RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Distal Student Outcome Measures & Preparation program Effectiveness

Bruce D. Baker
Professor
Dept. of Educational Theory, Policy and Administration
Graduate School of Education
Rutgers University
Bruce.baker@gse.rutgers.edu

RUTGERS

Summary of Issues

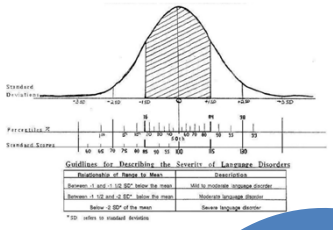
- General Reliability/Validity concerns regarding estimates at individual teacher level (classroom level)
- Attribution of preparation to teacher in the field
- Small share of all graduates & practicing teachers to whom the metrics apply
- Permeable state borders & incompatible/incomplete data systems (& varied outcome measures)
- **Non-random distribution of graduates**
  - **& low quality covariates**
- **Low quality data & models adopted in practice by states**
- **Bad resulting incentives/accountability!**

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

General Issues in Reliably & Validity of Estimating Individual Teacher "Effect" (on student test score differences)

RUTGERS

Distilling Signal from Noise



Total Variation

Unknown & Seemingly Unpredictable Error (Random)

Predictable Variation (Persistent Effect/Stable Component)

Attributable to Teacher?

Attributable to other Persistent Conditions?

A small portion of which might emanate from the preparation program?

Difficult if not implausible to accurately parse

Including who/where you teach!

RUTGERS

VAM Reliability/Validity Concerns

- Different years
  - Many year-over-year correlations below .20 (only some around .4 to .5)
- Different tests
  - Different tests of same subject reveal different results (correlation similar to year over year)
- Different kids (same teacher & year)
  - Same teacher across sections in same year correlation similar to year over year (many under .2)
- Spillover
  - Correlation across teams of teachers working with same kids
- Seasonality
  - Spring-Fall gains as big (in reading) as Fall-Spring (Gates MET)

RUTGERS

How Clear are the Signals?

- Haertel:
  - My first conclusion should come as no surprise: Teacher VAM scores should emphatically *not* be included as a substantial factor with a fixed weight in consequential teacher personnel decisions. The information they provide is **simply not good enough to use in that way**. It is not just that the information is noisy. Much more serious is the fact that the scores may be systematically biased *for* some teachers and *against* others… (p. 23)
  - https://www.ets.org/Media/Research/pdf/PICANG14.pdf
- Rothstein on Gates MET:
  - Hence, while the report's conclusion that teachers who perform well on one measure "tend to" do well on the other is technically correct, **the tendency is shockingly weak**. As discussed below (and in contrast to many media summaries of the MET study), this important result casts substantial doubt on the utility of student test score gains as a measure of teacher effectiveness.
  - http://nepc.colorado.edu/files/TTR-MET-Rothstein.pdf

RUTGERS

Common Responses

- Value-added in a given year is still the best predictor of itself a year later
  - That is, VA estimates are more associated with VA estimates the next year than are other metrics (observation protocols, etc.)
  - Circular validity test
- Year-over-career correlations are much higher than year over year
  - That is, if we have multiple years of estimates, they are a more reliable predictor of the next year
  - This, of course, requires waiting multiple years to be able to make any prediction, hence use to inform decisions. This is not a useful argument!
- VAM is as good a year-over-year correlation as baseball batting average
  - But baseball stats geeks agree that's specifically why batting average isn't a useful/reliable indicator of hitting skill[1]
  - And only the higher reported correlations (not all those brushed under the rug) are that high
- It's in its early stages and will get better…
  - Most problems with method result from basic structural issues of American education system, coupled with limitations regarding interpretation of underlying tests/measurement and scaling.
  - That's not likely to change.

[1] http://www.beyondtheboxscore.com/2011/9/1/2393318/what-hitting-metrics-are-consistent-year-to-year

RUTGERS

"Stretching" Variance

- We start w/student testing data, where differences in higher versus lower gain students may result from only a handful of questions (on a 50 item test) answered correctly.
- These 50 item tests are often converted to scale scores, say, with a mean of 200, and 3 or 4 question difference resulting in 10 to 20 point variations.
- We then use gains on these scale scores in our models (oft rescaled again for example, spread to percentiles, etc.)
- We then get variance across teachers on the order of +/- .3 standard deviations for most teachers (because we force that outcome).
- We then spread this variance into percentile ranks of teachers
- But all of that variance, from the "best" to "worst" teachers might be stretched from a few questions difference on a 50 point test!
  - And we wonder why it's so noisy!?

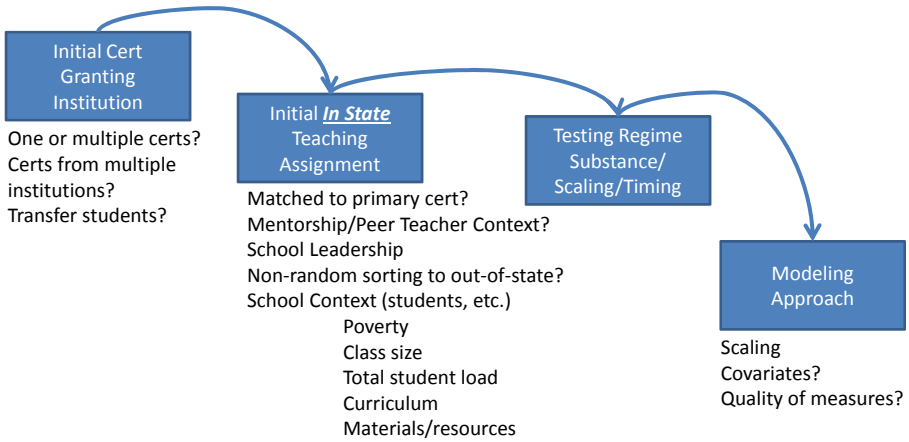
RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Attribution

System Structural & Data Linking Issues

RUTGERS

Attribution



RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Research on Isolating Program Effects

In Systems where Randomization is Entirely Infeasible

RUTGERS

Parsing Variation within/Between Programs

- Koedel, C., Parsons, E., Podgursky, M., & Ehle, M. (2012). *Teacher Preparation Programs and Teacher Quality: Are There Real Differences Across Programs?* (No. 1204). http://econ.missouri.edu/working-papers/2012/WP1204_koedel_et_al.pdf
  - We compare teacher preparation programs in Missouri based on the effectiveness of their graduates in the classroom. The differences in effectiveness between teachers from different preparation programs are very small. In fact**, virtually all of the variation in teacher effectiveness comes from within-program differences between teachers**. Prior research has overstated differences in teacher performance across preparation programs for several reasons, most notably because some sampling variability in the data has been incorrectly attributed to the preparation programs.

