Simulated Instruments for Addressing Descriptive and Causal Policy Questions

Vivian C. Wong
David Martin
University of Virginia

Coady Wing
Indiana University

Research Context

- Over the last 20 years of education reform, the major questions have been about how to hold schools "accountable" for student achievement, and whether accountability policies "work."
  - Pundits criticize NCLB as a "one-size-fits-all" approach to improving student achievement
  - No Child Left Behind replaced by Every Student Succeeds Act

- Even now, empirical research on NCLB has been challenged
  - Unclear how states responded to NCLB federal mandate?
  - Unclear the full impact of NCLB?

Overview of approach

- Proposed method provides descriptive information about implementation of state policies
  - Synthesizes states' decisions about a set of policies into a single quantitative measure
  - Provides a quantitative summary for each state, for each year of the policy
  - Is independent of population characteristics of the state (helpful for impact analysis!)
- Measure helps address questions such as:
  - How did states implement policies (for a specific year)? How much variation was there across states and over time?
  - What is the impact of states' implementation of the policy on outcomes?
- The measure is applied to NCLB context, but may work in other contexts
  - "Simulated instruments" originally introduced to describe states' marginal tax rates
  - Gruber and colleagues applied approach to examine effects of Medicaid expansion
  - We are using to describe and examine state pre-K eligibility rules

Under NCLB, schools were held “accountable” based on whether subgroups and schools met states’ annual Adequate Yearly Progress (AYP) Requirements

Requirements: Percent Proficiency, Participation and Graduation rates, Safe Harbor, and Confidence Interval, and Confidence Interval around Safe Harbor Targets

Schools either:

Met/Exceeded requirements and made AYP; or

Failed requirements and failed AYP

Variation in State Accountability Rules

Snapshot of States AYP rules for 2007-08				
	Pennsylvania	Alaska	Tennessee	Texas
Participation requirement	95%	95%	95%	95%
Minimum subgroup size	40	20	45	50
State AMO				
Elem Math	56	66.1	86	50
Elem ELA	63	77.2	89	60
Confidence Interval (CI)	95%	99%	95%	No
Safe Harbor (SH) Rule	Yes	Yes	Yes	Yes
SH-CI	75%	75%	No	No
Actual AYP School Failure Rates	28%	41%	20%	15%

Implementation of States' Accountability Policies

- States differed on accountability "stringency" under NCLB
  - Stringent rules made it harder for schools to meet annual accountability requirements
  - Less stringent rules made it easier for schools to make AYP rules

- One (inferior) option: Use the percent of schools in each state that failed AYP as stringency measure for the year

	Pennsylvania	Alaska	Tennessee	Texas
Actual AYP School Failure Rates	28%	41%	20%	15%

Actual % AYP School Failure Rate as a State Stringency Measure

- But states' AYP failure rates depend on the difficulty of states' standards *and* the population characteristics of the state
- A "good" measure of state implementation stringency separates states' policies from characteristics of schools and students in the state

Simulated AYP Failure Rates as a Stringency Measure

Snapshot of States AYP rules for 2007-08				
	PA	AK	TN	TX
Participation requirement	95%	95%	95%	95%
Minimum subgroup size	40	20	45	50
State AMO				
Elem Math	56	66.1	86	50
Elem ELA	63	77.2	89	60
Confidence Interval (CI)	95%	99%	95%	No
Fixed sample of schools' simulated failure rates	*27%*	*54%*	*70%*	*33%*

- Begin by taking a fixed sample of schools and their "input characteristics"
- Ask: *"What proportion of the fixed sample of schools would fail AYP if the same schools were held to accountability standards in other states?"*

Now:

- Differences in *simulated* stringency rates are based on differences in state rules, and
- *Not* on differences in state characteristics
- High simulated failure rates mean rules are more stringent, lower failure rates mean its easier for schools to make AYP

Calculating *Simulated* School AYP Failure Rates*

1. Using publicly available information, code all AYP policies for every state and year between 2003 and 2011
2. Using database of all AYP policies, construct an "AYP calculator" that determines whether a school fails AYP in a particular state and year
3. Feed "fixed baskets" of students/schools into calculator to construct measures of AYP stringency for each state and year

Basic idea: For a fixed sample of schools, what fraction of schools would meet AYP standards across different states and years?

