

# Improving school performance measures to raise the bar: A framework for diagnostic use of data

ED Evidence Summit, November 2023: Evidence in Action

Brian Gill  
Director  
REL Mid-Atlantic

# How can we make data *useful* to educators and policymakers?

**Data should be used as a flashlight, not a hammer.**

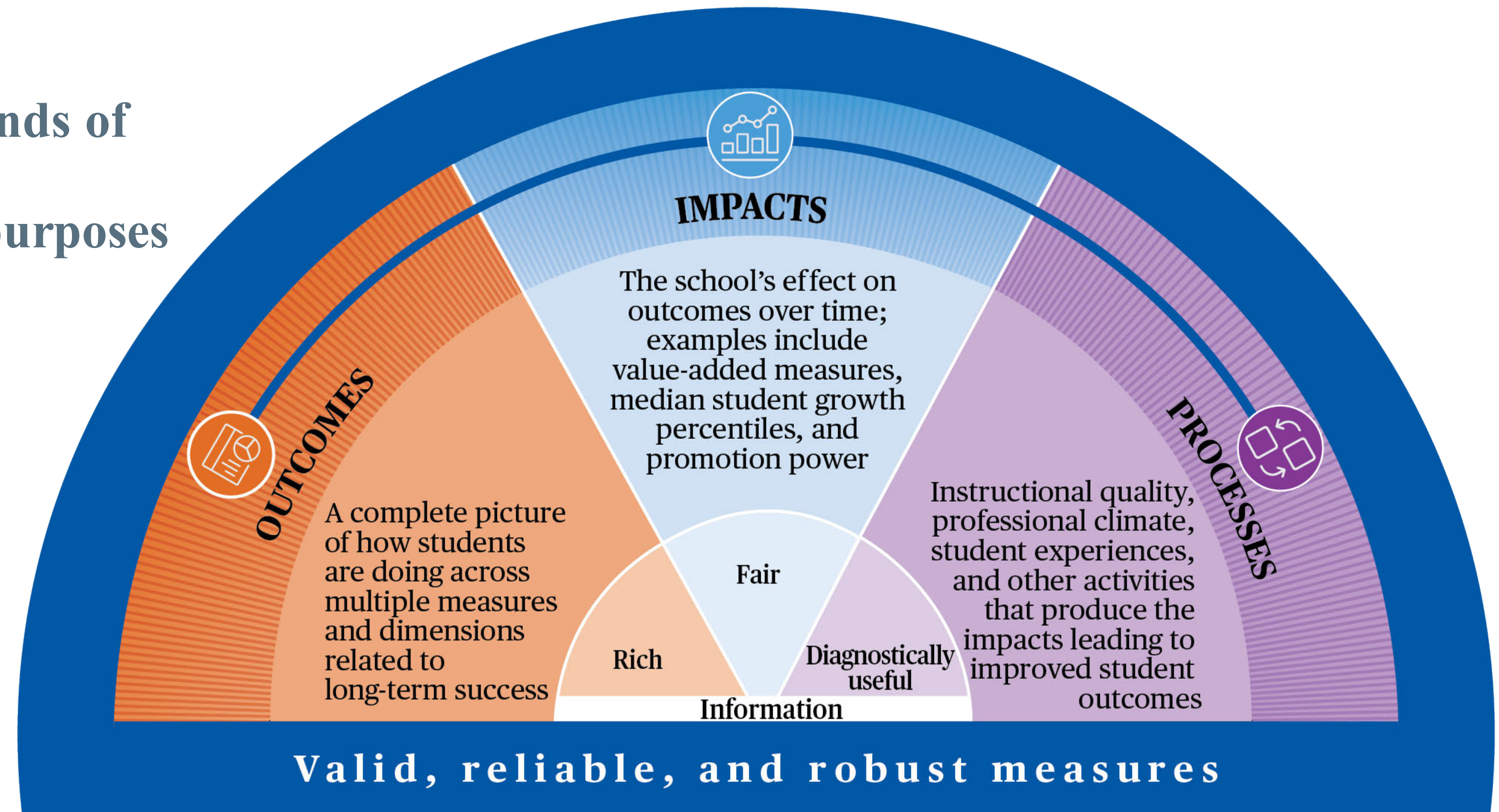
Miguel Cardona,  
U.S. Secretary of Education

Policymakers and educators are implored to be “data-driven” and “evidence-based”—but data often aren’t actionable

- Helping educators and policymakers make sense of data and use it for improvement is a major focus of the Regional Educational Laboratories (RELs)
- REL Mid-Atlantic has developed a framework of school performance measures that collectively create a powerful flashlight, for diagnostic and accountability purposes—promoting accuracy and equity

## Raising the bar requires distinct kinds of data—on *outcomes*, *impacts*, and *processes*—for distinct diagnostic purposes

- Achieving academic excellence and creating pathways for college and career success require rich information on *outcomes* and fair information on school *impacts*
- Boldly improving learning conditions requires useful information on *processes* in schools



### Measuring performance can drive improvement only if the measures themselves are good.

<b>Valid</b>	They must measure what decisionmakers perceive them as measuring, without substantial systematic bias that would lead to mistaken inferences.
<b>Reliable</b>	They must not be susceptible to a large amount of random variation, leading to instability, misdiagnosis, and (over time) a lack of credibility.
<b>Robust</b>	They must be resistant to unintended consequences, including corruption of the measure and neglect of important unmeasured outcomes.