RUTGERS

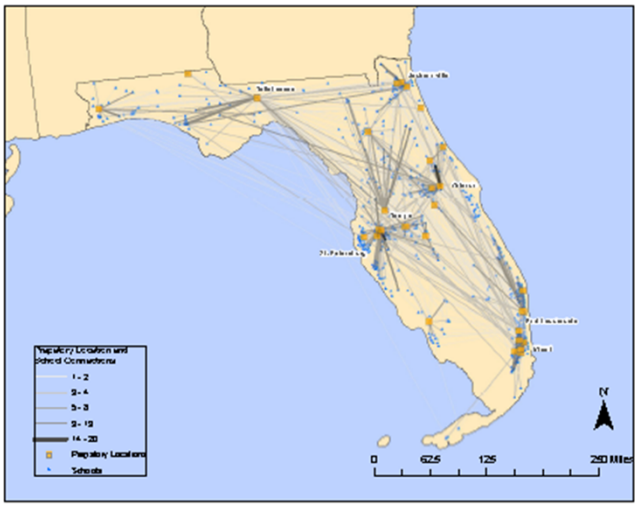
Sensitivity to Alternative Specifications

- Mihaly, K., McCaffrey, D. F., Sass, T. R., & Lockwood, J. R. (2012). Where You Come From or Where You Go?
  - In this paper we consider the challenges and implications of controlling for school contextual bias when modeling teacher preparation program effects.
  - Because teachers from any one preparation program are hired in more than one school and teachers are not randomly distributed across schools, failing to account for contextual factors in achievement models could bias preparation program estimates.
  - We find that some preparation program rankings are significantly affected by the model specification.
    - No Covariates
    - Covariates
    - School Fixed Effect

RUTGERS

Figure 1: Preparation Program and School Connections

Mihaly, K., McCaffrey, D. F., Sass, T. R., & Lockwood, J. R. (2012). Where You Come From or Where You Go?



RUTGERS

Sensitivity to Alternative Specifications

- Applying alternative specifications
  - The authors found that the less good alternatives were, to no surprise, less good- ***potentially biased***.
  - The assumption being that the school fixed effect models are most correct
    - which doesn't, however, guarantee that they are right!
  - the constraint imposed to achieve the "best case" model in this study is a constraint that is unlikely to ever be met for more than a handful of large teacher prep institutions concentrated in a single metropolitan area (or very large state like Florida).

RUTGERS

Substantial resorting of ranks with changes to model specification

Likely a function of stretching small underlying variance into ranks to begin with

Mihaly, K., McCaffrey, D. F., Sass, T. R., & Lockwood, J. R. (2012). Where You Come From or Where You Go?

Table 7: Preparation Program Rankings and Ranking Quartiles - Inexperienced Teachers

	No Schl FE		With Schl FE	
Program ID	*Rank*	*Rank Quartile*	*Rank*	*Rank Quartile*
20	1	1	6	1
32	2	1	32	4
17	3	1	3	1
4	4	1	9	2
7	5	1	13	2
28	6	1	2	1
13	7	1	7	1
12	8	1	14	2
2	9	2	17	2
19	10	2	4	1
16	11	2	22	3
10	12	2	12	2
5	13	2	23	3
14	14	2	29	4
6	15	2	25	3
18	16	2	11	2
8	17	2	15	2
31	18	3	1	1
1	19	3	20	3
24	20	3	5	1
3	21	3	19	3
25	22	3	21	3
29	23	3	18	3
30	24	3	16	2
9	25	3	24	3
11	26	4	10	2
26	27	4	8	1
22	28	4	30	4
15	29	4	28	4
23	30	4	26	4
27	31	4	31	4
21	32	4	27	4
33	33	4	33	4

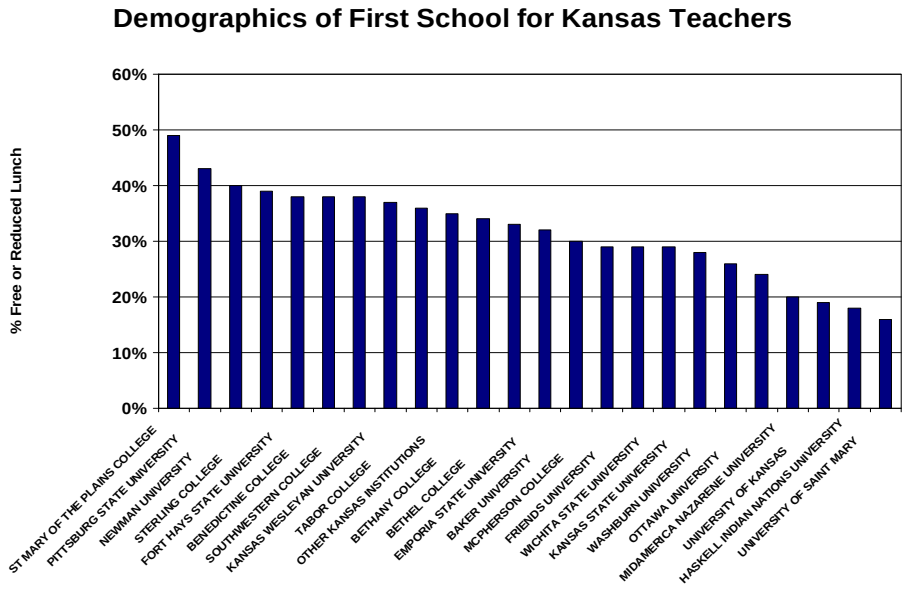
Note: Rankings based on program estimates in Table 6.
Programs ordered by "No Schl FE" rankings

