Using Bayesian stabilization to improve reliability in performance measures in education

October 23, 2024

Lena Rosendahl
Senior Data Scientist
Mathematica

Jennifer Starling
Statistician
Mathematica

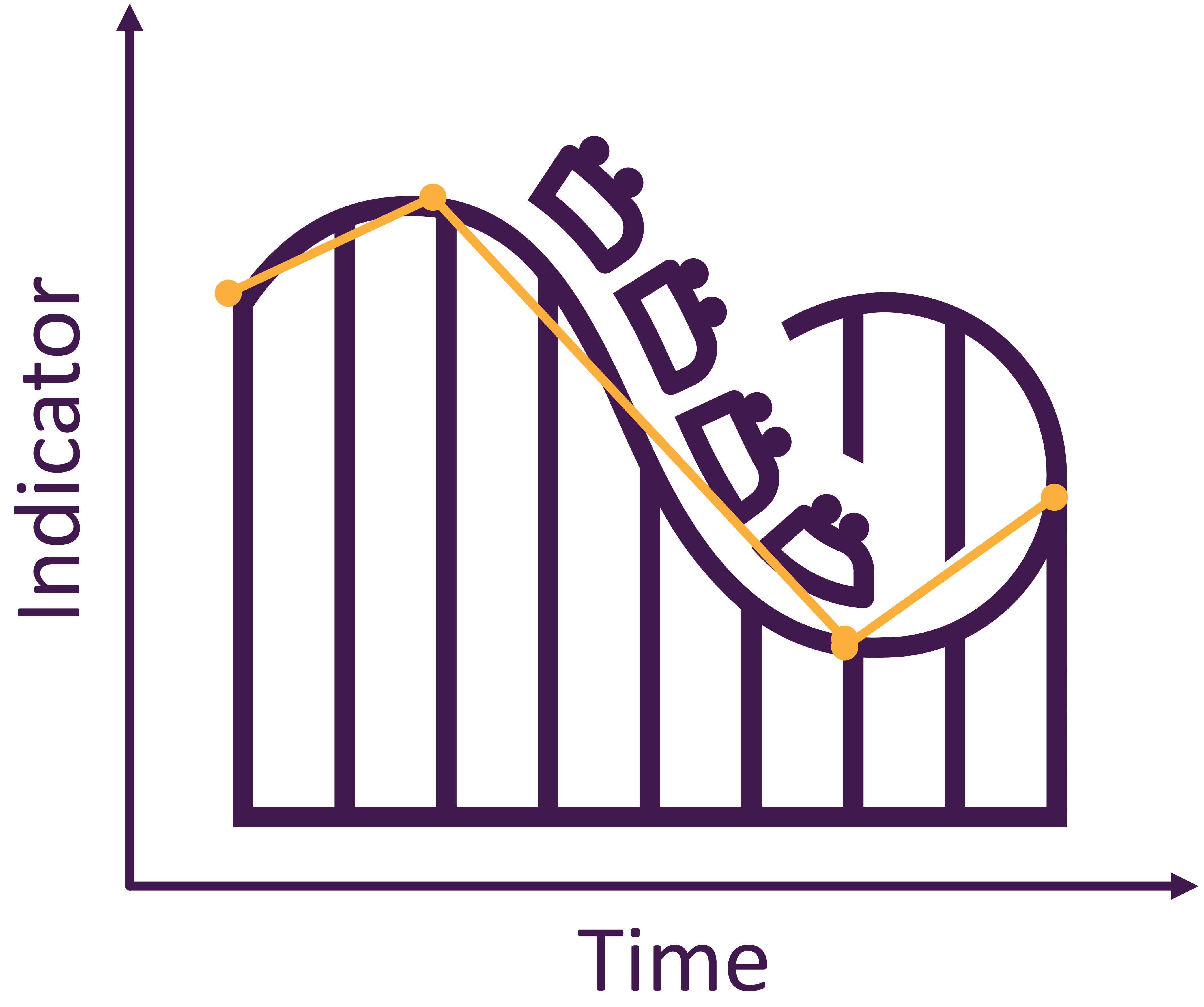
Brian Gill
Director, REL Mid-Atlantic
Mathematica

Lauren Forrow
Senior Researcher
Mathematica

The problem: Unreliable performance measures

Performance measures based on small amounts of data are often unreliable

- It’s often important to measure performance of demographic groups, for which numbers might be small.
- Measurement error is the random difference between what is true and what is measured.
- It has a larger effect on measurements made using smaller amounts of data.
- Error can cause instability over time in indicators used as performance measures.