Source: Gill (2022, 2023)

# 1: Student **outcomes**

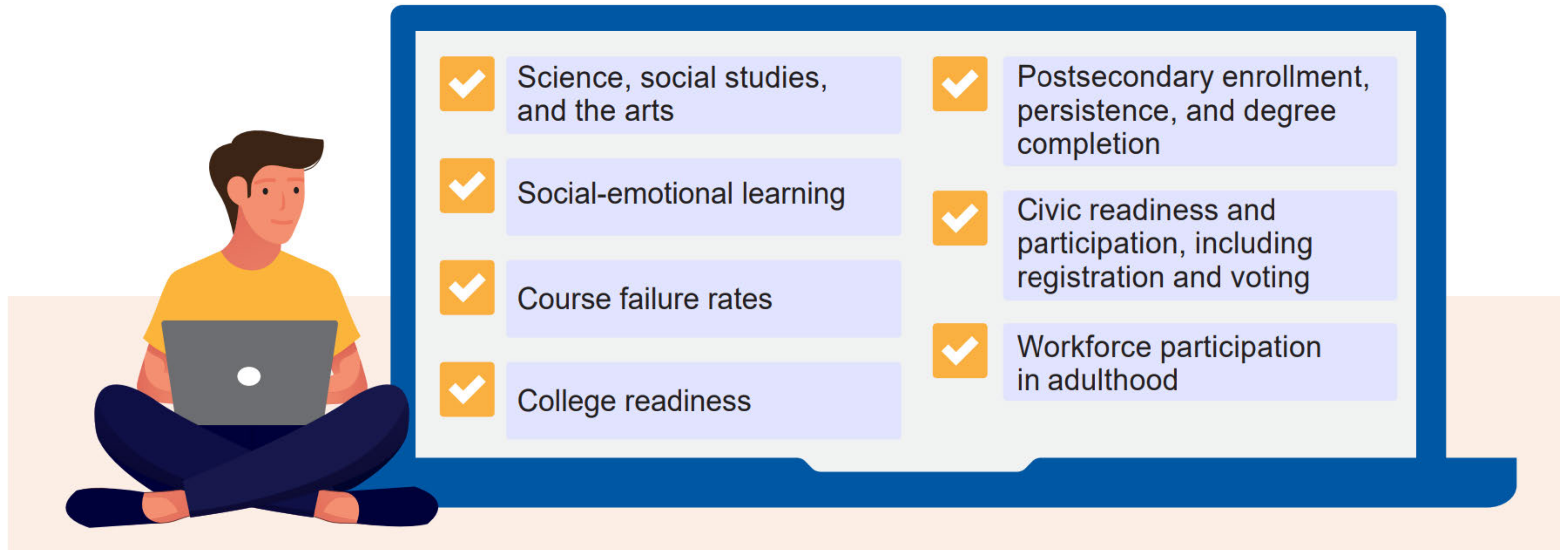
*How are all the kids doing, across different demographic groups?*

*Are they achieving academic excellence, across a comprehensive range of subject matter?*