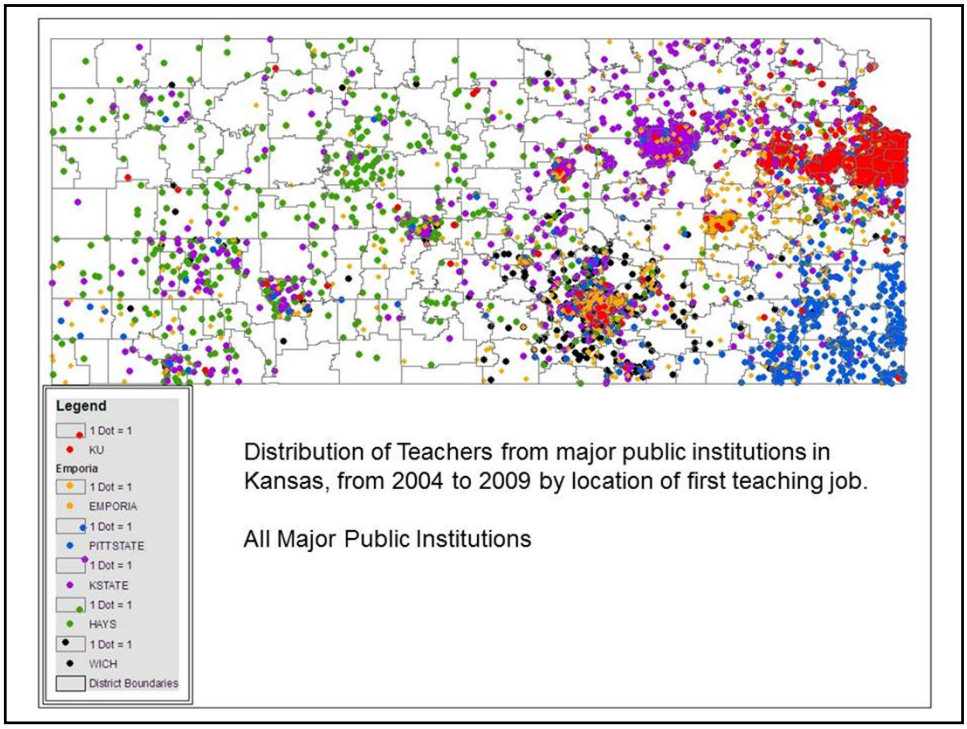
RUTGERS

Needed for Comparing Effects

- Sufficient sample size across preparation programs
  - Of recent graduates within the same school
  - Teaching the same subject
  - At the same grade level
  - Same number of years out of preparation program
  - Serving comparable groups of students
- Better COVARIATES
  - Current batch of dichotomous measures far to crude (insensitive to substantive differences)
- These conditions simply don't exist!

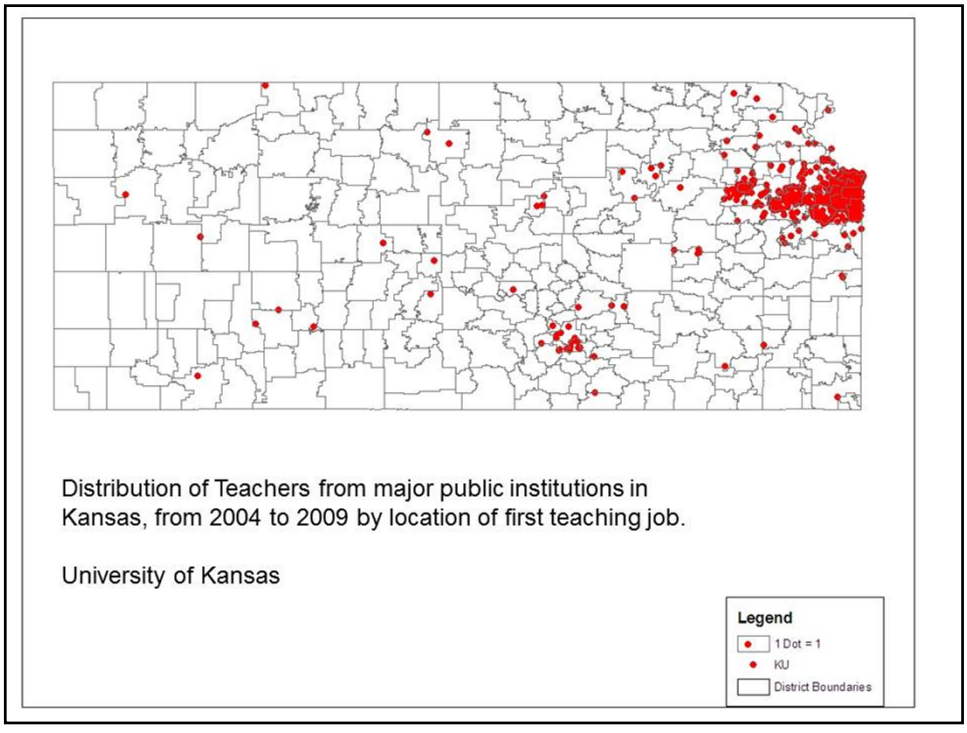
RUTGERS





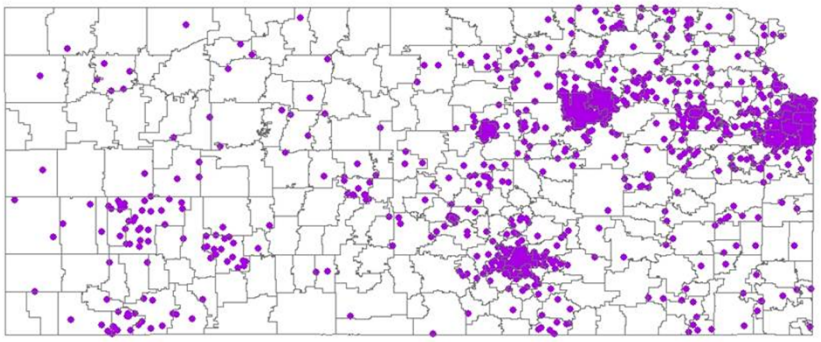
Distribution of Teachers from major public institutions in Kansas, from 2004 to 2009 by location of first teaching job.