* Similar to method used by Gruber & Simon (2007) to evaluate Medicaid expansion on crowd-out effects

Incorporating State Test Difficulty in Measure

- Evidence of variation in test difficulty across states and time, especially compared to national benchmark NAEP (NCES State Mapping reports) (Taylor et al. 2010)

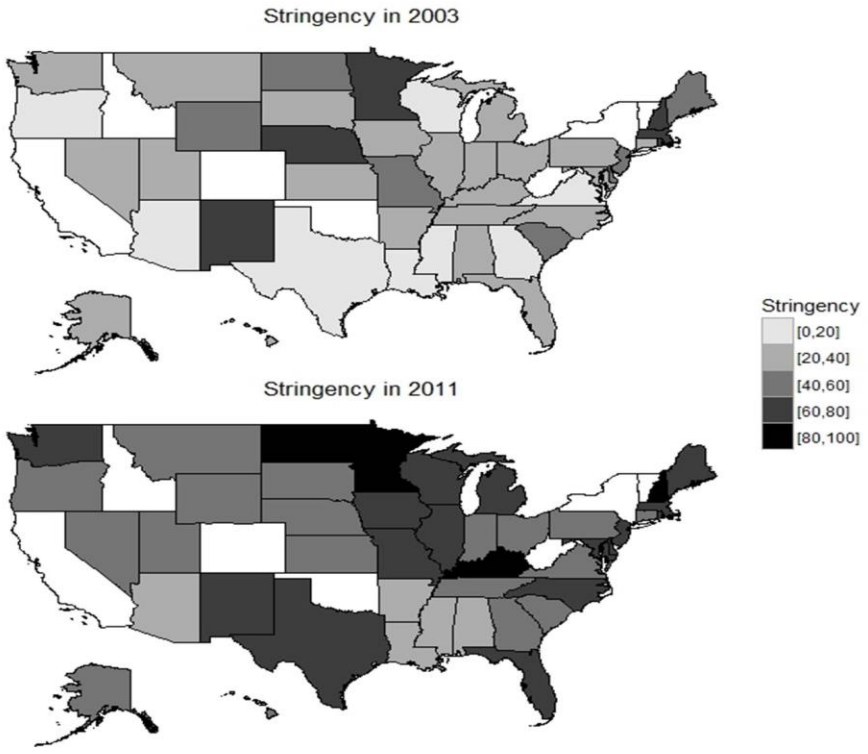
- NCES maps state proficiency cutoffs onto NAEP scale scores for 4^{th} and 8^{th} grade students

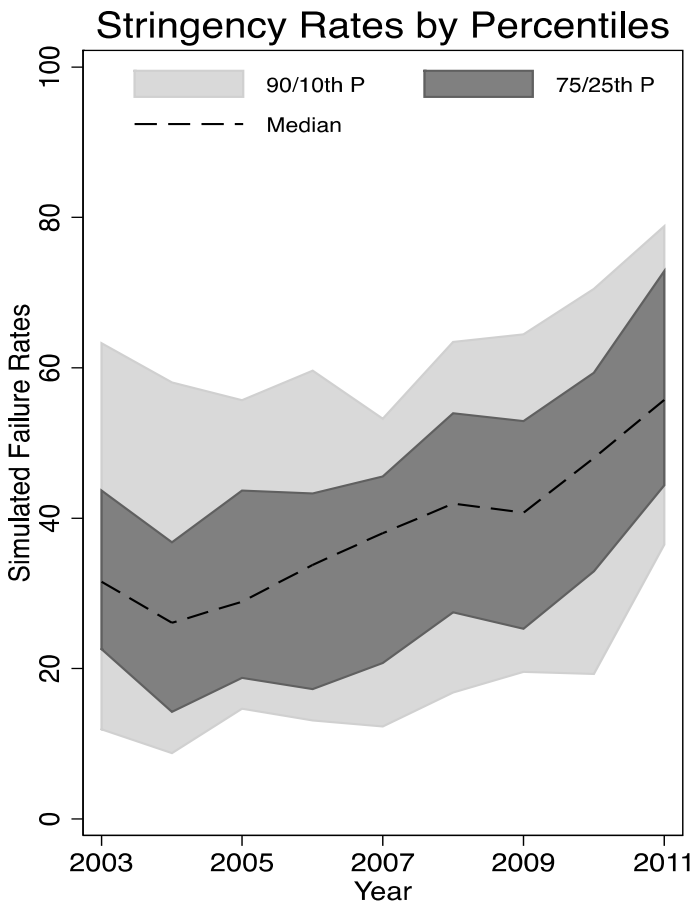
- We define the "fixed sample" as students in the NAEP sample and compare their NAEP scale scores to NAEP equivalent state cutoff scores
  - States with easier tests have lower NAEP equivalent cutoff scores
  - States with harder tests have higher NAEP equivalent cutoff scores
- Now, simulated stringency rates incorporate state accountability rules & test difficulty