*How is their mental health and well-being?*

# Measures of student outcomes should be broader and richer

Math and reading tests don't measure everything schools want kids to learn

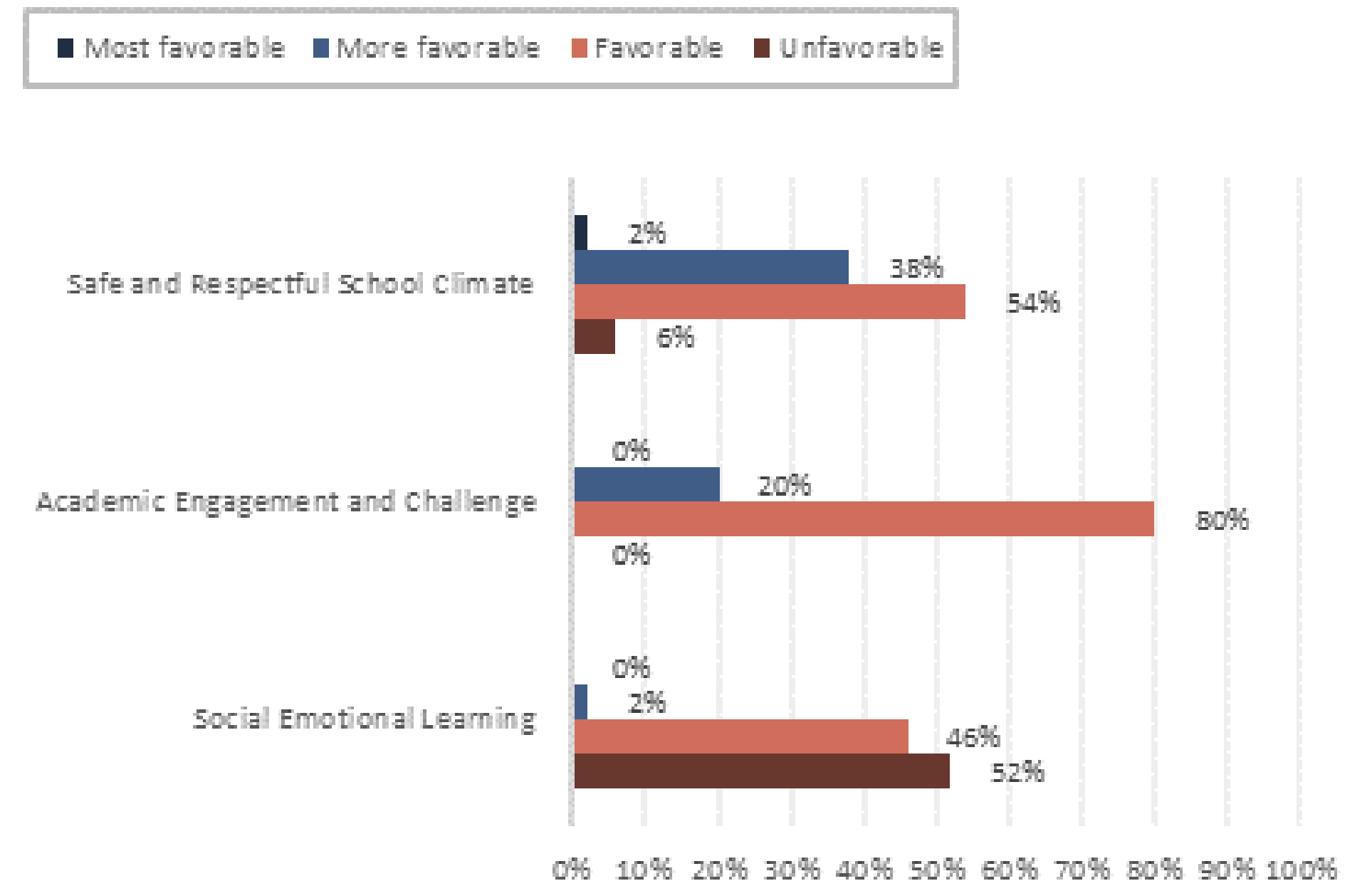
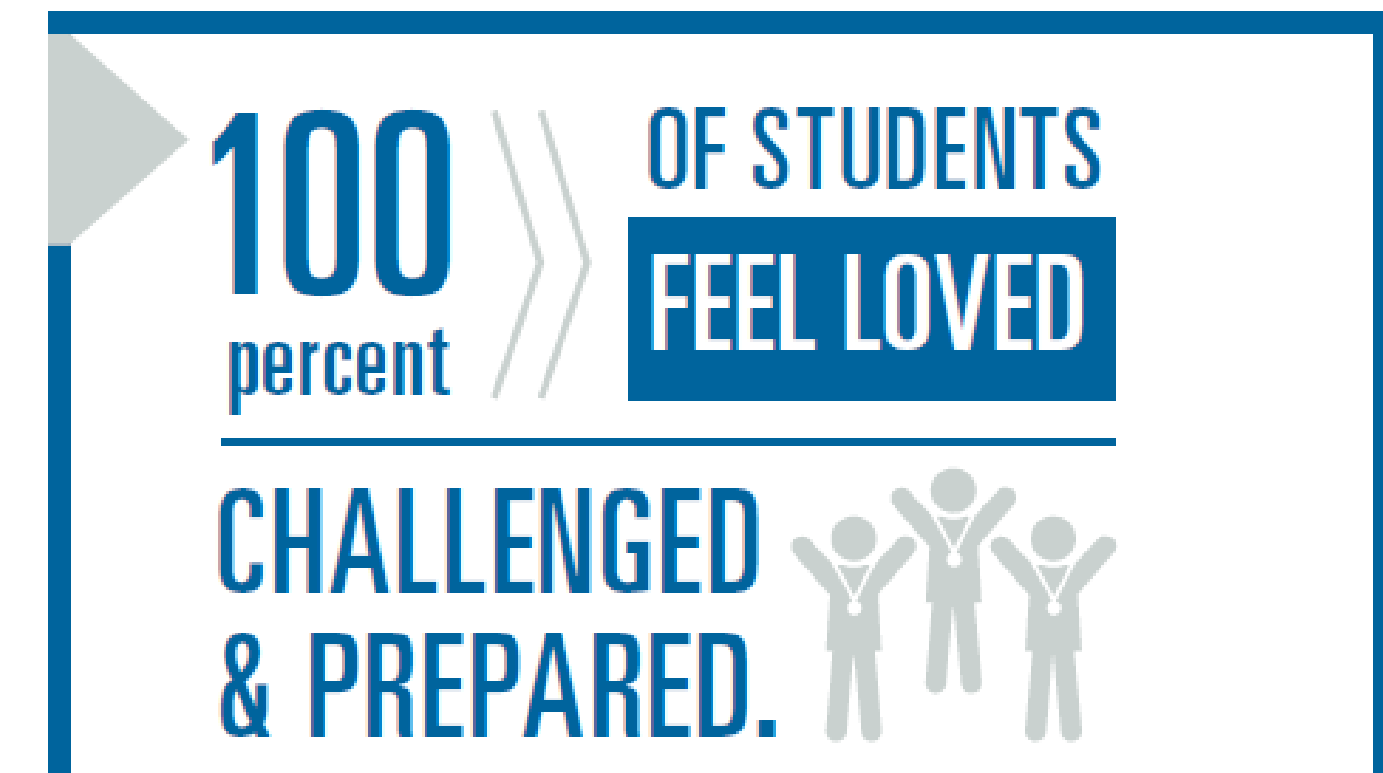


The illustration shows a person with dark hair, wearing a yellow t-shirt and dark blue pants, sitting cross-legged on a light orange surface. They are holding a silver laptop. To their right is a large, blue-bordered rectangular box containing a list of eight student outcome measures, each preceded by a white checkmark inside an orange square.

- ✓ Science, social studies, and the arts
- ✓ Social-emotional learning
- ✓ Course failure rates
- ✓ College readiness
- ✓ Postsecondary enrollment, persistence, and degree completion
- ✓ Civic readiness and participation, including registration and voting
- ✓ Workforce participation in adulthood

# REL Mid-Atlantic assessed social-emotional learning in PA and DC

- DC Public Schools: Analyzed student survey data to create index of whether students feel “loved, challenged, and prepared” (Kautz et al., 2021)
  - Not used for formal accountability, but districtwide measure is publicly reported
- Pennsylvania: Analyzed student and staff survey data to create school climate index (Amos & Xue, 2021)
  - Not used for formal accountability: Schools opt into survey participation



# Surveys can produce useful SEL information at the school level

- It is possible to get good response rates from students (and staff)
- Schoolwide SEL measures show good psychometric characteristics
- SEL varies systematically at the school level
  - There is more variation within schools than between schools, but between-school variation is meaningful

# Civics: Preparation for citizenship is foundation of public education

- Horace Mann, 1855: A well-functioning democracy requires an educated citizenry
  - “Education must be universal . . . The qualification of voters is as important as the qualification of governors, and even comes first, in the natural order . . . The theory of our government is . . . that every person... shall become fit to be a voter. Education must bring the practice as nearly as possible to the theory.”
- NAEP results in civics and history show lots of room for improvement
- Civic purpose of education remains relevant today
  - Raising the bar in civics is urgent



