submission/AOAS/user/submissionFile/22608?confirm=7a18d69c`. in press.

Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

Lindsay C. Page. Principal stratification as a framework for investigating mediational processes in experimental settings. *Journal of Research on Educational Effectiveness*, 5(3):215–244, 2012.

Steven Ritter, Ambarish Joshi, Stephen Fancsali, and Tristan Nixon. Predicting standardized test scores from cognitive tutor interactions. In *EDM*, pages 169–176, 2013.

Questions?
Comments?
Ideas?

Thank you!!

Next Step: Full-On Latent Variables?

- High-D usage data
- Are there clusters?
- Do treatment effects vary by cluster?
- Who knows?

Methodological Problem: Modeling Usage Data

One idea: Regress posttest scores on usage data
Answers question: does usage predict posttest scores.
Some problems:

- Nothing is causal
- Don't use experimental design
- Don't use control group
- Doesn't speak to causal mechanisms—what's driving the effect

What about "Mediation Analysis?"

- Need to assume no mediator-outcome confounding
- Usage is collinear with treatment

There is another way...

Methodological Problem: Modeling Usage Data

One idea: Regress posttest scores on usage data
Answers question: does usage predict posttest scores.
Some problems:

- Nothing is causal
- Don't use experimental design
- Don't use control group
- Doesn't speak to causal mechanisms—what's driving the effect

What about "Mediation Analysis?"

- Need to assume no mediator-outcome confounding
- Usage is collinear with treatment

There is another way...

Methodological Problem: Modeling Usage Data

One idea: Regress posttest scores on usage data
Answers question: does usage predict posttest scores.
Some problems:

- Nothing is causal
- Don't use experimental design
- Don't use control group
- Doesn't speak to causal mechanisms—what's driving the effect

What about "Mediation Analysis?"

- Need to assume no mediator-outcome confounding
- Usage is collinear with treatment

There is another way...