RAND EDUCATION

The (Sometimes Harsh) Reality of Longitudinal Student Achievement Modeling

J.R. Lockwood

November 7, 2005

This material is based on work supported by the National Science Foundation under Grant No. ESI-9986612 and the Department of Education Institute of Education Sciences under Grant No. R305U040005. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of RAND or these organizations.

Outline

- Background for case study
- Challenges:
  - Variance partitioning
  - Missing data
  - Structuring hypothesis about cumulative effects
  - Model estimation with nuisance correlations
- Closing thoughts

Case Study: The RAND Mosaic II Project

- **Study goal: Examine relationships between "reform-oriented teaching practices" and student mathematics and science achievement**
  - **Hands-on and investigative activities**
  - **Exploration of students' thinking**
- **Mosaic I: Found small positive effects of one year of exposure to reform teaching**
- **Mosaic II: Improved methods for assessing reform teaching, additional consideration of open-ended assessments, followed students through three years of exposure**

Data Structure

Basic design replicated in 5 cohorts

MC = Multiple Choice OE = Open-Ended

Cohort 1	Mathematics	Grades 3-5	SAT9 MC
Cohort 2	Mathematics	Grades 7-9	SAT9 MC
Cohort 3	Mathematics	Grades 6-8	SAT9 MC (PR,PS) + OE
Cohort 4	Science	Grades 3-5	SAT9 MC + OE
Cohort 5	Science	Grades 6-8	SAT9 MC + OE

- MC administered in Years 1, 2 and 3 in all cohorts
- OE administered in some years for some cohorts

Specific Data Elements

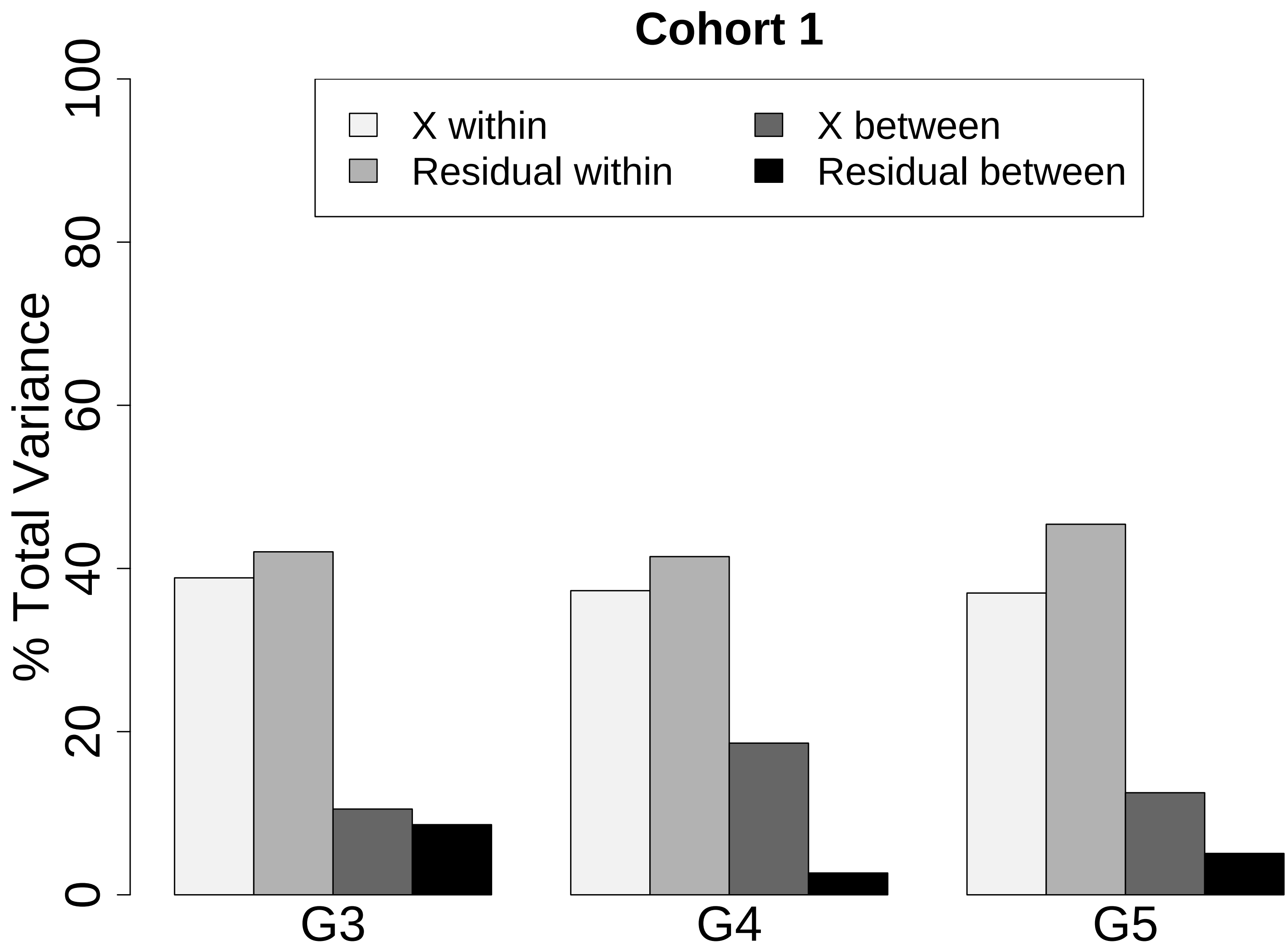
- Student Level:
  - Assessment scores (usually scaled) from Years 1-3
  - Assessment scores from districts and/or state tests from "Year 0", the year prior to the study
  - Background variables generally including: race, FRL, LEP, special education, gifted, and age (used to proxy for "behind cohort")
  - Links to teachers in Years 1-3
- Teacher Level:
  - Measures of teaching practices and background characteristics obtained from multiple methods (surveys, lesson logs, instructional vignettes)