# To promote equity, **subgroup** measures need to be reliable

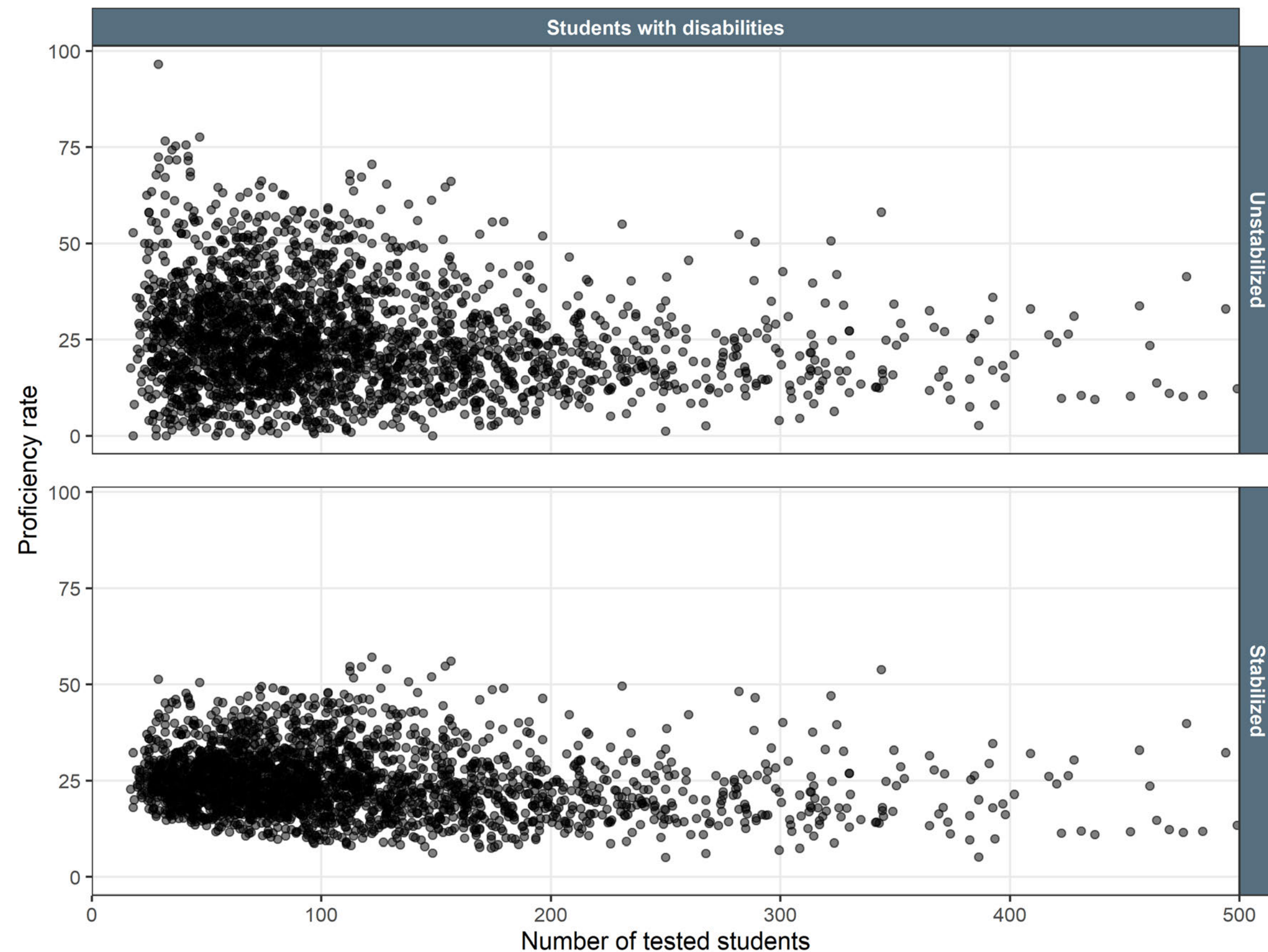
- ESSA seeks to promote equity by requiring reports on student subgroups in schools
- But subgroups are often small—and measures for small subgroups are often unreliable: they have substantial measurement error
- States set “minimum  $n$ -sizes” to mitigate the reliability problem: 10, 20, or 30 students
- Setting a minimum  $n$ -size involves a tradeoff between accuracy and equity:
  - Large  $n$ -size improves reliability but makes small subgroups invisible
  - Small  $n$ -size counts more students in subgroups at the cost of reliability

*States can solve this problem—simultaneously promoting accuracy and equity—through Bayesian stabilization*

# Bayesian stabilization improves accuracy by learning from patterns in the data

- Learning about a school's performance in one year from its historical trend
- Learning about one school from other schools across the state
- Stabilization goes by many names:
  - (Bayesian) hierarchical modeling
  - (Bayesian) shrinkage
  - (Bayesian) random effects modeling
  - Reliability adjustment
- REL Mid-Atlantic study demonstrates improvement in accuracy using subgroup data from Pennsylvania

# Stabilization dramatically improves reliability for small subgroups (in PA)



**Unstabilized rates** showed a funnel pattern, with more variation for smaller subgroups

This pattern likely reflects **measurement error**.

**Stabilized rates** reduced random variation for small subgroups

Increased consistency in variability across subgroup size suggests that **stabilization improves statistical reliability**.