All Major Public Institutions



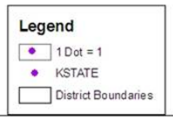
Distribution of Teachers from major public institutions in Kansas, from 2004 to 2009 by location of first teaching job.

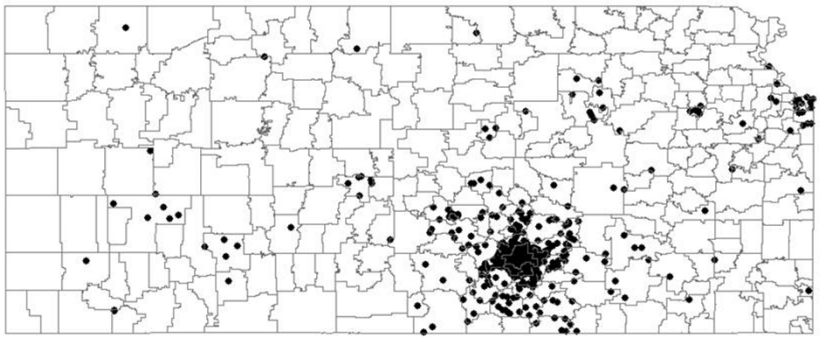
University of Kansas



Distribution of Teachers from major public institutions in Kansas, from 2004 to 2009 by location of first teaching job.

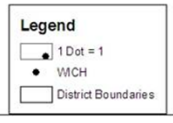
Kansas State University

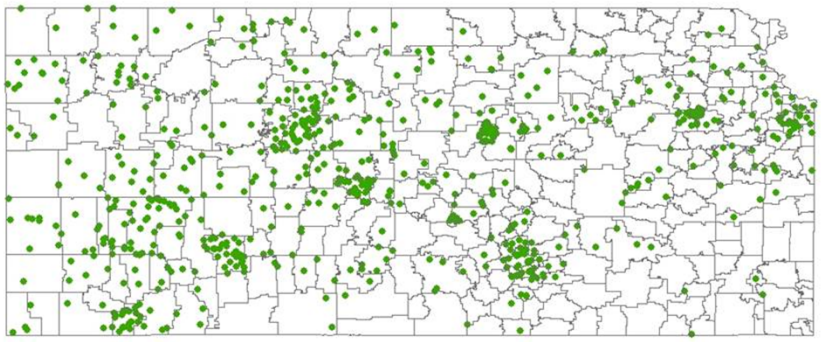




Distribution of Teachers from major public institutions in Kansas, from 2004 to 2009 by location of first teaching job.

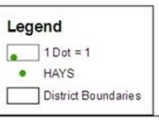
Wichita State University

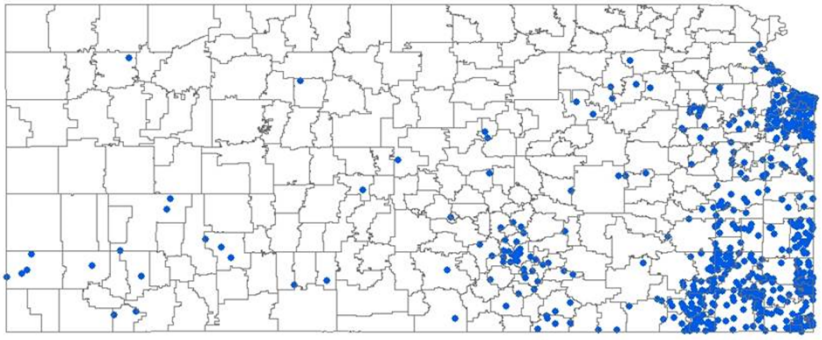




Distribution of Teachers from major public institutions in Kansas, from 2004 to 2009 by location of first teaching job.