- **Challenges:**
  - **Variance partitioning**
  - **Missing data**
  - **Structuring hypothesis about cumulative effects**
  - **Model estimation with nuisance correlations**

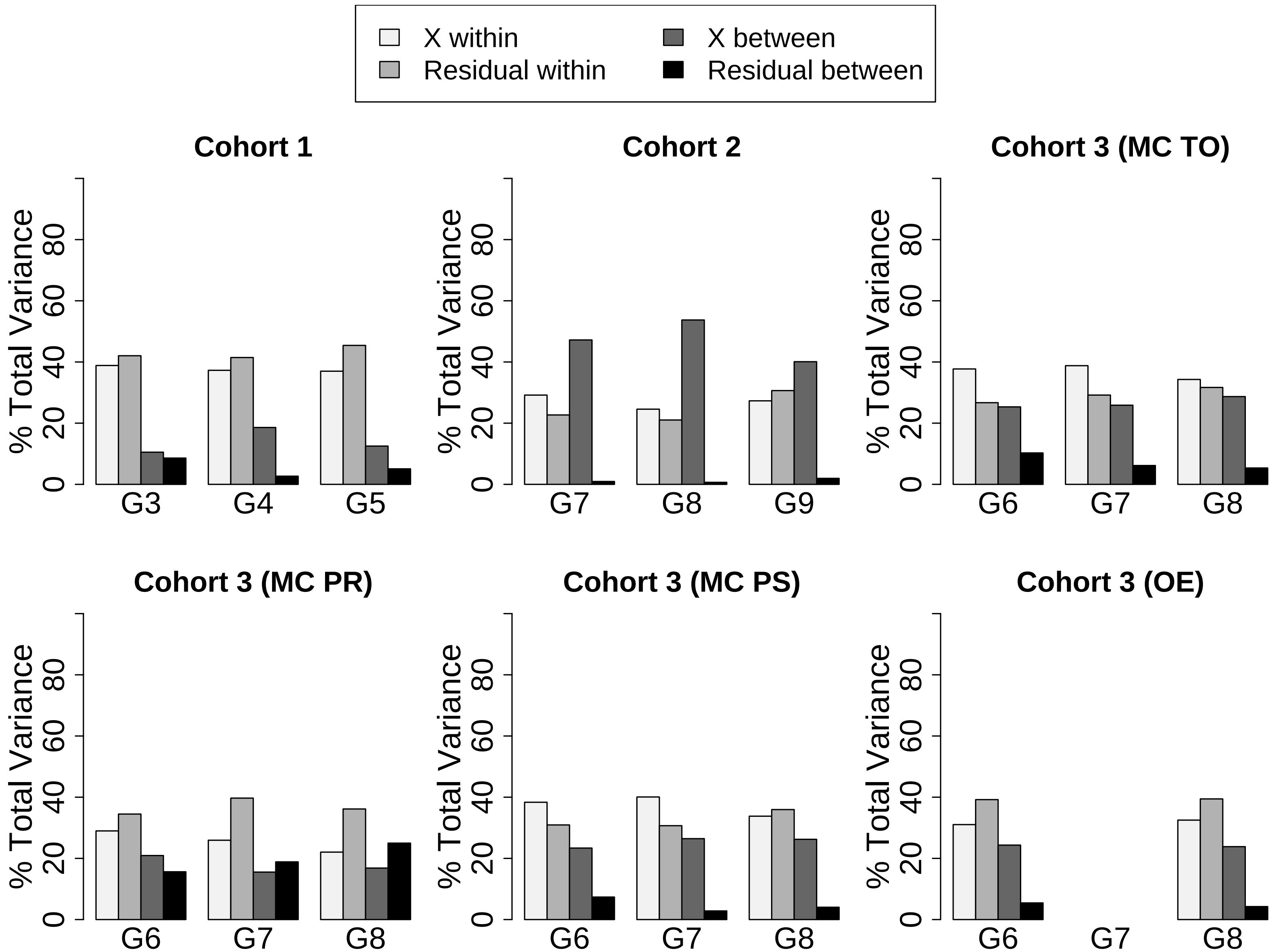
Exploratory Analyses on Scores Suggests Small Explainable Variance

- **For each cohort, year, and outcome, decompose variance of level scores into four sources:**
  1. **Background variables and Year 0 scores *within teachers* (“X within”)**
  2. **Unexplained variance *within teachers***
  3. **Aggregate background variables and year 0 scores *between teachers* (“X between”)**
  4. **Unexplained variance *between teachers***
- **(3 + 4) bounds R^2 of main effect of a teacher-level predictor with respect to achievement levels**
- **Ideally: (3 + 4) is big and 4$>>$3**
- **Empirically: (3 + 4) $<<$ (1 + 2) and 4 is generally tiny**