Stabilized results for 10-19 students show variability comparable to 100+ students

*With stabilization, states can reduce n-sizes, improving accuracy AND equity*

## 2: **Impacts** on student outcomes

*How much does the school contribute to how kids are doing?*

*Is it accelerating learning for every student?*

*Is it ensuring that every student has a pathway to college and career?*

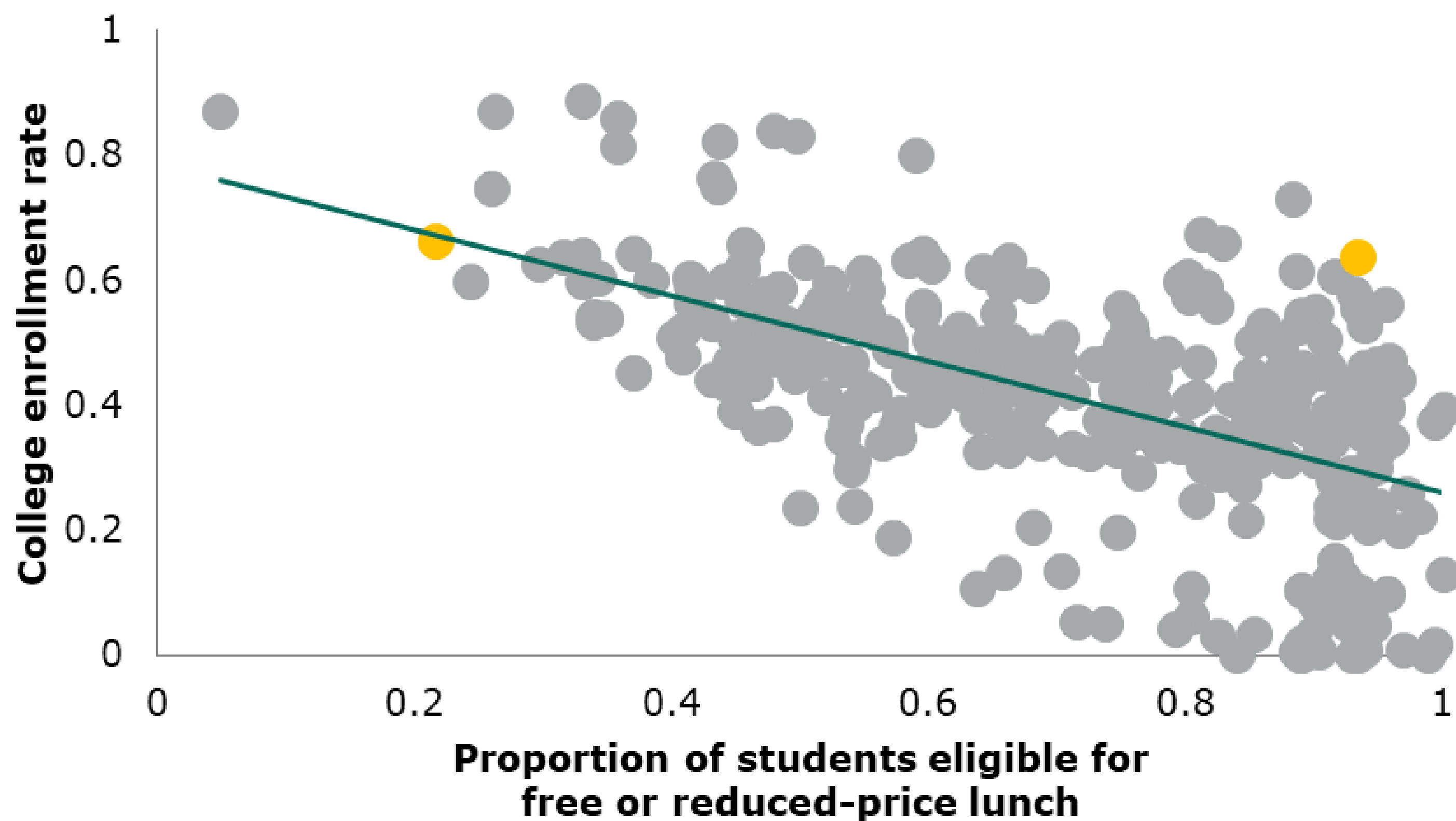
# Impacts must be distinguished from raw outcomes

- NCLB critics recognized raw proficiency results conflated school performance/student characteristics
  - Critiques prompted development of value-added and student growth measures
- Expanding outcomes included in accountability systems risks re-creating NCLB's flaw in failing to distinguish school's contribution
  - Why not apply statistical techniques to identify school impacts on SEL, graduation, college enrollment?
- Accountability arguments about relative weight of outcomes vs impacts (status vs growth) miss the point: Outcomes and impacts are diagnostic for different purposes
  - Low-status/high-growth school could end up with same overall rating as high-status/low-growth school, but they need very different interventions

# Promotion power measures separate schools' contributions from other factors for non-test outcomes

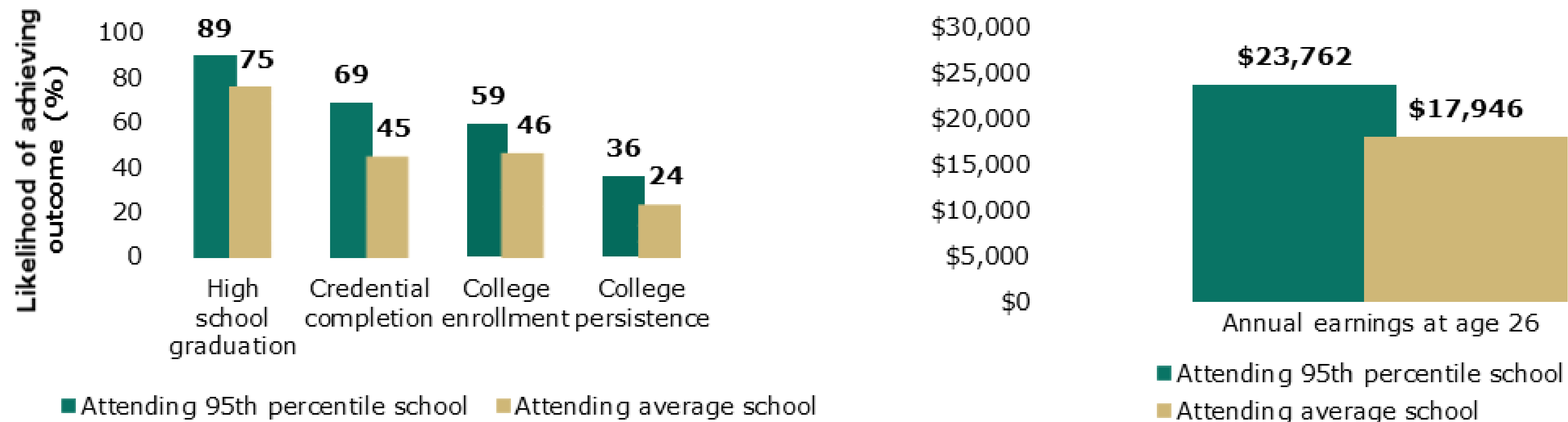
- Use statistical methods similar to those in value-added/growth models (Deutsch et al., 2020; Gross et al., 2021)
- Account for differences in advantages and disadvantages of students served
  - Poverty, prior achievement, IEP and ELL status, anything relevant and measured prior to high-school entry
- Aims to be fair to schools and provide better diagnostic info to districts and states
- Explored in partnerships with state education agencies in Louisiana and DC
- Relatedly, Mathematica is measuring colleges' impacts on earnings (in work with Postsecondary Commission)