State agencies try to reduce measurement error by setting minimum group sizes, trading accuracy against equity

- Decisions are made with more accurate, reliable scores, but:
  - Small groups are invisible to performance measurement processes.
  - Doesn’t remove measurement error, which affects every measurement.
- **Resources may not go to the students who need them most.**

Strategy: Bayesian stabilization

Bayesian stabilization is a data-driven method to reduce error

- Stabilization can reduce measurement error by learning from patterns in the data.
  - Learning about one school from other schools across the state.
  - Learning about a school's performance in one year from its historical trend.
- The amount of stabilization a data point receives depends on:
  - How much information (sample size) it provides.
  - How extreme it is.
- Learning from other schools increases the precision and plausibility of the estimates – especially for small numbers of students.

	Less extreme value	**More extreme value**
Less information	Some adjustment	Most adjustment
More information	Little adjustment	Some adjustment

Data point receives: Less adjustment ⟷ More adjustment

Two case studies:
Stabilizing school performance indicators in Pennsylvania and New Jersey schools

Case study 1:
Pennsylvania

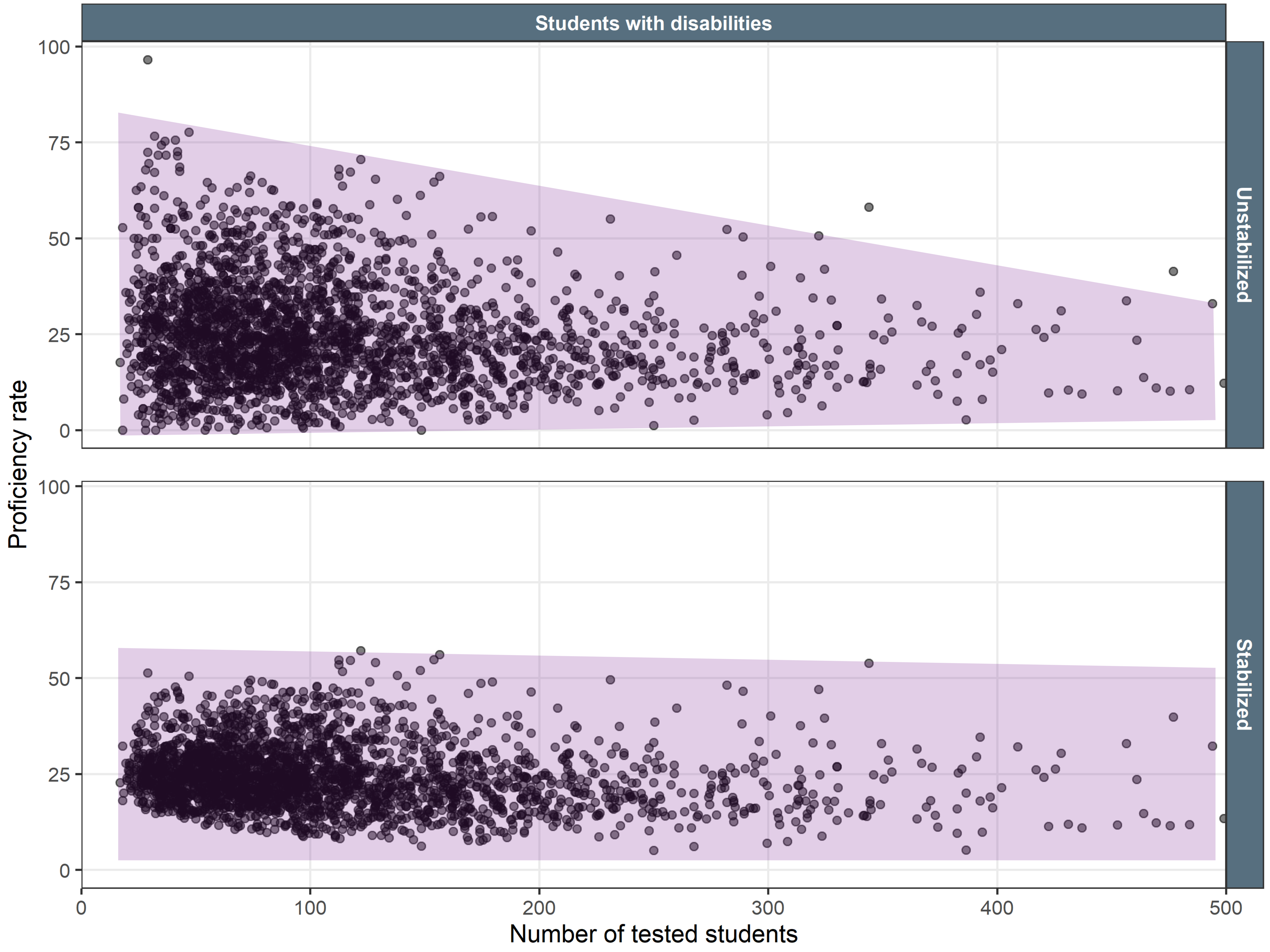
Case study 2:
New Jersey

We piloted stabilization using data from Pennsylvania schools

- To answer the research question: Does stabilization improve the reliability of subgroup academic proficiency rates used to identify low-performing schools?
- Pennsylvania Department of Education (PDE) provided two years of proficiency data for each subgroup and school.
- The team created models that:
  - Align with PDE's rules for identifying low-performing schools.
  - Combine proficiency rates from both years.
  - Learn from the same subgroup in different schools.

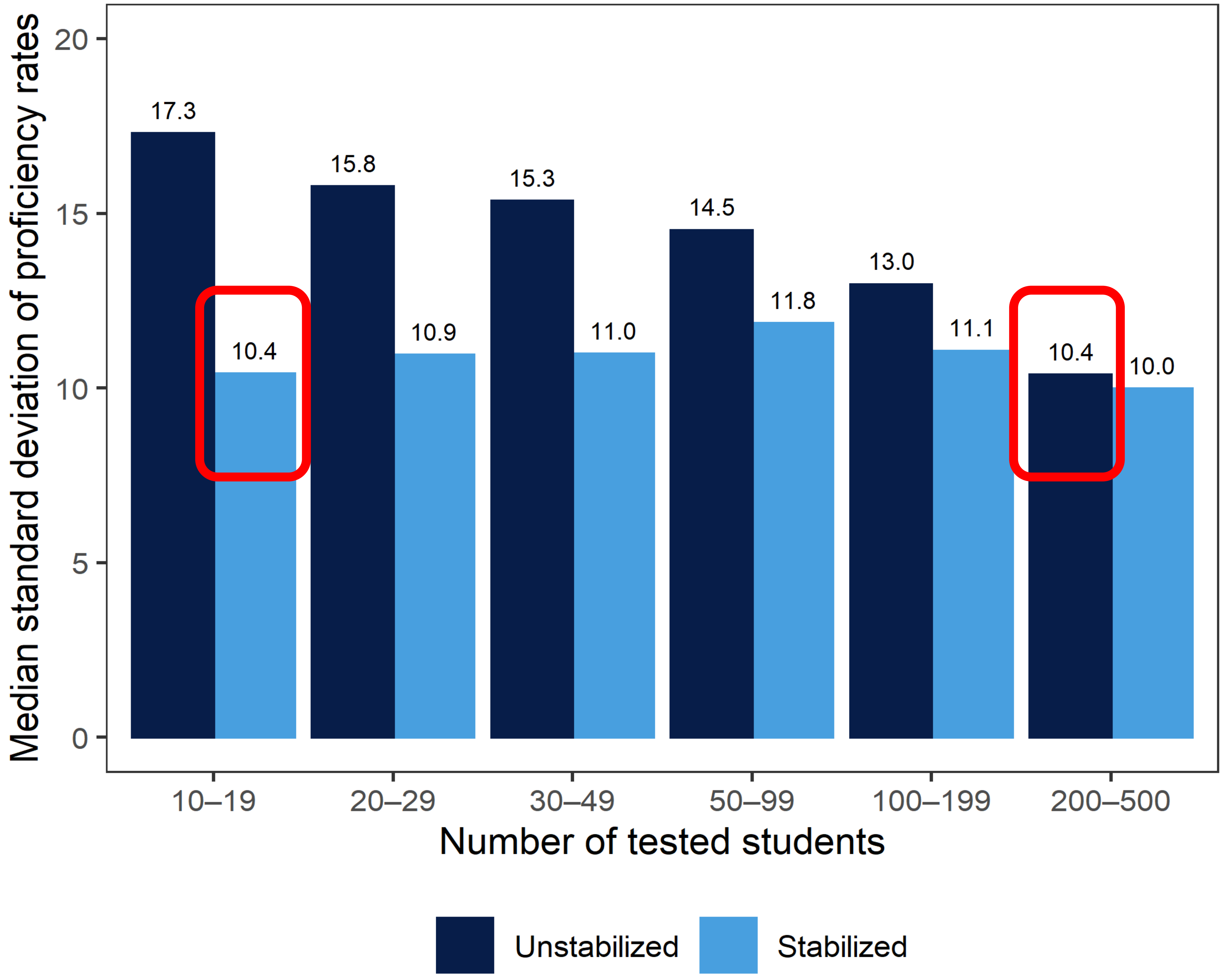
Case study 1: Pennsylvania

Findings suggest that stabilization improved statistical reliability



- **Unstabilized rates** showed a funnel pattern.
  - Small groups showed more variance.
  - Larger groups showed less variance.
  - The difference is likely due to measurement error.
- **Stabilized rates** had more uniform variance across group sizes.
- This suggests **stabilization improves statistical reliability**.