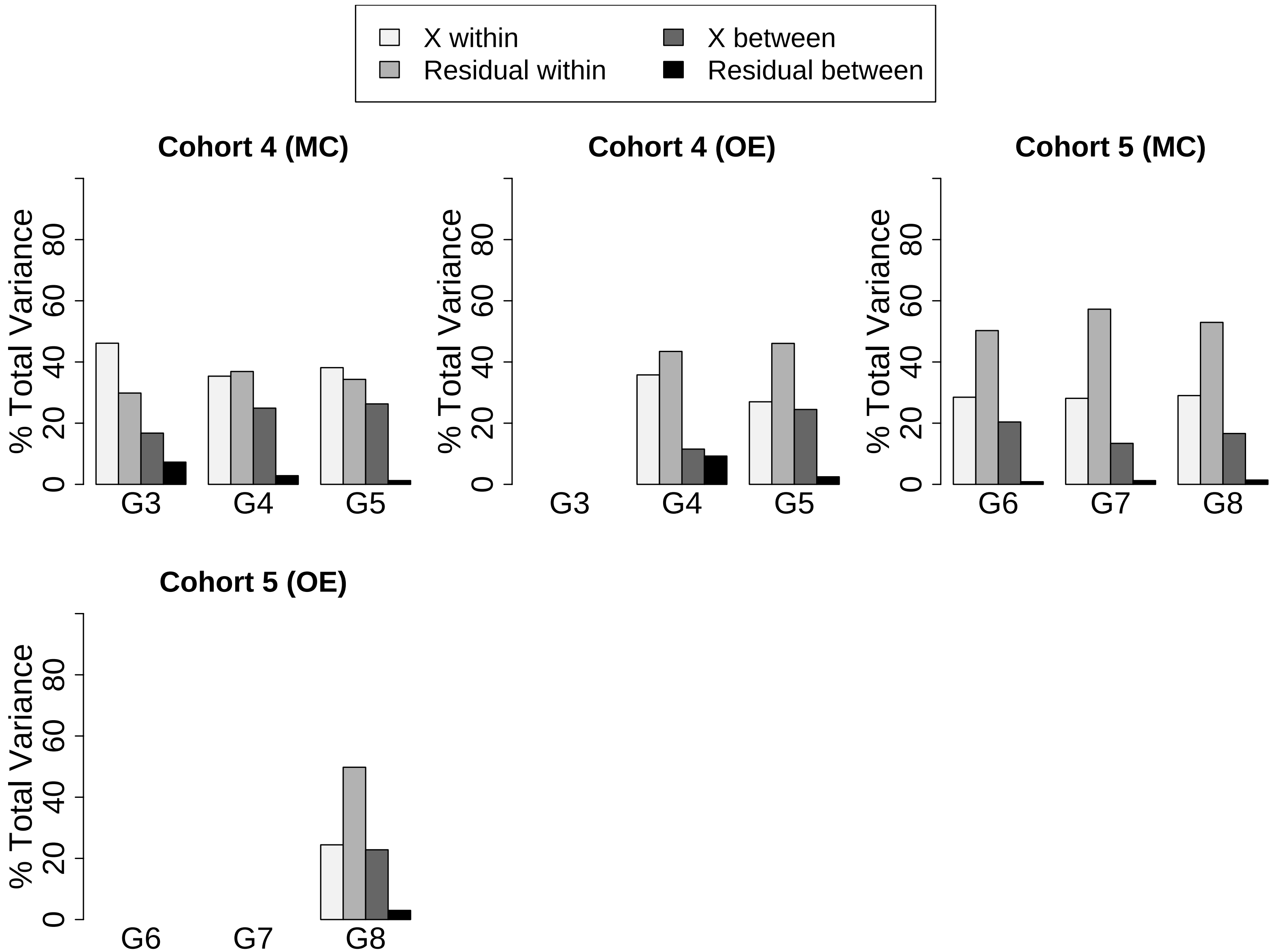
Variance Decomposition of Level Scores (Example)