High schools can have the same student outcomes with very different promotion power; raising the bar requires distinguishing them



Source: Deutsch et al. (2020)

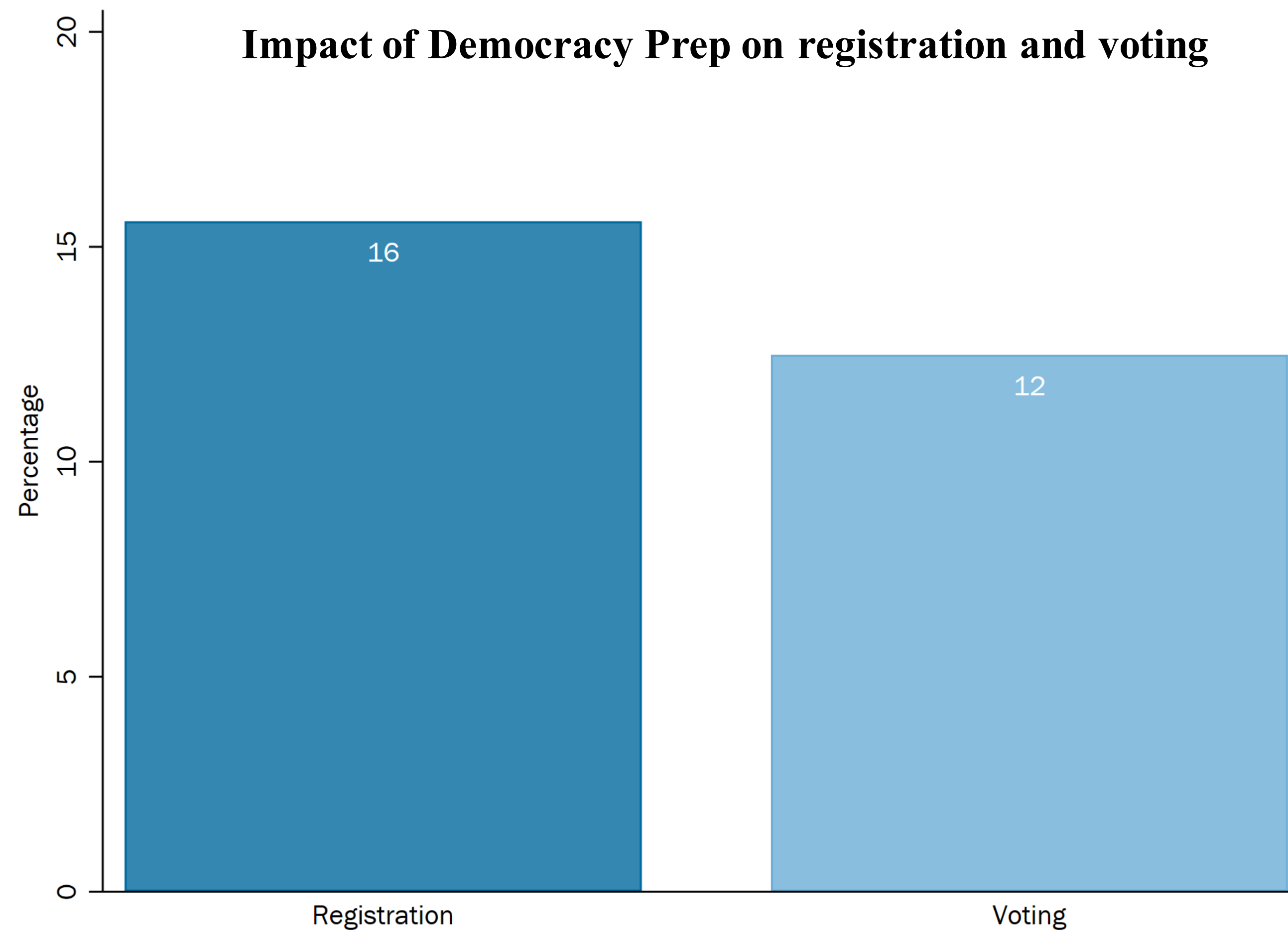
# Attending a high school with high promotion power can substantially improve a student's long-term success



Source: Deutsch et al. (2020)



# Schools can have large impacts on registration and voting



Source: Gill et al. (2020)

### 3: Educational **processes**

*What is happening in the school?*

*Is it boldly improving learning conditions?*

*Is it delivering a comprehensive and rigorous education for every student?*

# Process measures can help identify areas for possible intervention

- Even if an impact measure provides a valid and reliable measure of performance, it is a black box: doesn't tell us how or why
- Process measures and mechanisms might include indicators of students' opportunity to learn
  - Observations of instructional practice
  - Climate surveys
  - Student participation indicators from learning management system
  - School inspections
  - Exclusionary discipline
  - Class size, teacher qualifications and experience
  - Quality of curriculum and materials
  - Diversity, equity, inclusion measures