Stabilization may make it possible for Pennsylvania to include smaller subgroups in performance measurement processes



- For stabilized rates, the median standard deviation was relatively consistent across subgroup size categories.
- For very small subgroups, the stabilized standard deviation was close to the standard deviation for the largest groups.
- This indicates that stabilization may make it **possible to include smaller groups in performance measures**, without sacrificing statistical reliability.

We expanded on this work in New Jersey

- To answer the research questions:
  - Does stabilization reduce overrepresentation of small groups in the extremes of score distributions?
  - When applied to multiple indicators, does stabilization change which schools are designated for support and improvement?
- New Jersey of Education (NJDOE) provided data for all indicators from up to five school years.
  - Data availability varied by indicator.
- We created one model that could be applied across multiple indicators and used it to test how stabilization may change the list of "low-performing" schools.

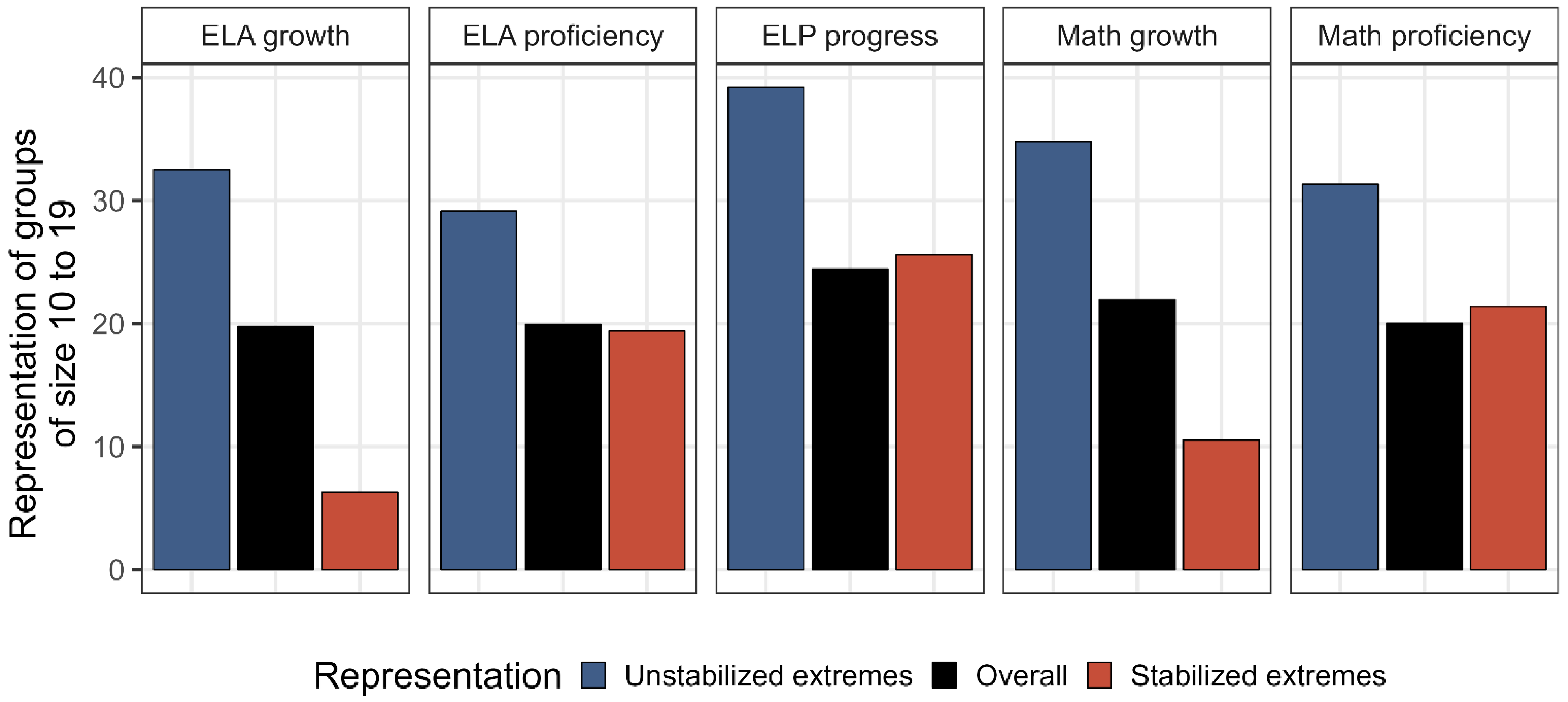
Case study 2: New Jersey

Stabilization improved reliability of test-based indicators and changed the list of low-performing schools

- Reliability measured by reducing **overrepresentation of small groups** in the extremes of the score distributions.