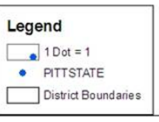
Fort Hays State University

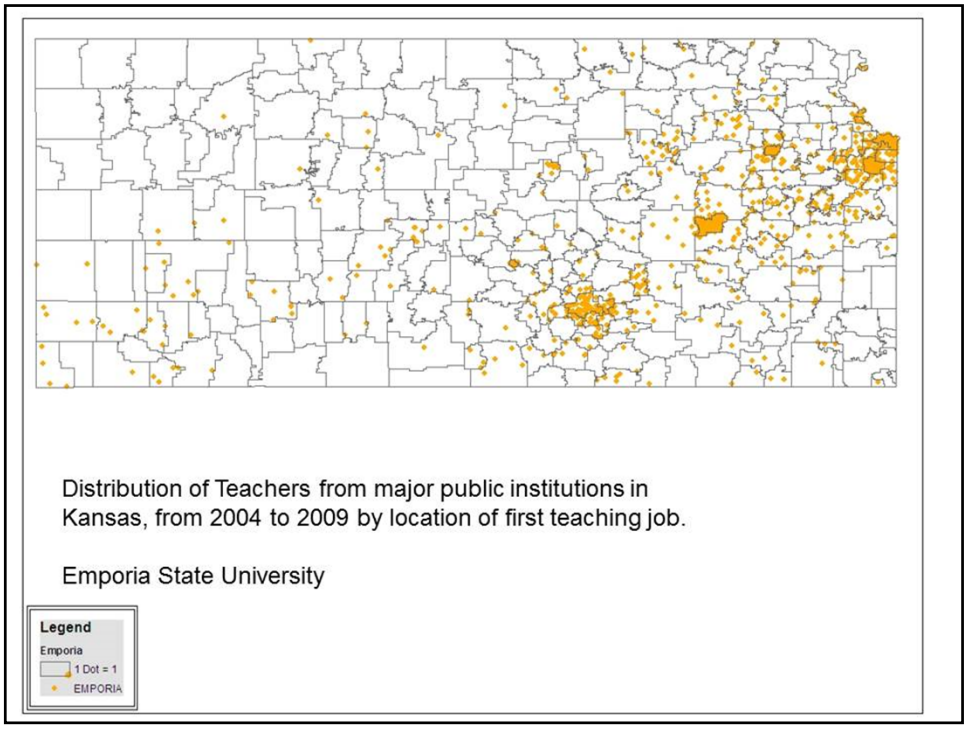




Distribution of Teachers from major public institutions in Kansas, from 2004 to 2009 by location of first teaching job.

Pittsburg State University





RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Inadequate Demographic Covariates

When "Dummy Variables" are just too Dumb

RUTGERS

Relatively thorough model compared to many state models (and all Growth Percentile Models).

Much richer than NY State model.

The Model Mathematically Factors In Measurable Characteristics To Calculate Predicted Gain

Student characteristics	Classroom characteristics	School characteristics
✓Prior year reading	✓Average prior year reading and math	✓Average classroom characteristics
✓Prior year math	✓Percent free or reduced price lunch	✓Average class size
✓Free or reduced price lunch	✓Percent special education status	✓Total tested by grade/subject
✓Special education status	✓Percent English Language Learner status	✓Year starting and ending school
✓English Language Learner status	✓Average number of suspensions and absences (prior)	Teacher Characteristics (used when comparing teachers to peer teachers)
✓Number of suspensions and absences (prior year)	✓Percent of students retained in grade	✓ Years of experience ✓ Years teaching in the same grade and subject
✓Student retained in grade	✓Percent attended summer school	
✓Attended summer school	✓Class size	
✓New to school	✓Percent by race	
✓Race	✓Percent by gender	
✓Gender		
✓Prior year teacher		

NYC Department of Education

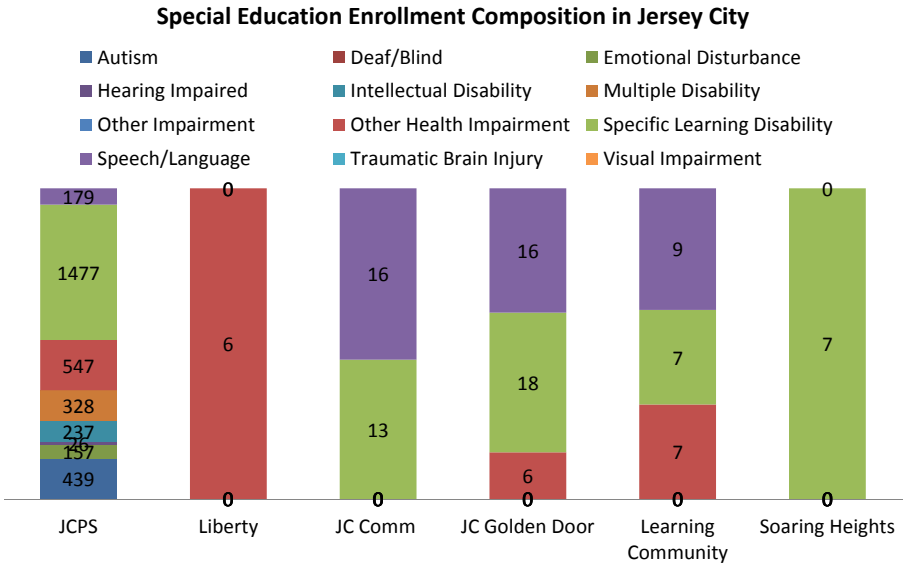
10

RUTGERS

Crude Dummy Variables

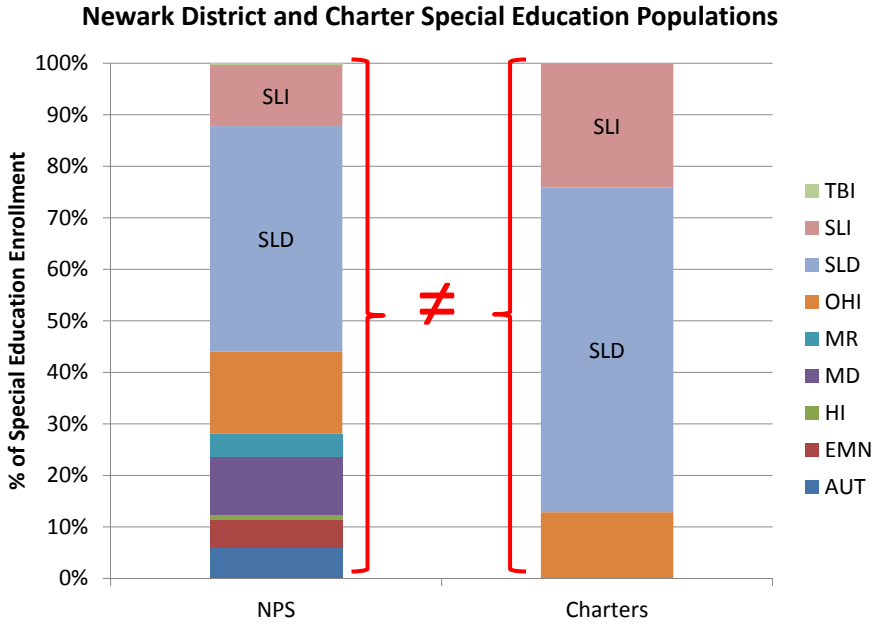
- Qualifies for National School Lunch Program
  - <130% income threshold for poverty
  - <185% income threshold for poverty
  - Includes most kids in some states and nearly all in some districts – thus picks up little no variance in income across settings.
  - Uses same income thresholds across regions, despite different economics of urban/rural areas
  - Problematic when used as individual or school level covariate
- Special Education/Has IEP
  - Composition of special education population may vary widely, especially in labor markets with large charter share.
  - Problematic when used as individual or school level covariate
- ELL status (or not)