# Process measures might compensate for weaknesses of outcome and impact measures

- Some impacts may be impossible to measure
  - We can produce acceptably valid and reliable estimates of the value-added of schools, but...
  - Nobody knows how to measure the impact of an individual principal on student outcomes (Chiang, McCullough, Lipscomb, Gill, 2016)
- Outcome/impact measures, even if valid and reliable, are thin/incomplete
  - But we can recognize good schools and good teaching from observation
  - Evidence indicates that even students can recognize good teaching (Raudenbush & Jean, 2015; Chaplin et al., 2014)

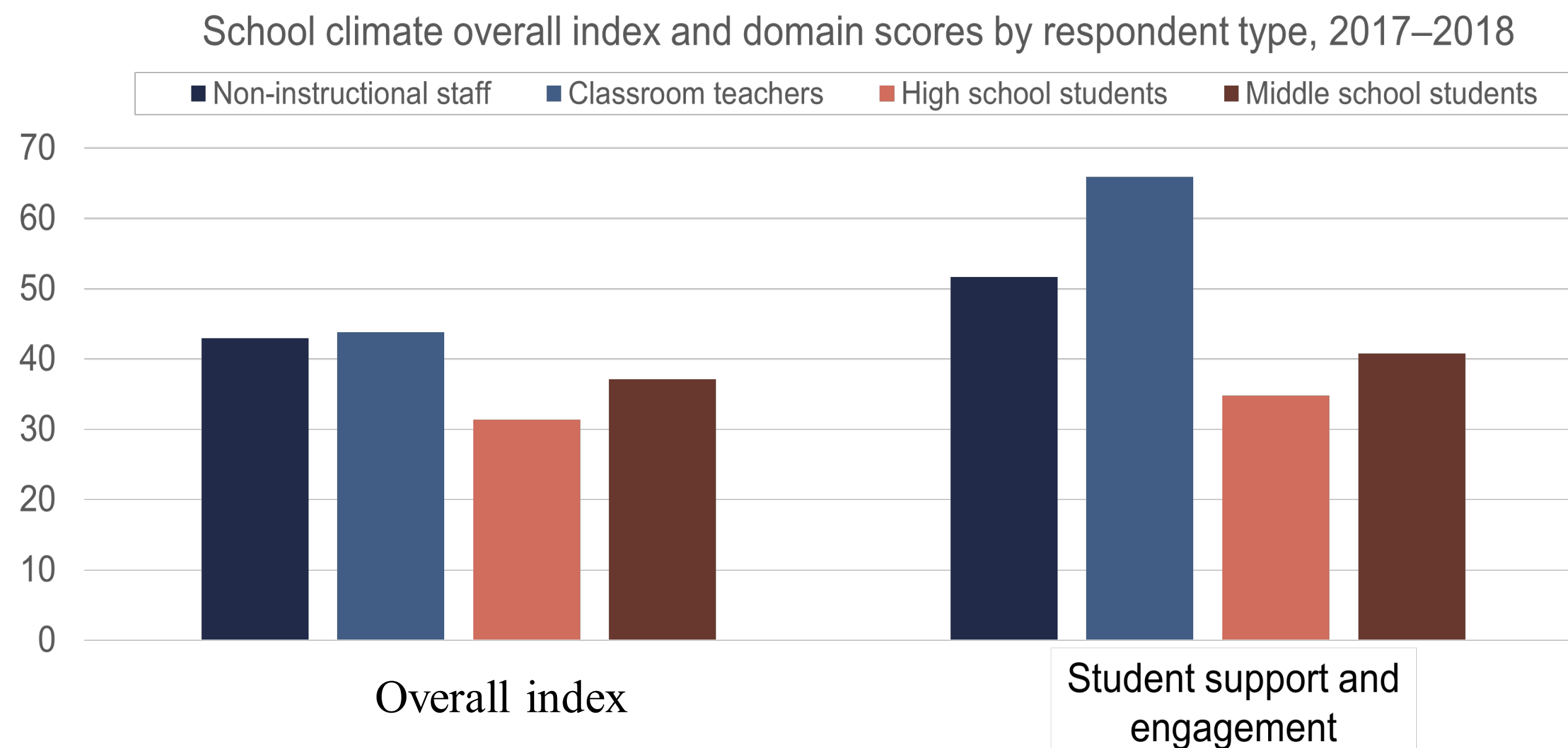
# School climate changes substantially with a new principal