Variance Decomposition of Level Scores: Mathematics

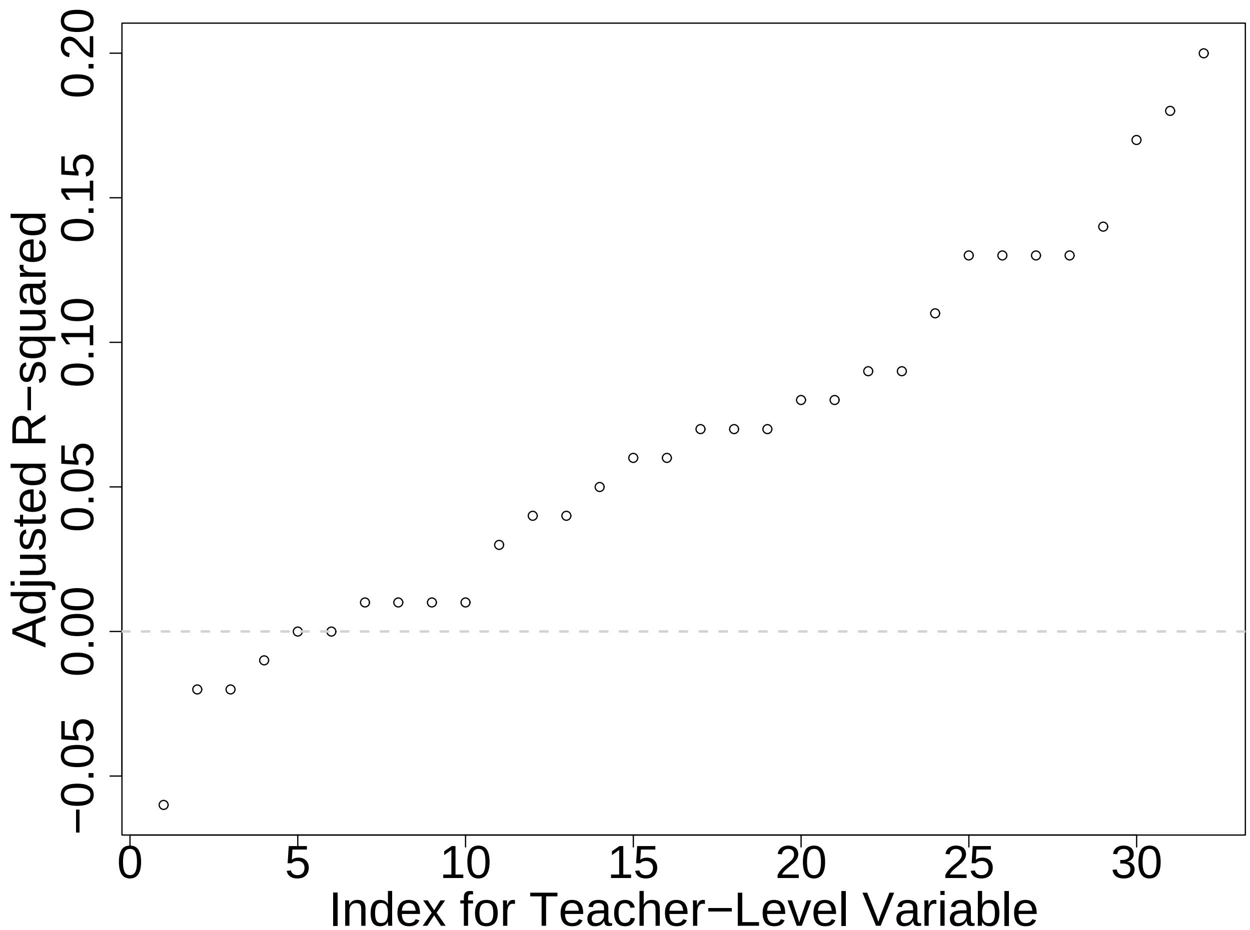


Variance Decomposition of Level Scores: Science



Can Strong Relationships Between Aggregate X and Teacher Predictors Leave Hope of Big Effects?

Not Really



- **Challenges:**
  - Variance partitioning
  - **Missing data**
  - Structuring hypothesis about cumulative effects
  - Model estimation with nuisance correlations

Missing Data Compounds Quickly in Longitudinal Studies

- **Many levels of missing data:**
  - **Student test scores (mobility, absenteeism)**
  - **Student covariates**
  - **Student-teacher links**
  - **Unit-level teacher non-response**
  - **Item-level teacher non-response**
- **Compounds roughly geometrically across years**
- **Particular challenge with longitudinal models and cumulative effects is the need for full information on *exposure history***

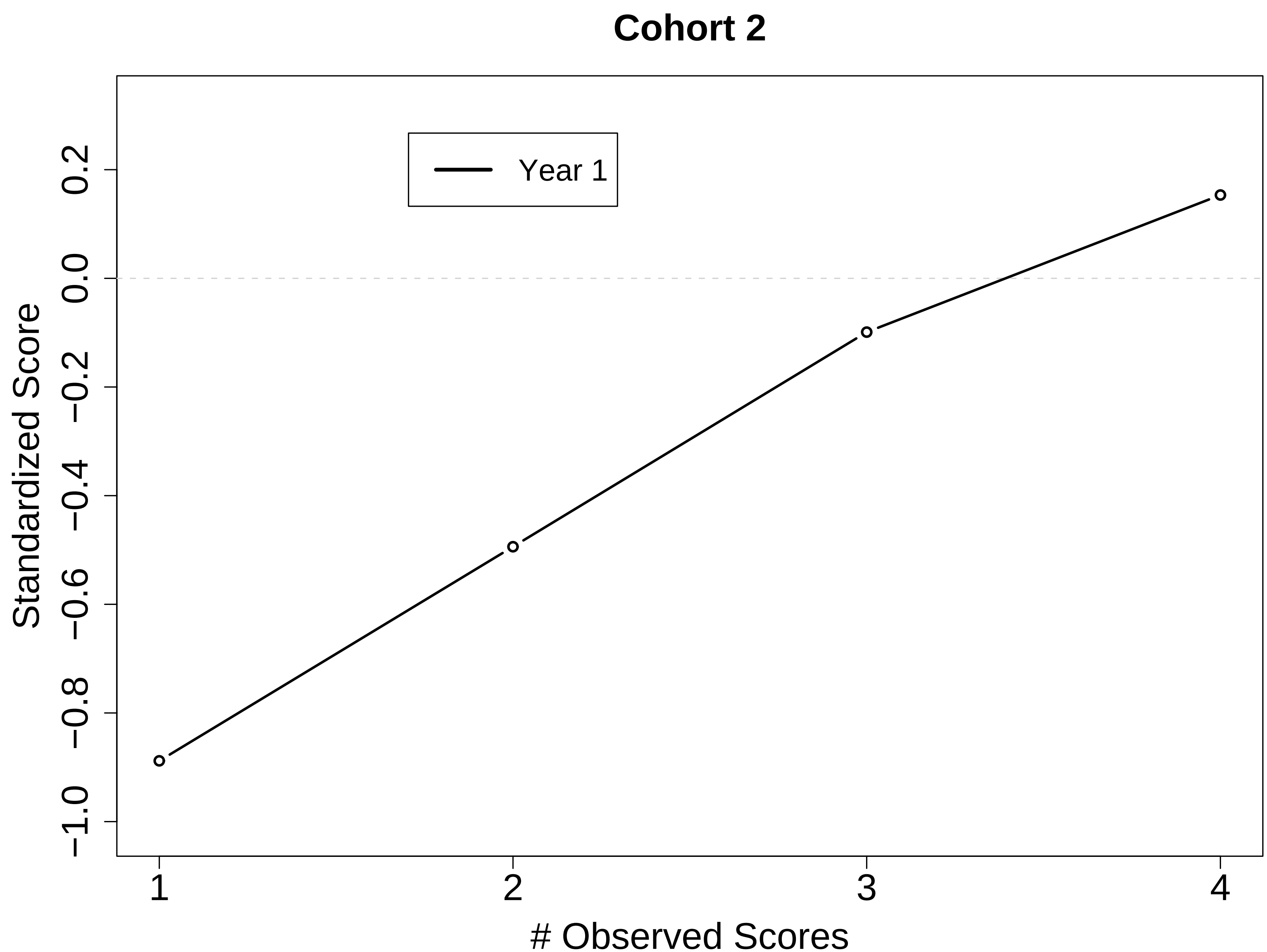
Relatively Small Fraction of Students Have Complete Information

