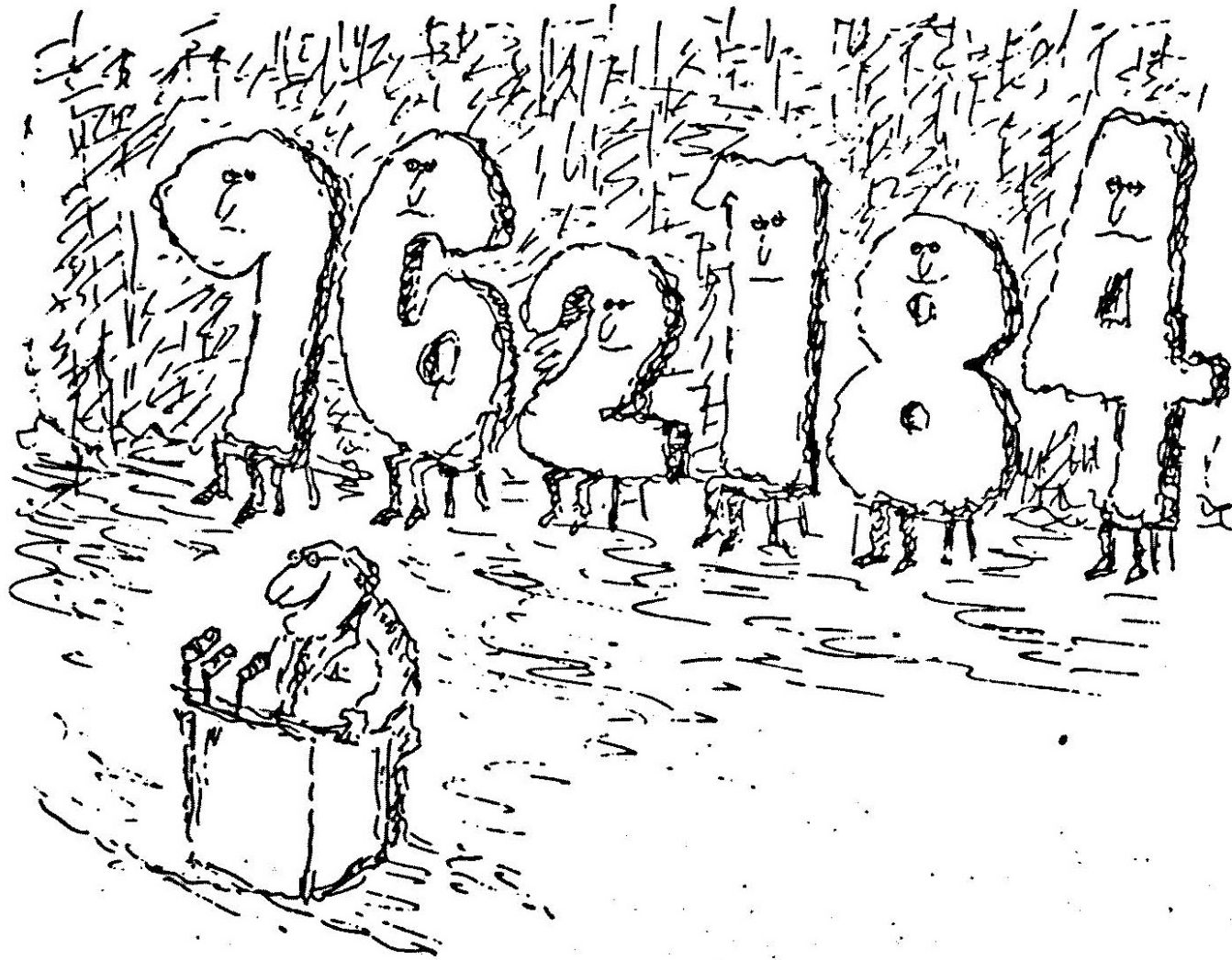


Comments on the Inter-Temporal Stability of Teacher Effect Estimates



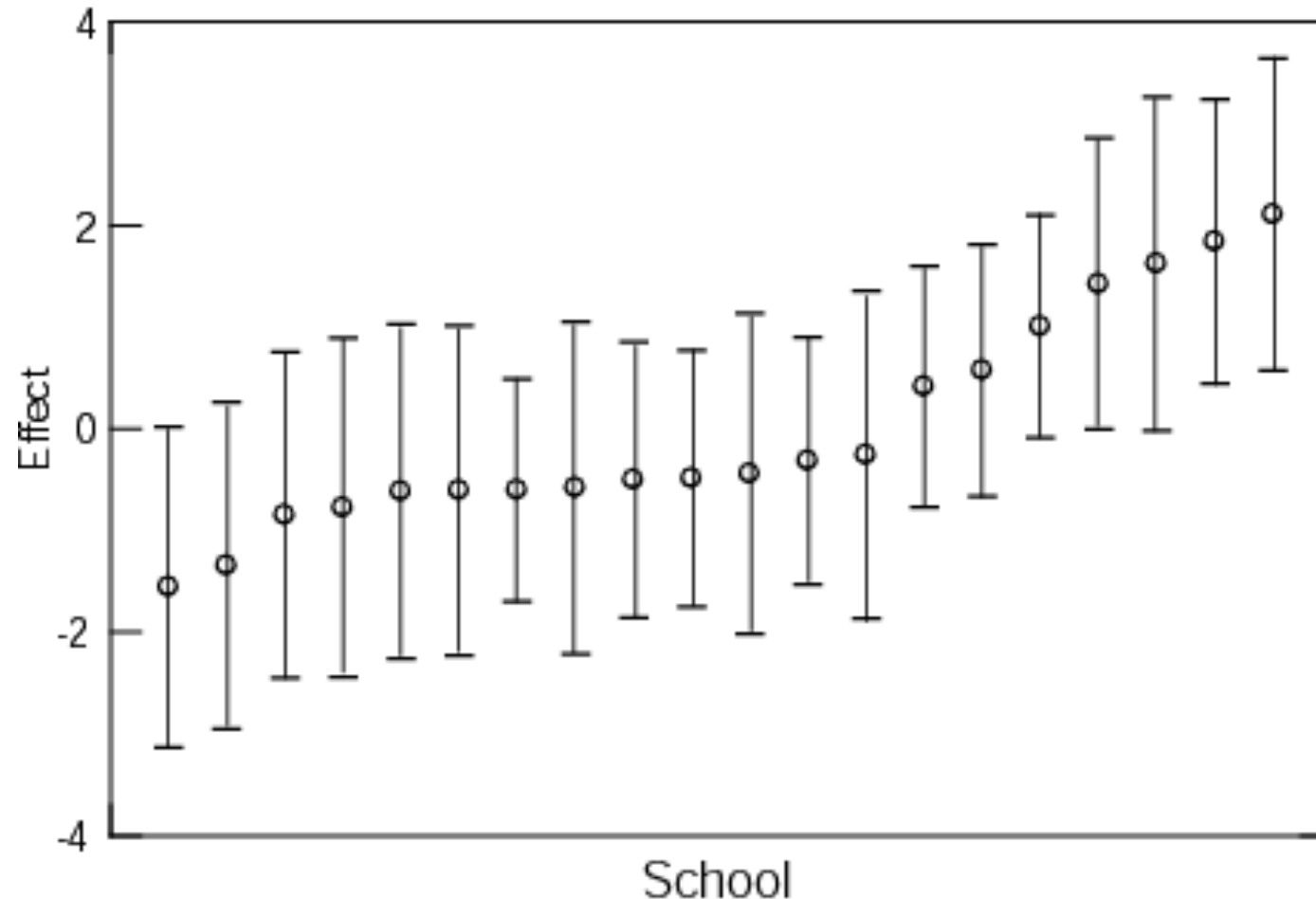
"Tonight, we're going to let the statistics speak for themselves."

Jee-Seon Kim
University of Wisconsin, Madison

Estimating Teacher Effectiveness—Why is it So Difficult?