- Of 72 schools identified as lowest-performing, 17 would move off the list after stabilization, replaced by 16 others.
  - Fewer smaller schools were identified when using stabilized test indicator data.

Stabilization alleviated overrepresentation in the extremes of the score distribution for groups of 10-19 students



Conclusions and future directions

Stabilization can support progress toward accuracy and equity

- In our study in PA, stabilization improved the statistical reliability of school performance indicators enough to include groups of 10-19 students in the performance measurement process.
- In our study in NJ, stabilization reduced overrepresentation of small groups in the extremes of score distributions and changed which schools were identified as low-performing.
- Applying stabilization can reduce measurement error and may help states ensure that resources go to the students who need them most.

Challenges and supports

Challenges

Communication: Bayesian stabilization increases the complexity of performance assessment systems, so adoption will have to go hand in hand with enhanced communication to stakeholders.

Implementation: For states that don't have strong technical departments or resources to devote to training and computing, conducting Bayesian analyses may be a challenge.

Supports

Multiple communication tools: REL Mid-Atlantic and IES developed an infographic and blog posts that can support discussions on this topic.

Infographic

The Accuracy 4 Equity (A4E) tool: REL Mid-Atlantic and IES are developing a free tool to support this process in a transparent, intuitive manner. It is expected to be available on the IES website in early 2025.

Studies and resources are available at IES

PA Report

NJ Report

Infographic

References

[1] Forrow, L., Starling, J., & Gill, B. (2023). *Stabilizing subgroup proficiency results to improve the identification of low-performing schools* (REL 2023-001). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/rel/Products/Publication/106926

[2] Rosendahl, M., Gill, B., & Starling, J. E. (2024). *Stabilizing school performance indicators in New Jersey* (REL 2025-009). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/rel/Products/Publication/108130

Disclaimer

This presentation 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.

Appendix: PA model specification

Likelihood: $\bar{y}_j \sim N\left(\alpha_0 + \alpha_j, \frac{\sigma^2}{\bar{n}_j}\right)$

Priors:

$$\alpha_0 \sim N(0,1)$$
$$\alpha_j \sim N(0,\sigma_\alpha^2)$$
$$\sigma_\alpha, \sigma \sim N^+(0,1)$$

We fit this model separately to data for each subgroup.

Notation

Data

- $\bar{y}_j$ is the combined two-year proficiency rate for school j.
- $\bar{n}_j$ is the average number of tested students across years for school j.

Parameters

- α_0 is the overall average proficiency rate.
- α_j is the difference between school j's proficiency rate and the overall average.
- σ^2 is residual variance, weighted by sample size.

Appendix: NJ model specification

$$y_{j,t} = \alpha + \alpha_j + (\beta + \beta_j)t + (\gamma + \gamma_j)C_t + \epsilon_{j,t}$$

$$\epsilon_{j,t} \sim N\left(0, \frac{\sigma^2}{n_{j,t}}\right)$$

- $\alpha \sim N(0,1)$: **Overall intercept**, representing the average intercept for all schools.
- $\alpha_j \sim N(0, \sigma^2{}_{\alpha})$: **School-specific intercept**, representing the difference between overall performance for school j and the overall performance of schools on average.
- $\beta \sim N(0,1)$: **Overall slope**, representing the average change over time for all schools.
- $\beta_j \sim N(0, \sigma^2{}_{\beta})$: **School-specific slope**, representing the difference between average change over time for all schools and change over time for school j.
- C_t: **Indicator variable**, which is 0 for years preceding the COVID-19 pandemic (years before 2020) and 1 for years during and after the pandemic.
- $\gamma \sim N(0,1)$: **Overall effect of the COVID-19 pandemic**, representing the average effect for all schools.
- $\gamma_j \sim N(0, \sigma^2{}_{\gamma})$: **School-specific effect of the COVID-19 pandemic**, representing the difference between overall COVID-19 effects and COVID-19 effects for school j.
- $\sigma, \sigma_{\alpha}, \sigma_{\beta}, \sigma_{\gamma} \sim N^{+}(0,1)$: **Variance terms.**