

Modeling | Measuring Growth

Arie van der Ploeg

Principal Researcher, American Institutes for Research
avanderploeg@air.org

School and District Accountability Design Team
October 17, 2011

Modeling | Measuring Growth

You can't always get what you want

Arie van der Ploeg

Principal Researcher, American Institutes for Research
avanderploeg@air.org

School and District Accountability Design Team
October 17, 2011

Modeling | Measuring Growth

*You can't always get what you want
Sometimes you get what you need*

Arie van der Ploeg

Principal Researcher, American Institutes for Research
avanderploeg@air.org

School and District Accountability Design Team

October 17, 2011

What do we want for students?

- Academic skills and knowledge
- Critical thinking ability
- Readiness for college and career
- Appreciation of arts and literature
- Social skills and work ethic
- Good citizenship and community responsibility
- Good physical and emotional health
- Etc.

Rothstein, R. (2008). *Grading Education: Getting Accountability Right*. Washington: Economic Policy Institute.

What do we want from the system?

- Improved teaching and learning
- Accurate accountability determinations
- Allocate support and professional development to staff appropriately
- Identify most/least effective educators
- Informative, actionable data about schools, programs, staff, and students
- Broad understanding of school productivity

What do we speak about when we speak 'growth'?

| | Status | Change |
|---------------|---|--|
| Achievement | What percent of students meet standards? How high do students score on state tests? | Are percentages rising or falling with time? Are scores rising? |
| Effectiveness | Is each student learning satisfactorily? Are individual students making sufficient progress from grade to grade? | Is a school becoming more effective? Are successive cohorts accelerating in learning? |

Carlson, D. (2002). *Focusing State Educational Accountability Systems: Four Methods of Judging School Quality and Progress*. Dover, NH: Center for Assessment.

Some basic 'growth' calculations

| Grade | 2009 | 2010 | 2011 |
|-------|------|------|------|
| 4 | 24 | 30 | 29 |
| 5 | 28 | 32 | 26 |

| | Status | Change |
|---------------|--|---|
| Achievement | <p>What percent of students meet standards? How high do students score on state tests?</p> <p>24</p> | <p>Are percentages rising or falling with time? Are scores rising?</p> <p>$30 - 24 = 6$</p> |
| Effectiveness | <p>Is each student learning? Are individual students making desired progress from grade to grade?</p> <p>$32 - 24 = 8$</p> | <p>Is a school becoming more effective? Are successive cohorts accelerating in learning?</p> <p>$(26 - 30) - (32 - 24) = -12$</p> |

Seems easy enough. Why isn't it?

A The question a parent will ask:

Which school is best for my child?

B The question an educational leader will ask:

Which school makes more learning?