	1	2	3	4	5
Total # Students	2415	5173	3460	1864	7827
Y1 Scores (%)	71	84	59	63	76
∩ Y2,Y3 Scores (%)	52	63	24	32	54
∩ Demographics (%)	52	63	24	32	54
∩ Y0 Scores (%)	45	54	22	NA	48
∩ Links to Responders (%)	25	23	15	16	35
∩ Links to Complete Resp. (%)	19	12	10	11	18

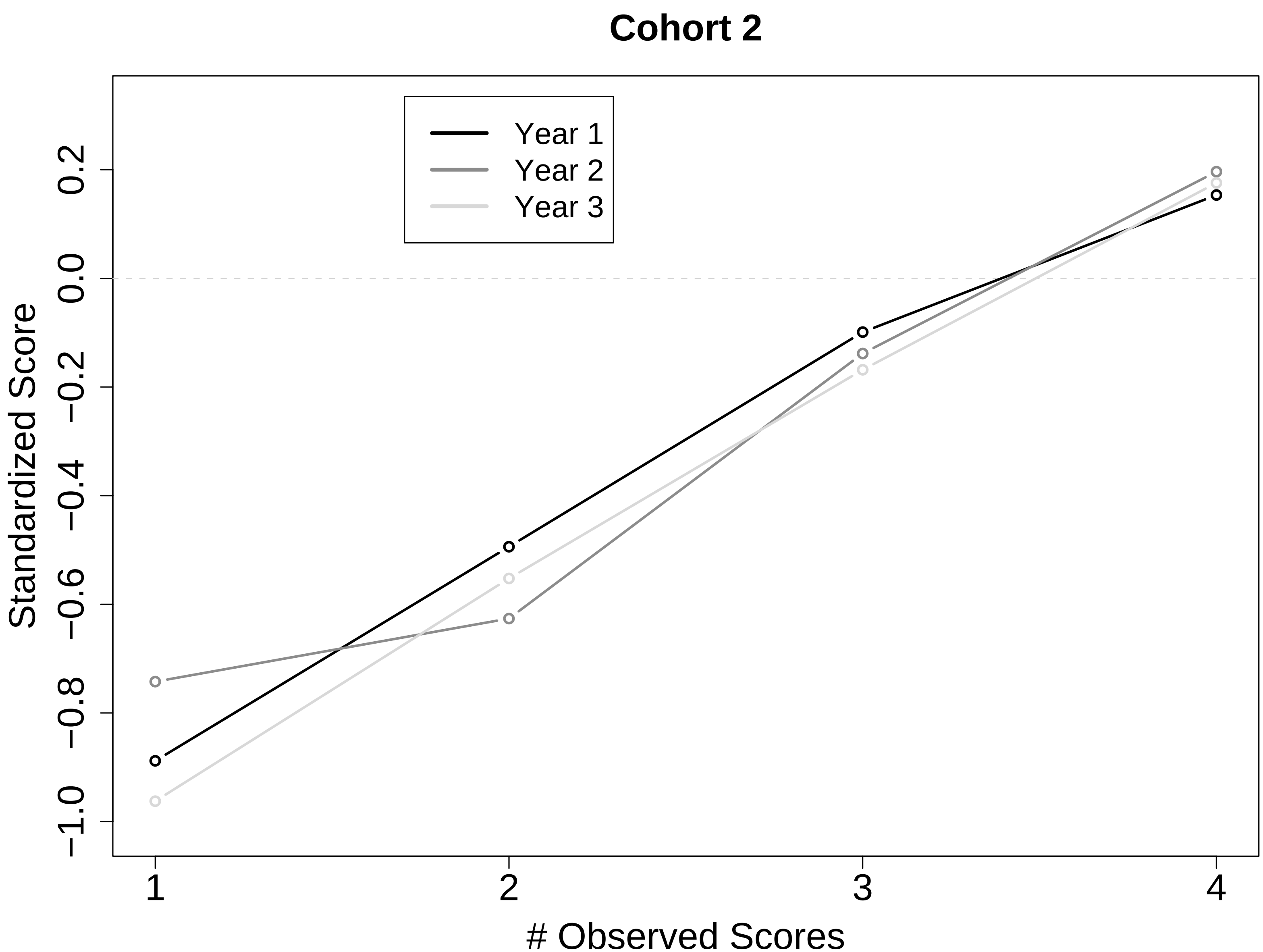
Complete Case Analysis?