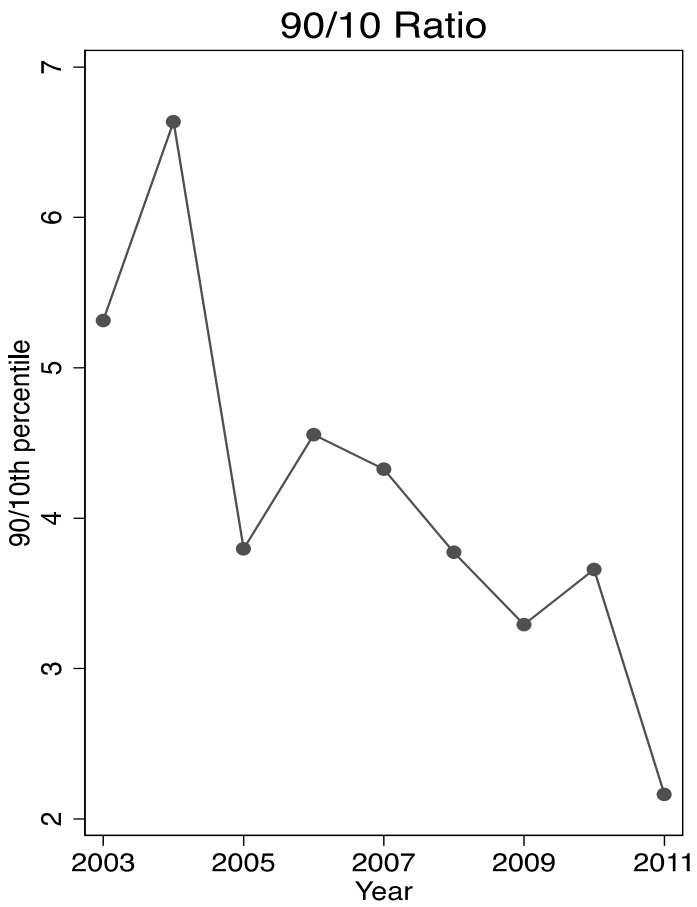
Advantages and Disadvantages of Simulated Failure Rates as Implementation Measure

- Advantages of Measure
  1. Synthesizes states' decisions about accountability policies into a single quantitative measure of stringency
  2. Provides an annual measure for each state's accountability rules
  3. Is independent of population characteristics of the state (useful for causal analysis!)
- Limitations of Measure
  1. May not account for all AYP rules that are unobserved
     - Validation checks indicate calculator performs well
  2. Measure may be sensitive to characteristics of fixed sample
     - Check results with alternative fixed samples

Under NCLB, accountability stringency ratcheted up rules but also became less discrepant







Using the Stringency Measure for Impact Evaluations

Experimental ideal

- Random assignment stringency rates to states

$$Y_{st} = \beta_0 + \beta_1 Stringency_{st} + \epsilon_{st}$$

Where, *Y* is schools' actual AYP failure rates for state *s* at time *t*.