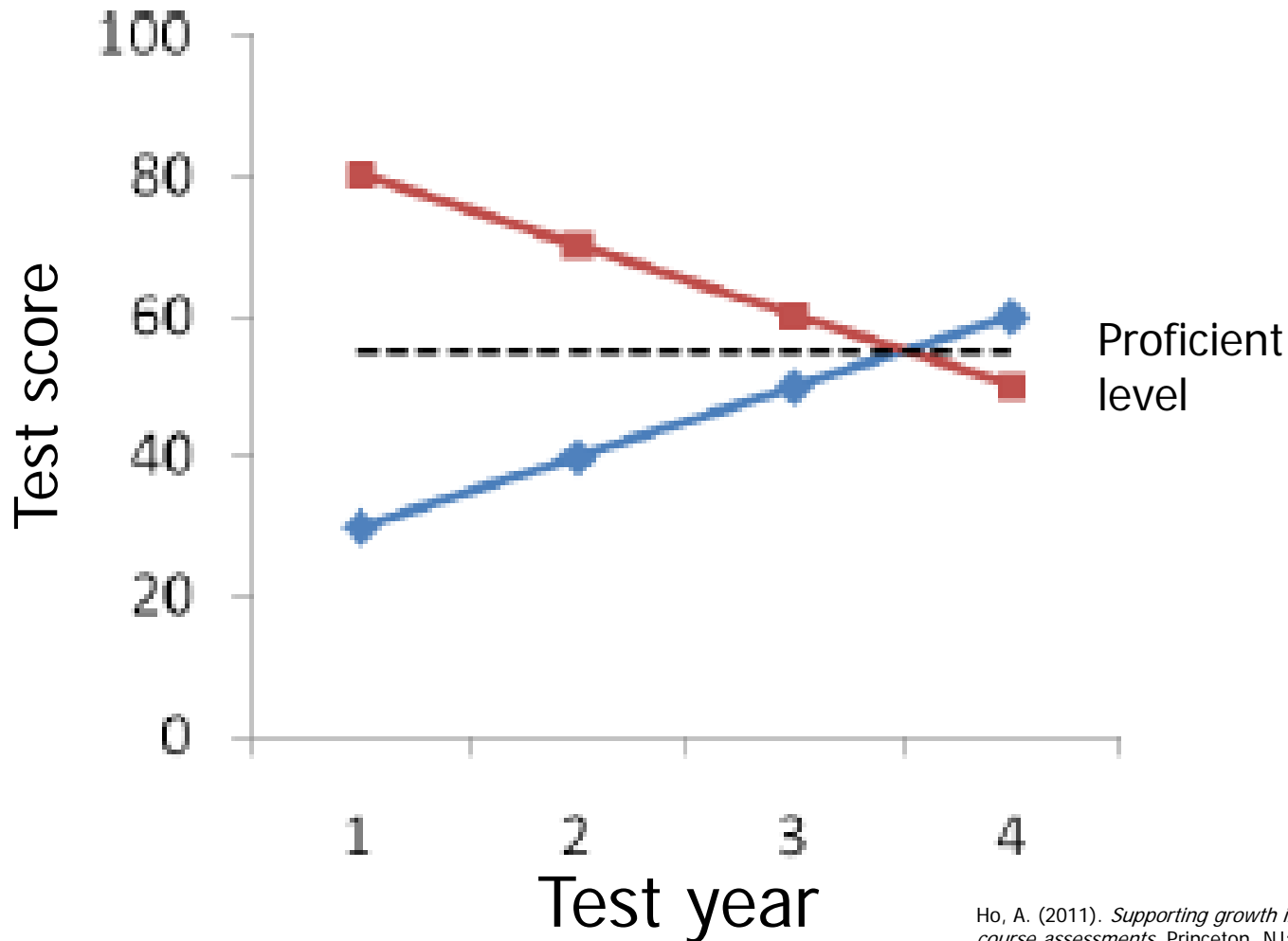
Causal answer to

A requires randomization of children to schools.

B requires randomization of schools to regimes of educational practice.

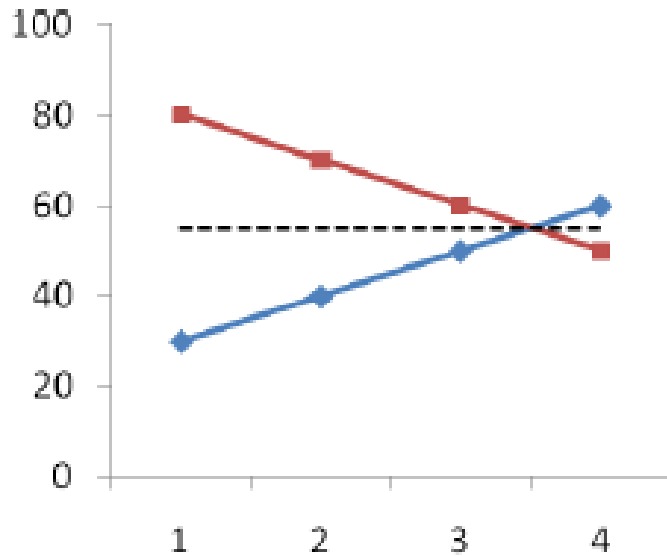
Raudenbush, S.. (2004). *Schooling, statistics, and poverty: Can we measure school improvement?*. Princeton, NJ: Educational Testing Service.

Growth to Standard



Ho, A. (2011). *Supporting growth interpretations using through-course assessments*. Princeton, NJ: Educational Testing Service

Growth to Standard: Contradictory results



Trajectory model regressions calculate best fit line for each student. Hence, blue student will be proficient at time 5; red student not.

Prediction model estimates a best general regression equation, using five data points from all students of a prior cohort. Red student's prediction for time 5 likely will be above the proficiency threshold; blue student's below it.