Source: Kozakowski, Gill, & Shiferaw, 2021

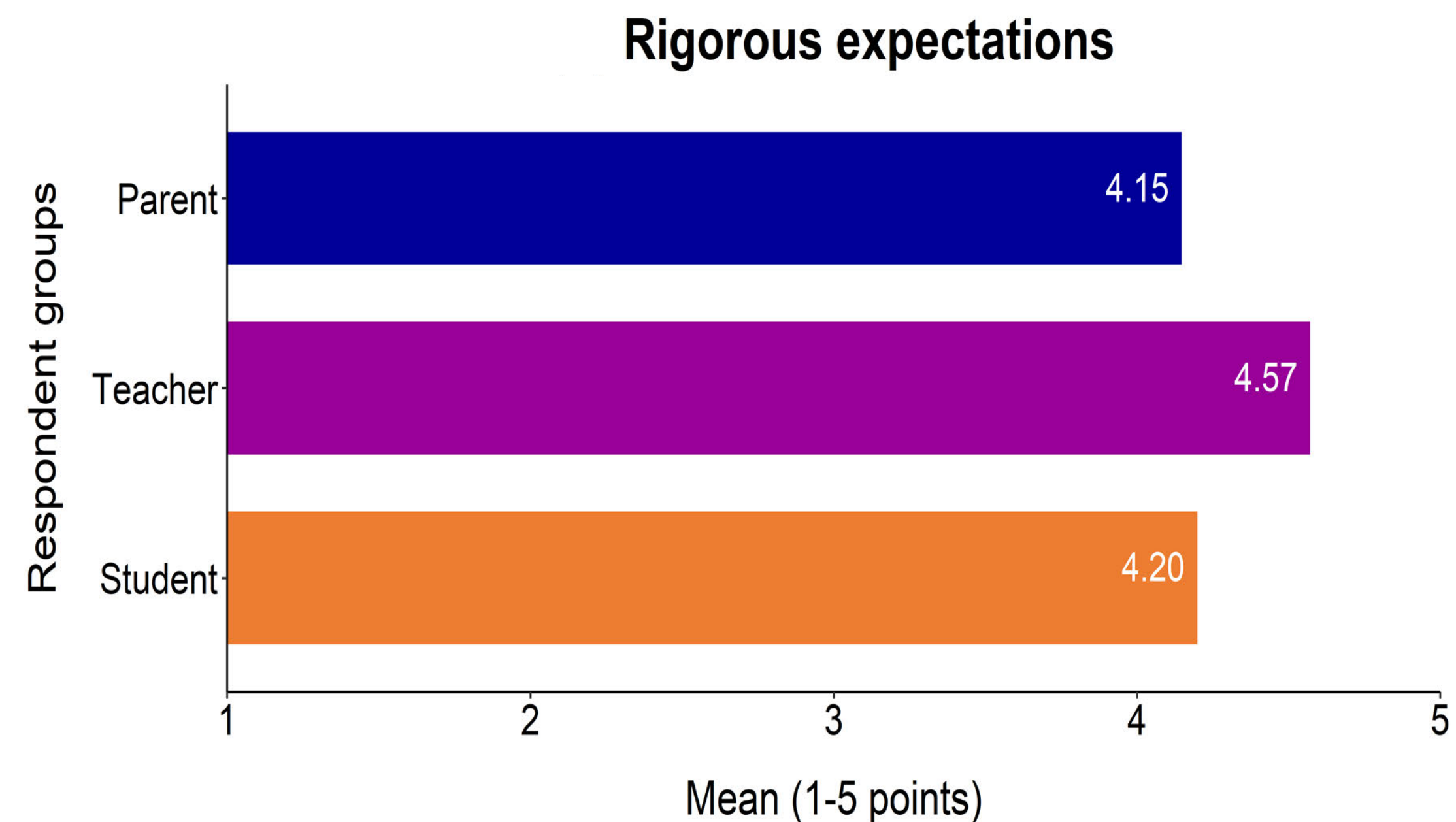
# School climate: Student and staff ratings are correlated but also differ systematically

## Pennsylvania



Source: Amos & Xue (2021)

## DCPS



Source: Kautz et al. (2021)

# Implications for policy and practice

# 1: Broaden discussion of performance beyond student outcomes and beyond formal consequences

- Outcomes are not the only measures that matter: examine impacts and processes too—but keep them separate
- Beware the descent into Taylorism
  - Distorting effects of high stakes (Campbell's Law); some measures are more corruptible than others
  - Don't be fooled by randomness of unreliable measures
- Accountability can be created without high stakes (Gill et al., 2016)
  - Information alone can motivate
  - Lower stakes may lower the temperature
  - Transparency can create accountability



## 2: Recognize the limits of data and measurement

- Good measures can identify:
  - Student and school needs (how are the students doing?)
  - Underperforming schools (what is the school contributing?)
  - Broken processes (what is happening in the school?)
- But measurement alone can't tell an educator or administrator what to do
- Actions to improve must be informed by expert knowledge from the field

# Disclaimer

This work was funded by the U.S. Department of Education's Institute of Education Sciences (IES) under contract 91990022C0012, with REL Mid-Atlantic, administered by Mathematica. The content of the presentation does not necessarily reflect the views or policies of IES or the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. government

# Citations

- Amos, L., & Xue, Y. (2021, June). *Development of Pennsylvania Department of Education school climate index summary*. Regional Educational Laboratory Mid-Atlantic, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). *Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh Public Schools*. Regional Educational Laboratory Mid Atlantic, Institute of Education Sciences, U.S. Department of Education.
- Chiang, H., McCullough, M., Lipscomb, S., & Gill, B. (2016). *Can student test scores provide useful measures of principals' performance?* (NCEE Report No. 2016-002). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.
- Deutsch, J., Johnson, M., & Gill, B. (2020). *The promotion power impacts of Louisiana high schools*. Mathematica.
- Gill, B. (2022). What should the future of educational accountability look like? *Journal of Policy Analysis and Management*, vol. 41(4), 1232–1239.
- Gill, B. (2023). Outcomes, impacts, and processes: How a three-part measurement strategy can improve schools' performance. *School Administrator*, January 2023, 16–20.
- Gill, B., Lerner, J. S., & Meosky, P. (2016). Re-imagining accountability in K-12 education: A behavioral science perspective. *Behavioral Science & Policy*, 2(1), 57–70.
- Gill, B., Ruble Whitesell, E., Corcoran, S. P., Tilley, C., Finucane, M., & Potamites, L. (2020). Can charter schools boost civic participation? The impact of Democracy Prep public schools on voting behavior. *American Political Science Review*, 114(4), 1386–1392. doi:10.1017/S000305542000057X.
- Gross, M., Shiferaw, M., Deutsch, J., & Gill, B. (2021). *Using promotion power for college- and career-ready SAT scores, graduation, and college enrollment to identify the effectiveness of public high schools in the District of Columbia* (REL 2021-098). Regional Educational Laboratory Mid-Atlantic, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Kautz, T., Feeney, K., Chiang, H., Lauffer, S., Bartlett, M., & Tilley, C. (2021). Using a survey of social and emotional learning and school climate to inform decisionmaking (REL 2021-114). Regional Educational Laboratory Mid-Atlantic, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Kozakowski, W., Gill, B., & Shiferaw, M. (2021). *Exploring the potential role of staff surveys in school leader evaluation* (REL 2021-117). Regional Educational Laboratory Mid-Atlantic, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Raudenbush, S., & Jean, M. (2015). To what extent do student perceptions of classroom quality predict teacher value added? In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 170–202). Jossey-Bass.