Combine Measure with Differences in Differences Strategy

- Simulated failure rates ensure that population characteristics (schools and students) are independent from state accountability stringency
- But other factors could be related to states' adoption of accountability policies, and their outcomes (e.g. states may have a strong tradition of teacher unions)
- Settle for differences-in-differences approach with the simulated stringency rate, and

$$Y_{ist} = b_1 \overline{A}_{st} + X_{st} b + q_s + d_t + e_{ist}$$

state and year fixed effects, and a vector of time varying covariates

Advantages and Limitations

- Advantages
  - The strength here is the introduction of quantitative measure that provides state-by-year summaries of policy decisions
  - Helps that the measure is independent of population characteristics of state
  - State and Year fixed effects address other omitted confounders that do not change differentially across states and time
- Disadvantages
  - Assumptions for traditional “differences-in-differences” still apply
    - States would follow a “common path” through time absent of variations in accountability stringency
  - Similar checks as regular DiD: Inclusion of time varying covariates, balancing tests, assess pre-intervention trends, time varying treatment effects

Extensions

- Gruber and colleagues applied method for looking at expansion of Medicaid eligibility on a host of different outcomes
- Method maybe useful in other areas of education, but has not been applied
  - Example: We are currently developing a calculator for examining the impacts of state preK expansion on labor market outcomes
- Method requires careful documentation, coding, and recording of state policies across time
  - Coding preK eligibility policies much more challenging than NCLB context
- Availability of administrative data of state rules may provide many opportunities for describing, understanding, and evaluating policy impacts!

Extra Slides

Issue 1: Incorporating Test Difficulty in the Measure

- NCES maps state proficiency cutoffs onto NAEP scale scores for 4th and 8th grade students
- We define the “fixed sample” as students in the NAEP sample and compare their NAEP scale scores to NAEP equivalent state cutoff scores
  - States with easier tests have lower NAEP equivalent cutoff scores
  - States with harder tests have higher NAEP equivalent cutoff scores
- Now, simulated stringency rates incorporate state accountability rules & test difficulty

Issue 2: Does Not Accurately Reflect All Accountability Rules

The calculator:

- Accounts for all AYP rules about:
  - Test difficulty, participation rates, proficiency thresholds, other academic indicators (e.g. attendance, graduation, writing, and science), minimum subgroup size, confidence intervals, safe harbor, confidence intervals around safe harbor, multiyear averages
- But does not *yet* account for:
  - Growth models, performance indexes, alternative/modified tests for students with disabilities and Limited English Proficiency students
    We currently omit 7 states from the analyses

Validation of AYP Calculator

- For AYP calculator to work, it should mimic state AYP policies accurately for determining whether schools make AYP or not
- We validate our calculator by feeding *actual state populations of schools* through the calculator, and comparing our predicted pass rates to states' reports of actual pass rates
- We have done this for Pennsylvania and Texas for two years each. More validations states to come…

	Actual AYP Pass Rates	**Predicted AYP Pass Rates Based on our AYP Calculator**
Pennsylvania (2004)	86%	86%
Pennsylvania (2008)	72%	73%
Texas (2004)	83.4%	84.8%
Texas (2008)	66.1%	64.2%

Issue 3: Check robustness of measure and results to alternative fixed samples

- Main results are from the NAEP fixed sample
  - Allows us to incorporate test difficulty in stringency measure (imperfectly)
  - NAEP includes national representative sample of schools
- But there are limitations of the NAEP fixed sample
  - Stringency measure based on 4th and 8th grade standards only
  - NAEP equivalent scale scores not available every year, so we interpolated when possible

- Alternative sample: Pennsylvania schools
  - PA schools included heterogeneous samples with inputs needed for calculator
  - Allowed us to include all grades (including HS) in the fixed sample
  - No need to interpolate proficiency standards for states
- Limitations of PA fixed sample
  - PA fixed sample does not address differences in test difficulty across states