Ho, A. (2011). *Supporting growth interpretations using through-course assessments*. Princeton, NJ: Educational Testing Service

Models, purposes, incentives

- Growth-to-standard models are not “growth models in the traditional statistical or psychometric sense. [They are] concerned with improving student and school outcomes, not with understanding the functional form of growth trajectories or [with ascribing] growth to various value-added causal effects” (Ho, 2011).
- Statistical models are seductive. But, what matters is how results are used: “All models are wrong but some are useful” (Box, 1979).
- The issue becomes one of understanding the variations in utility across persons, roles, and purposes.
- In other words, get incentives right: Evaluate “growth models not only by the inferences they support but also by the incentives they create” (Ho, 2011).

VAMs and covariates

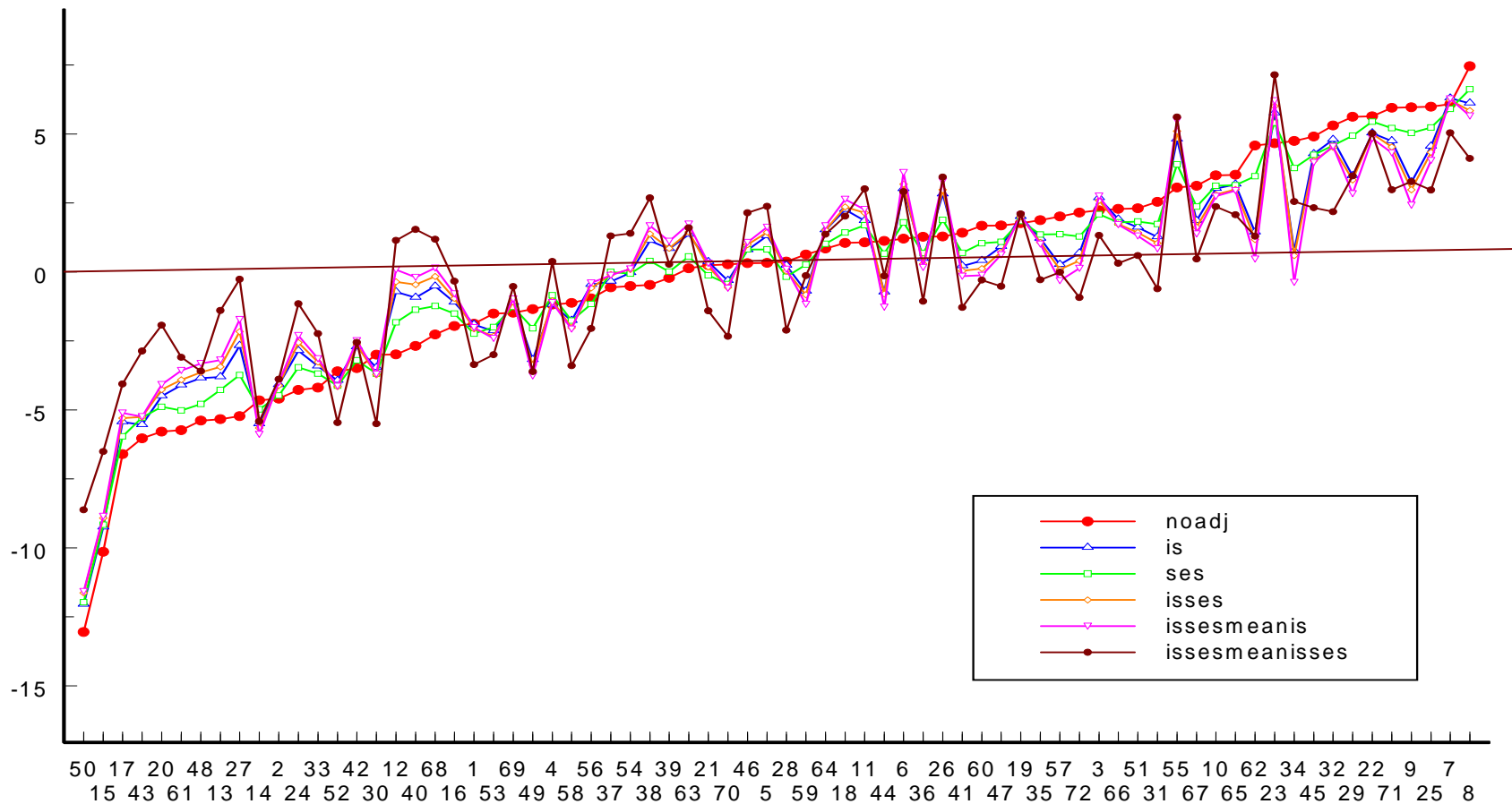
Rank order correlations among several models