RUTGERS



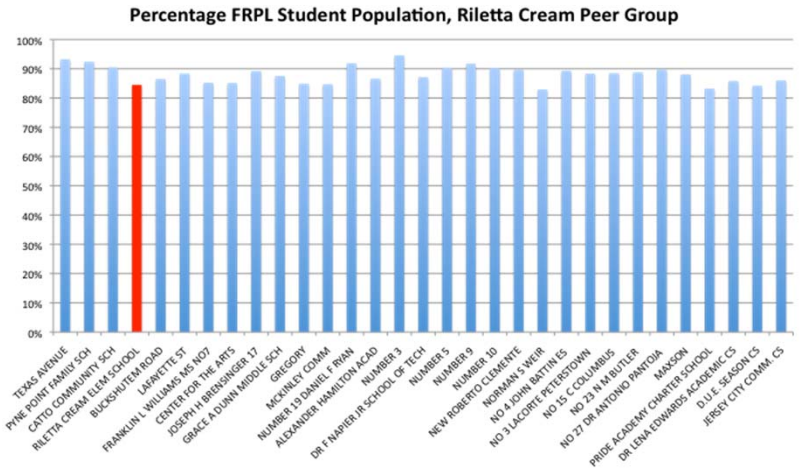
Data from: http://www.nj.gov/education/specialed/data/ADR/2012/EligibilitybyPlacement/PlacementByElig6-21.xls

RUTGERS



RUTGERS

Income/Poverty Variation in High Poverty Setting

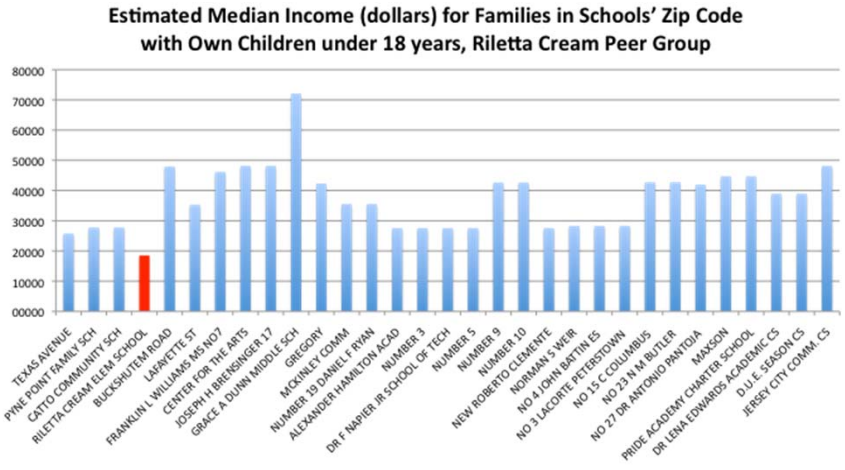


Source: U.S. Census Bureau, American Community Survey,
http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_5YR_S1903&prodType=table

Source: Weber & Lamonde [RU GSE] 2013

RUTGERS

Income/Poverty Variation in High Poverty Setting

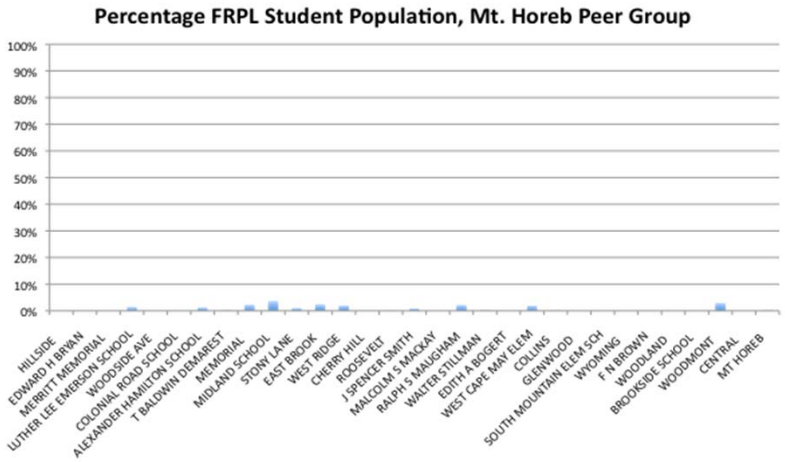


Source: U.S. Census Bureau, American Community Survey,
http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_5YR_S1903&prodType=table

Source: Weber & Lamonde [RU GSE] 2013

RUTGERS

Income/Poverty Variation in Low Poverty Setting

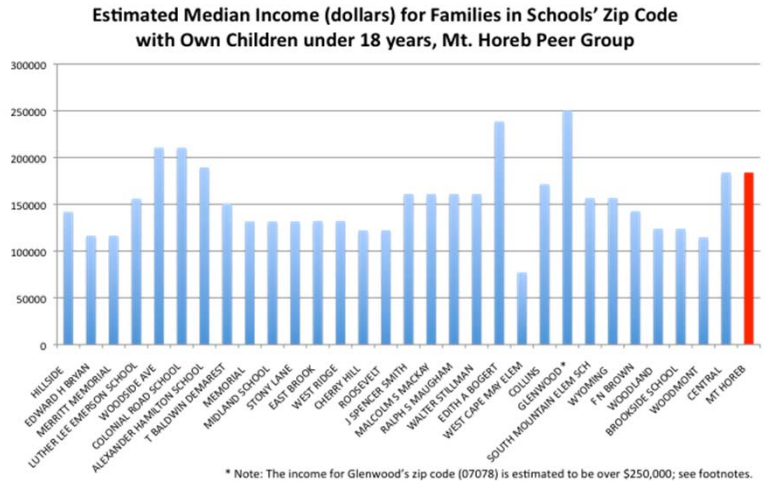


Source: U.S. Census Bureau, American Community Survey,
http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_5YR_S1903&prodType=table