- Loss of power
- Lack of generalizability: complete cases are truly a selected sample

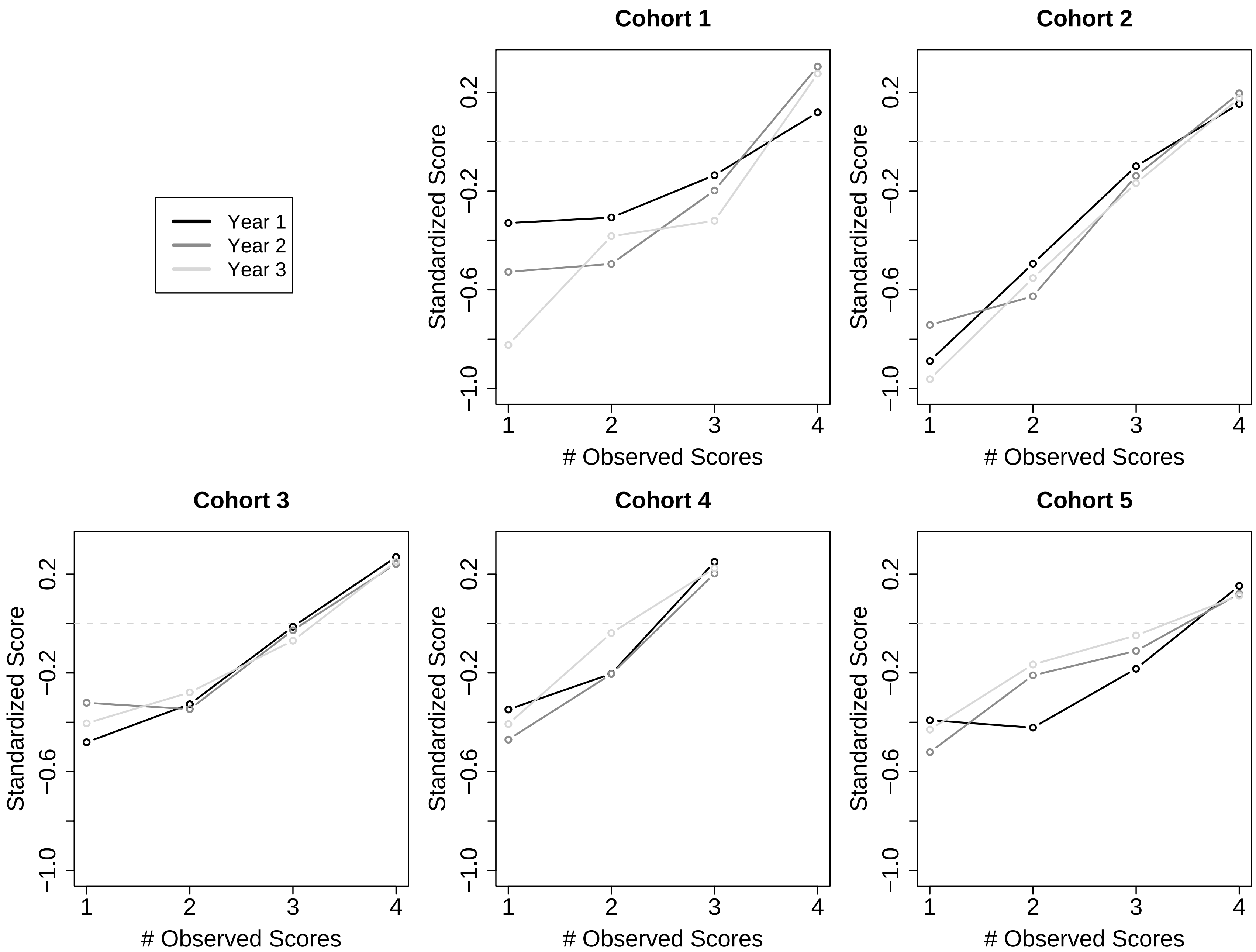
Average Scores Conditional on Number of Observed Scores



Average Scores Conditional on Number of Observed Scores



RAND


Year 1
Year 2
Year 3
Cohort 1
Cohort 2
Cohort 3
Cohort 4
Cohort 5
Standardized Score
Observed Scores
0.2
−0.2
−0.6
−1.0
1
2
3
4

Calibrating the Information Contained in Missing Scores

- **Variance decomposition implies the maximal R^2 of a teacher-level predictor is on the order 10% or less**
- **For a continuous predictor, corresponds to a Cohen-type effect size of about 0.33**
- **In these data, approximately equal to the predictive value of knowing the student had one unobserved score sometime over 4 years**