- ❑ I agree with Lockwood, McCaffrey, and Sass' view on the careful use of value-added estimates.
- ❑ It is difficult to isolate the effects of teachers in complex educational environments (e.g., omitted variables, confounders).
- ❑ As a consequence, a model may be misspecified:
The independence assumption between predictors and random components are often violated → *omitted variable bias*
- ❑ Even if a model is correctly specified, the precise estimation of random entities (teachers, schools, etc.) is more difficult than fixed parameters.
 - “Estimation” of a random variable is referred to as *prediction*.
 - A prediction interval is wider than a confidence interval.

Rankings of Schools and 95% Prediction Intervals



- Optimal (unbiased & minimum variance) best linear unbiased prediction (BLUP) of value-added school effects at level 3 (Frees & Kim, 2006)
- Still, almost all prediction intervals overlap.
- Similar issues are discussed in Longford's (1986), Goldstein (1993), Raudenbush and Willms (1995), and Snijders and Bosker (1999).

Inter-Temporal Stability of Teacher Effects

- ❑ There exist time-varying and time-constant teacher effects.
- ❑ Ideally, inter-temporal variability should mainly reflect time-varying teacher effects (preparation, motivation, health, mobility, etc.).
- ❑ However, low year to year correlations (0.3 to 0.5) are more likely due to difficulty in measuring teacher qualities precisely.

Difficulties in Teacher Effectiveness Study

- ❑ Many educational studies are observational and the analyst doesn't always have the ability to "go and get all the important variables".
- ❑ Omitted variable bias and selection bias become particularly problematic when schools/teachers are heterogeneous.
- ❑ Randomized trials would reduce omitted variable bias and allow for causal inferences. However, there still exist heterogeneity issues. Estimates may be (close to) unbiased but are imprecise to draw strong conclusions.

A Synthesis of Experimental and Observational Studies

- ❑ One should try to remove systematic differences and selection bias by randomization whenever possible.
- ❑ One should try to get as many relevant variables as possible (e.g., student, neighborhood, and district variables) for the proper estimation of school and teacher effects (“auxiliary” variables). A simulation study showed that a omitted student variable severely affects the estimation of higher level effects (Kim & Frees, 2006).
- ❑ Also, if possible, compare homogenous populations (similar types of schools).
- ❑ Strategies for statistical analysis – VAM, model specification tests, robust estimators – should come after this effort of obtaining “as good as it gets” data.

Model Specification Tests & Robust Estimators

1. DIAGNOSIS

- ❑ Hausman test (Hausman, 1978): Fixed effects vs. Random effects (RE) estimators for panel data models (two-level random intercept models)
- ❑ Omitted variable tests in multilevel models (Kim & Frees, 2006):
Severity of omitted variable bias at multiple levels

2. TREATMENTS

- ❑ Alternative estimators for removing or reducing bias
- ❑ Meyer (yesterday). Conditional RE + multivariate shrinkage
- ❑ Raudenbush (yesterday). RE + adaptive centering
- ❑ Kim and Frees (2007). Generalized method of moments (GMM) estimator provides a “robust to efficient” continuum of estimators.
One can find the most efficient among unbiased ones.
- ❑ In NELS88, the effects of previous year score, SES, and minority would be overestimated, whereas the effects of several school variables would be underestimated if the efficient ML or GLS estimator were employed.

Concluding Remarks

- By exploiting hierarchical structure, we can test omitted variable bias and obtain robust estimators. However, it is very hard to estimate teacher or school effects accurately and precisely, even using multilevel models.
- These difficulties should be recognized during planning and designing a study, rather than at the data analysis stage.
- Sometimes, there is not enough information in the data to obtain unbiased estimates or to evaluate teachers or schools credibly. In these cases, it would be inappropriate to make strong inferences or high-stakes decisions, as Lockwood, McCaffrey, and Sass stated.
- Classifying teachers/schools into homogenous groups (e.g., teachers with non-overlapping prediction intervals) might be more reliable than ranking.
- I hope that the awareness of these difficulties can stimulate communication among policymakers, researchers, and practitioners on what should not be ignored in the future.