Source: Weber & Lamonde [RU GSE] 2013

RUTGERS

Income/Poverty Variation in Low Poverty Setting

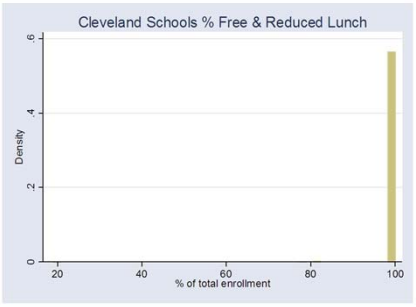


Source: U.S. Census Bureau, American Community Survey,
http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_5YR_S1903&prodType=table

Source: Weber & Lamonde [RU GSE] 2013

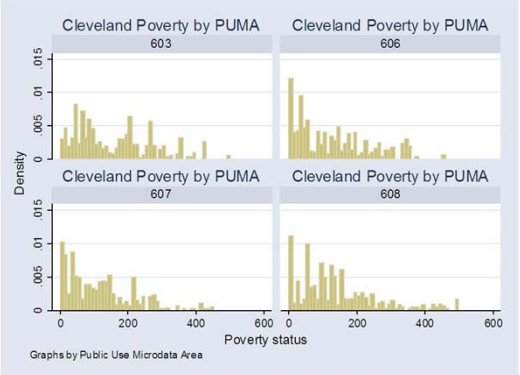
RUTGERS

Extreme Example-Cleveland Ohio



Schools data represented with dummy variable

Areas of city using full income distribution



RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Existing State "Teacher Effect" Measures are Very Low Quality

Growth Percentiles vs. VAMs, Persistent Bias and Noise

RUTGERS

External Review of NY State MGPs

But the study found that New York did not adequately weigh factors like poverty when measuring students' progress.

"We find it more common for teachers of higher-achieving students to be classified as 'Effective' than other teachers," the study said. "Similarly, teachers with a greater number of students in poverty tend to be classified as 'Ineffective' or 'Developing' more frequently than other teachers."

Andrew Rice, a researcher who worked on the study, said New York was dealing with common challenges that arise when trying to measure teacher impact amid political pressures.

"We have seen other states do lower-quality work," he said.

http://www.lohud.com/article/20131015/NEWS/310150042/Study-faults-NY-s-teacher-evaluations

RUTGERS

NY State Technical Report on MGPs

- AIR Technical Report
  - Finding:
    - Despite the model conditioning on prior year test scores, schools and teachers with students who had higher prior year test scores, on average, had higher MGPs. Teachers of classes with higher percentages of economically disadvantaged students had lower MGPs. (p. 1) http://schoolfinance101.files.wordpress.com/2012/11/growth-model-11-12-air-technical-report.pdf
  - Conclusion?
    - The model selected to estimate growth scores for New York State provides a **fair** and **accurate** method for estimating **individual teacher and principal effectiveness** based on specific regulatory requirements for a "growth model" in the 2011-2012 school year. p. 40 http://engageny.org/wp-content/uploads/2012/06/growth-model-11-12-air-technical-report.pdf

RUTGERS

The SGP/MGP Difference (from VAM)

- Two recent working papers compare SGP and VAM estimates for teacher and school evaluation and both raise concerns about the face validity and statistical properties of SGPs.
  - Goldhaber and Walch (2012) conclude: "For the purpose of starting conversations about student achievement, SGPs might be a useful tool, but one might wish to use a different methodology for rewarding teacher performance or making high-stakes teacher selection decisions" (p. 30).[6]
  - Ehlert and colleagues (2012) note: "Although SGPs are currently employed for this purpose by several states, we argue that they (a) cannot be used for causal inference (nor were they designed to be used as such) and (b) are the least successful of the three models [Student Growth Percentiles, One-Step VAM & Two-Step VAM] in leveling the playing field across schools" (p. 23).[7]

[6] Goldhaber, D., & Walch, J. (2012). Does the model matter? Exploring the relationship between different student achievement-based teacher assessments. University of Washington at Bothell, Center for Education Data & Research. CEDR Working Paper 2012-6.

[7] Ehlert, M., Koedel, C., &Parsons, E., & Podgursky, M. (2012). Selecting growth measures for school and teacher evaluations. National Center for Analysis of Longitudinal Data in Education Research (CALDAR). Working Paper #80.

RUTGERS

In the authors words…

Damian Betebenner:

"Unfortunately Professor Baker conflates the data (i.e. the measure) with the use. A primary purpose in the development of the Colorado Growth Model (Student Growth Percentiles/SGPs) ***was to distinguish the measure from the use: To separate the description of student progress (the SGP) from the attribution of responsibility for that progress."***

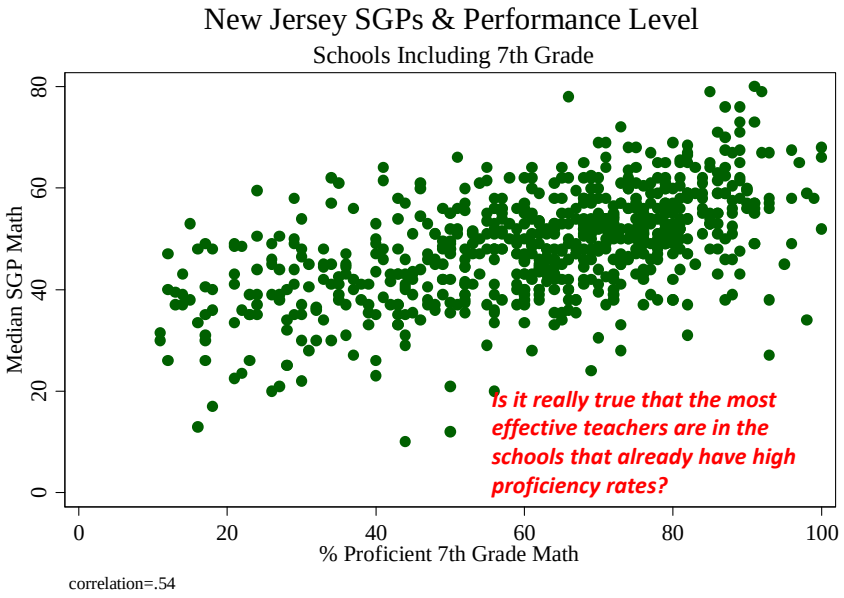
http://www.ednewscolorado.org/voices/student-growth-percentiles-and-shoe-leather