Moving Forward with Missing Data: Multiple Imputation

- **Logic of multiple imputation**
  - **Obtain K realizations of missing data D_{mis} by sampling from $p(D_{mis}|D_{obs})$ to create K replicates of $D_{full} = (D_{obs}, D_{mis})$**
  - **Fit models to each D_{full} as if data were fully observed**
  - **Pool estimates and standard errors across model fits to obtain global inferences that account for uncertainty due to D_{mis}**
- **Conceptually straightforward**
- **Practical challenge is positing $p(D_{mis}|D_{obs})$ in such a way to maintain fidelity to multivariate structure of observed data**

Implemented Multi-Stage Multiple Imputation Procedure

Relied heavily on Schafer's `norm` package for R Environment

- **Student-level demographics and year 0 scores: use approximate joint distribution of demographics, year 0 scores, future scores and future exposures to teacher-level variables**
- **Item-level missingness of teacher variables: use approximate joint distribution of item responses and classroom aggregates of student scores and variables**
- **Missing student-teacher links: create "pseudo-teachers" who receive donated covariate vectors from actual teachers, obtained by informed hotdeck**
- **Missing test scores from Years 1-3: Imputed "on the fly" during model estimation ("data augmentation")**

Missing Data Had Surprisingly Little Leverage on Inferences

- **Point estimates for key teacher-level predictors were relatively robust to using complete cases versus sequences of nested sets of students with increasingly poorer observed information**
  - **E.g. missing no test scores, versus missing at most 1 test score, versus missing at most 2 test scores**
- **Standard errors for estimates under multiple imputation were consistently smaller**
  (rough approximation: $SE_{MI}/SE_{complete} \approx 2/3$)
- **Reasonable imputations + Complex model that implicitly downweights incomplete cases = Robustness**

- **Challenges:**
  - Variance partitioning
  - Missing data
  - **Structuring hypothesis about cumulative effects**
  - Model estimation with nuisance correlations

An Additive Model For Cumulative Exposure

- **General notion: Achievement at time t is (partially) a function of exposure to practices *up to and including time* t**

$$Y_{i1} = f_1(stuff, exposure_{-\infty}, \ldots, exposure_1) + error$$
$$Y_{i2} = f_2(stuff, exposure_{-\infty}, \ldots, exposure_1, exposure_2) + error$$
$$Y_{i3} = f_3(stuff, exposure_{-\infty}, \ldots, exposure_1, exposure_2, exposure_3) + error$$

- **Here f_t is arbitrary but we will be somewhat less ambitious**
- **Let P_j be the measure of a particular teacher characteristic or teaching practice for teacher j**

$$Y_{i1} = else + \delta_{11} P_{j(i,1)} + error$$
$$Y_{i2} = else + \delta_{21} P_{j(i,1)} + \delta_{22} P_{j(i,2)} + error$$
$$Y_{i3} = else + \delta_{31} P_{j(i,1)} + \delta_{32} P_{j(i,2)} + \delta_{33} P_{j(i,3)} + error$$

Model Subsumes Some Plausible Alternatives

$$Y_{i1} = else + \delta_{11} P_{j(i,1)} + error$$

$$Y_{i2} = else + \delta_{21} P_{j(i,1)} + \delta_{22} P_{j(i,2)} + error$$

$$Y_{i3} = else + \delta_{31} P_{j(i,1)} + \delta_{32} P_{j(i,2)} + \delta_{33} P_{j(i,3)} + error$$

- $(\delta_{21} = \delta_{31} = \delta_{32} = 0)$**: No cumulative effects: exposure this year affects *level score* this year**
- $(\delta_{11} = \delta_{21} = \delta_{31})$ **and** $(\delta_{22} = \delta_{32})$**: "Complete" cumulative effects: exposure this year affects *gain score* this year**
- **General structure with all six parameters unknown allows the data to inform the appropriate degree of accumulation**
- **Key function of interest is** $\delta_{total} = (\delta_{31} + \delta_{32} + \delta_{33})$
- **Also of interest might be** $\delta_{current} = (\delta_{11} + \delta_{22} + \delta_{33})/3$

Full Fixed Effects Structure

$$Y_{i1} = \mu_1 + \boldsymbol{X}_i^{'}\boldsymbol{\beta}_1 + \boldsymbol{Z}_{i0}^{'}\boldsymbol{\gamma}_1 + \delta_{11}P_{j(i,1)} + error_{i1}$$

$$Y_{i2} = \mu_2 + \boldsymbol{X}_i^{'}\boldsymbol{\beta}_2 + \boldsymbol{Z}_{i0}^{'}\boldsymbol{\gamma}_2 + \delta_{21}P_{j(i,1)} + \delta_{22}P_{j(i,2)} + error_{i2}$$

$$Y_{i3} = \mu_3 + \boldsymbol{X}_i^{'}\boldsymbol{\beta}_3 + \boldsymbol{Z}_{i0}^{'}\boldsymbol{\gamma}_3 + \delta_{31}P_{j(i,1)} + \delta_{32}P_{j(i,2)} + \delta_{33}P_{j(i,3)} + error_{i3}$$

- **$\boldsymbol{X}_i$: student background variables**
- **$\boldsymbol{Z}_{i0}$: student achievement on year 0 assessments**