| | No adjustment | Student initial status | Student SES | Student initial status & SES | Student initial status, SES, & school_mean_initial_status |
|--------------------------|---------------|------------------------|-------------|------------------------------|---|
| Stud_IS | 0.93 | 1 | | | |
| Stud_SES | 0.98 | 0.97 | 1 | | |
| Stud_IS_SES | 0.91 | 1.00 | 0.96 | 1 | |
| Stud_IS_SES & MeanIS | 0.86 | 0.98 | 0.92 | 0.99 | 1 |
| Stud_IS_SES & MeanIS-SES | 0.75 | 0.88 | 0.85 | 0.90 | 0.90 |

Choi K., Yamashiro, K., Seltzer, M., & Herman, J. (2004). *Comparing like with like: The role of student and school characteristics in value-added models*. Los Angeles, CA: CRESST.

VAMs and covariates

Choi K., Yamashiro, K., Seltzer, M., & Herman, J. (2004). *Comparing like with like: The role of student and school characteristics in value-added models*. Los Angeles, CA: CRESST.



Test choice and quality matter

- Variation in effects due to math measures was large compared to choice of model; variation within teachers across measures was larger than variation across teachers.
- Three reading tests did “not rank individual teacher consistently.” Rank order correlations ranged from 0.15 to 0.58 across multiple models. Test timing and measurement error played critical roles.

Lockwood, J., McCaffrey, D., Hamilton, L., Stecher, B., Le, V., & Martinez, J. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement* 44(1, Spring), pp. 47-67.

Papay, J. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal* 48(1, February), pp. 163-193.

Risk of indicator corruption

“The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.

Achievement tests may well be valuable indicators of general school achievement **under conditions of normal teaching aimed at general competence**. But when test scores become the goal of the teaching process, they both lose their value as indicators of educational status and distort the educational process in undesirable ways.”

Campbell, D. (1976). *Assessing the impact of planned social change*. Hanover, NH: Public Affairs Center, Dartmouth College.

One researcher's perspective

- ❑ VAM is better than no VAM; GPM is better than no GPM.
- ❑ Using both can confuse—unless audiences and purposes are well defined and distinct.
- ❑ Substantive and technical quality of assessments matters more than implementation of VAM or GPM.
- ❑ Reporting and distribution of results matters much.
- ❑ Reporting should support inquiry, from multiple perspectives.
- ❑ Preparation for and continuous delivery of training for interpretation of and inquiry into results is critical.
- ❑ No accountability system is complete without a system of supports for the schools, teachers, and students targeted.

A focus on Wisconsin

- ❑ Wisconsin has set three goals: Literacy, Gaps, Transitions
- ❑ Accountability is about evaluating progress toward goals
 - ❑ What progress requires must be understood, and support provided
 - ❑ Longitudinal measurement is necessary
 - ❑ Indicators must serve to align incentives across multiple actors
 - ❑ The “right” indicator may be a myth, since goals may change and measurement technology will
- ❑ Primacy of local actors and agencies
- ❑ Accept that indicators are imprecise and contain error
- ❑ Collapsing distributions into categories throws information away
- ❑ Reliance only on a few numerical indicators is probably foolish

Building accountability for Wisconsin

- ❑ **Goals** define targets for action
- ❑ **Theory of action** suggests how goals will be achieved
- ❑ **Supports** describe the amount and nature of assistance, rewards, or sanctions for progress
- ❑ **Indicators** are the evidence base for accountability decisions
- ❑ **Business rules** specify measurement procedures
- ❑ **Decision rules** define criteria for determinations
- ❑ **Intended consequences** are what will happen as a consequence of determinations
- ❑ **Unintended consequences** are outcomes of accountability that are not expected



Arie van der Ploeg

312-238-2312

avanderploeg@air.org

**REL Midwest at American Institutes for
Research**

20 N. Wacker Dr., Suite 1231

Chicago, IL 60606