Stabilizing measures to reconcile accuracy and equity in performance measurement

James E. Pustejovsky

Society for Research on Educational Effectiveness
September 29, 2023
Arlington, VA

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Shrinkage estimators

Partial pooling

Borrowing information

Bayesian stabilization

Implementation differences

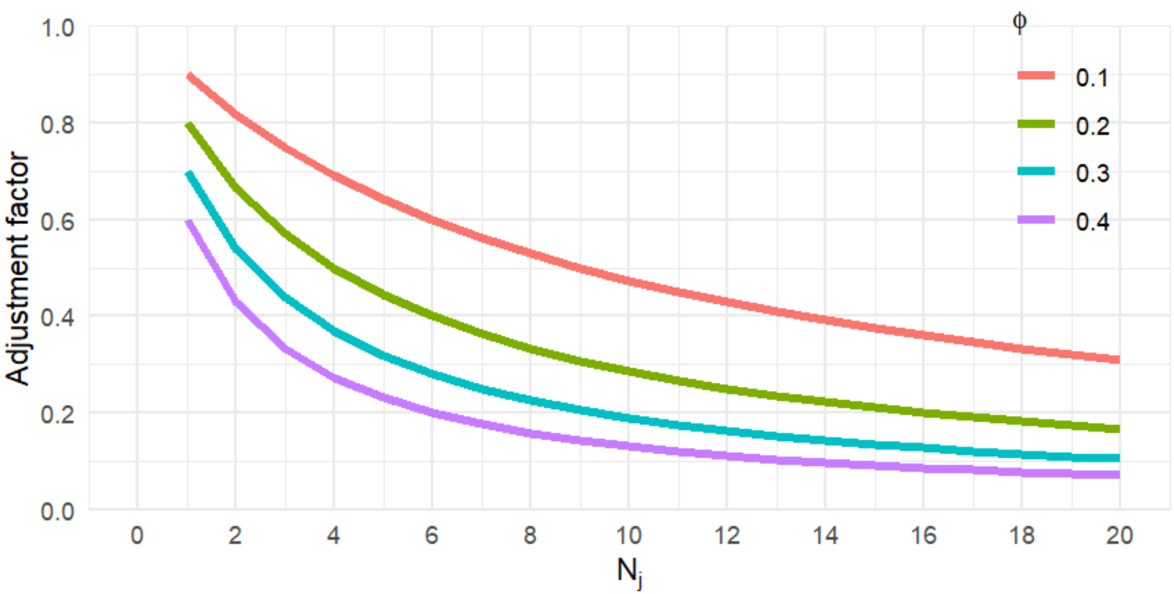
	McCaffrey	Forrow	Gellar
What scores to model?	Individual-level growth scores	Aggregate proficiency rates	Average scores
Where to borrow?	Past years, across subjects (**only within schools**)	Other schools	Other schools, past years, other subjects
What to condition on?	Response patterns	Specific groups of students	Subject, grade level, year
How to estimate?	Bespoke moment estimation	Full Bayes	Full Bayes

Implications of borrowing across schools

- Adjustments depend strongly on sample size. If borrowing across schools:

$$\hat{\mu}_j = \bar{y}_j + \frac{1}{1 + \frac{\phi}{1-\phi}\, N_j}\left(\hat{\mu}_\bullet - \bar{y}_j\right)$$

  - For school below the mean, reducing sample size will improve the proficiency/growth estimate.



Implications of subgroup analysis

Adjustment models within each of several specific groups:

$$Y_{ij}^g = \mu^g + v_j^g + e_{ij}^g, \qquad v_j^g \sim N(0, \tau_g^2), \quad e_{ij}^g \sim N(0, \sigma_g^2)$$

for groups $g = 1, \ldots, G$.

- Each group gets adjusted towards a different mean, by a different factor.

As a multivariate model:

$$\begin{pmatrix} Y_{ij}^1 \\ Y_{ij}^2 \\ \vdots \\ Y_{ij}^G \end{pmatrix} = \begin{pmatrix} \mu^1 \\ \mu^2 \\ \vdots \\ \mu^G \end{pmatrix} + \begin{pmatrix} v_j^1 \\ v_j^2 \\ \vdots \\ v_j^G \end{pmatrix} + \begin{pmatrix} e_{ij}^1 \\ e_{ij}^2 \\ \vdots \\ e_{ij}^G \end{pmatrix}, \qquad \begin{pmatrix} v_j^1 \\ v_j^2 \\ \vdots \\ v_j^G \end{pmatrix} \sim N(\mathbf{0}, \mathbf{T}), \qquad \begin{pmatrix} e_{ij}^1 \\ e_{ij}^2 \\ \vdots \\ e_{ij}^G \end{pmatrix} \sim N(\mathbf{0}, \mathbf{\Sigma})$$

Model building and model checking

- Estimates based on all of these models are **model-assisted**.
    - How to develop the model for specific applications?
    - How to check that the model is suitable for purpose?
- McCaffrey: further model development seems challenging because of bespoke moment estimation framework. Would shifting to more conventional tools help in developing more stable and interpretable models?
- Forrow and Gellar: With the Bayesian approach, **posterior predictive checks** are useful general-purpose technique for evaluating model fit. But what summary quantities should the analyst check?
- All: is correlation between school size and school performance level an issue for these estimators?