RUTGERS

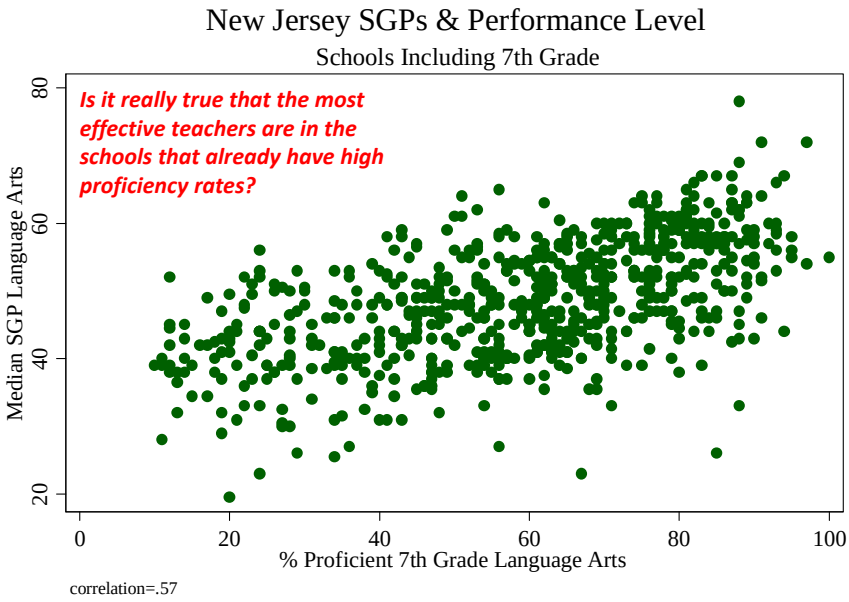
Parsing Words…

- When pressed on the point that GPs are not designed for attributing student gains to their teachers, those defending their use in teacher evaluation will often say…
  - "SGPs are a good measure of student growth, and shouldn't teachers be **accountable** for student growth?"
- Let's be clear here, one cannot be **accountable** for something that is not rightly **attributable** to them!
- If these measures aren't attributable to the teachers, then they certainly aren't attributable to the institutions that prepared them!

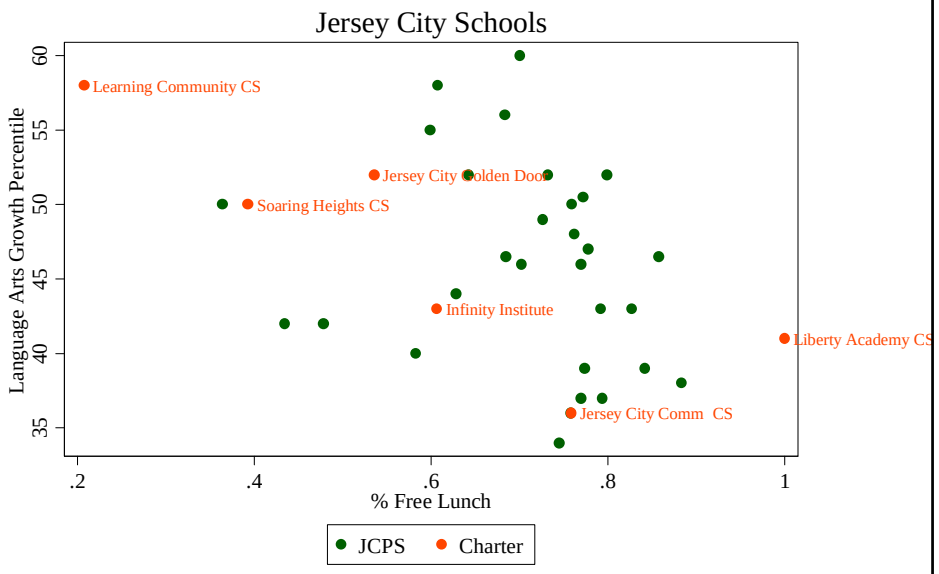
RUTGERS



RUTGERS



RUTGERS



RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

The Path Forward

Thoughtful Consideration of Data as one Tool within Evaluation Systems

RUTGERS

Difficult to Untangle

- Selection Effect
- Placement Effect
- Program Effect

RUTGERS

General Conclusion

- What might be useful for exploratory analysis, including raising questions about selection effects versus placement effects versus program effects, may not be useful for accountability purposes (rating and ranking).
- Some existing measures and state data systems inadequate for either.

RUTGERS

Accountability & Incentives

- Should we produce teachers for certain fields?
- Should we encourage teachers to work in certain schools?
  - Certain grades?
  - With certain populations?
- How can we increase correlation between our pre-service program quality measures and state accountability measures? (should we?)

RUTGERS

Reasonable Directions

- We must always have multiple model specifications and sensitivity analyses
  - Using alternate testing data (multiple different tests of same subjects)
  - Using alternative model variables/structures
  - We must use these sensitivity analyses to paint a richer picture than can be presented with any one model.
- It would be foolish to aggregate or average results across models as it would be to force a specific model!
- It would be absolutely wrong to base any high stakes policy determinations (accreditation, financial aid access) on the results of a single model or single aggregation across models.
- We must acknowledge and build into our evaluation process the fallibility of the outcome measures.
  - That is, we must **not** set up our evaluation systems to presume that twice removed measured student achievement gains are the ultimate validity check.
  - Any evaluation system integrating these model results MUST include the option to ignore/override them entirely.