- **Challenges:**
  - Variance partitioning
  - Missing data
  - Structuring hypothesis about cumulative effects
  - **Model estimation with nuisance correlations**

Error Terms Have Structure That Must Be Dealt With

- Correlation within students over time (unmeasured student effects)
- Correlation across students sharing a teacher this year (unmeasured teacher/classroom effects)
- Carry-over effects of past shared classrooms
- Addressing these nuisance correlations is necessary to obtain reasonable standard errors for parameters of interest

Parameterizing The Error Terms

$$error_{i1} = \theta_{j(i,1)} + \epsilon_{i1}$$

$$error_{i2} = \alpha_{21}\theta_{j(i,1)} + \theta_{j(i,2)} + \epsilon_{i2}$$

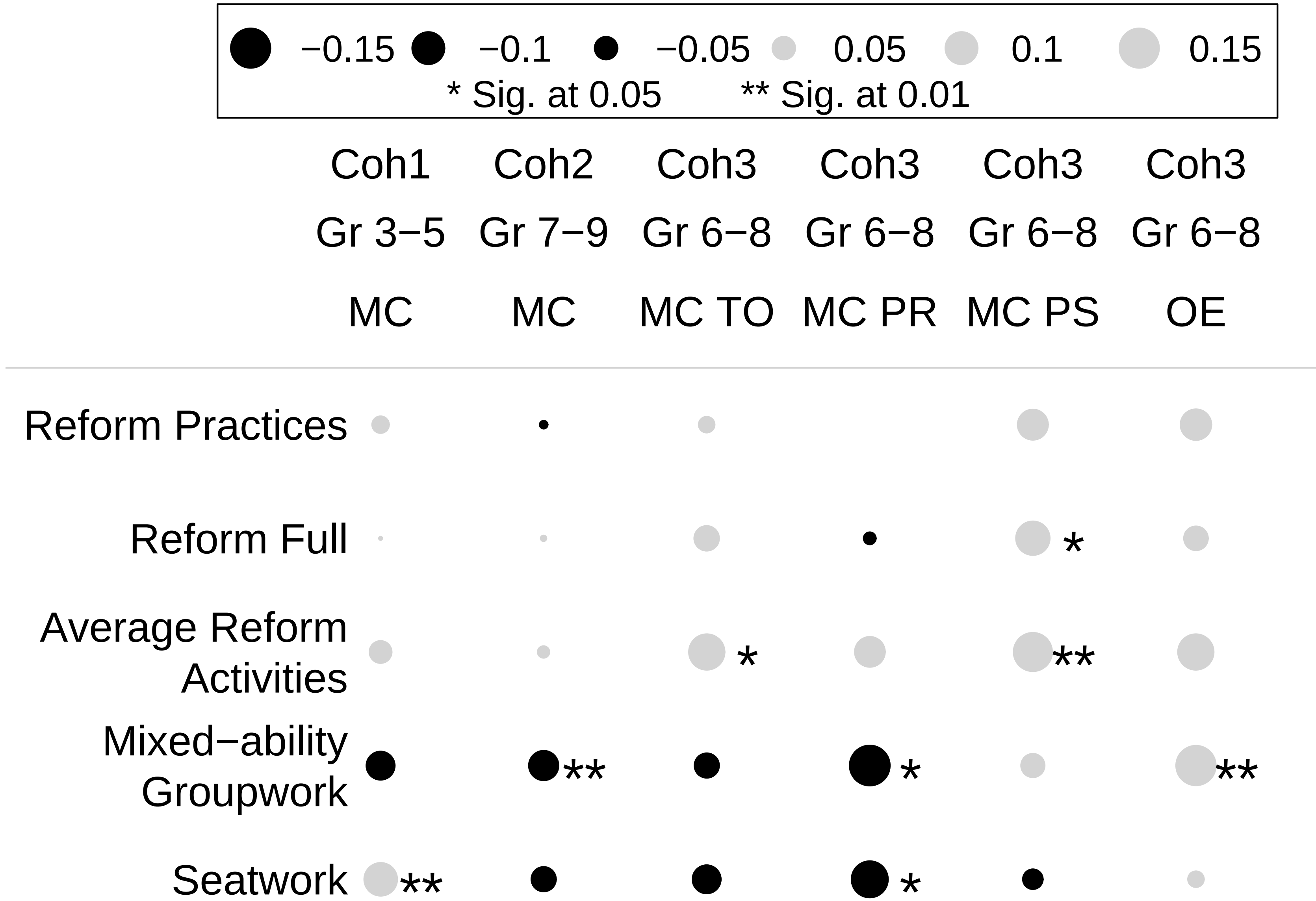
$$error_{i3} = \alpha_{31}\theta_{j(i,1)} + \alpha_{32}\theta_{j(i,2)} + \theta_{j(i,3)} + \epsilon_{i3}$$

- θ_j: **unobserved "teacher effects", treated as independent normal random effects with year-specific variance components**
- $(\alpha_{21}, \alpha_{31}, \alpha_{32})$ **"persistence parameters" that moderate the persistence of past unobserved teacher effects (estimated from data)**
- $(\epsilon_{i1}, \epsilon_{i2}, \epsilon_{13}) \sim N(0, \mathbf{\Sigma})$ **independently across students**
  - **Unstructured covariance proxies for omitted student variables**

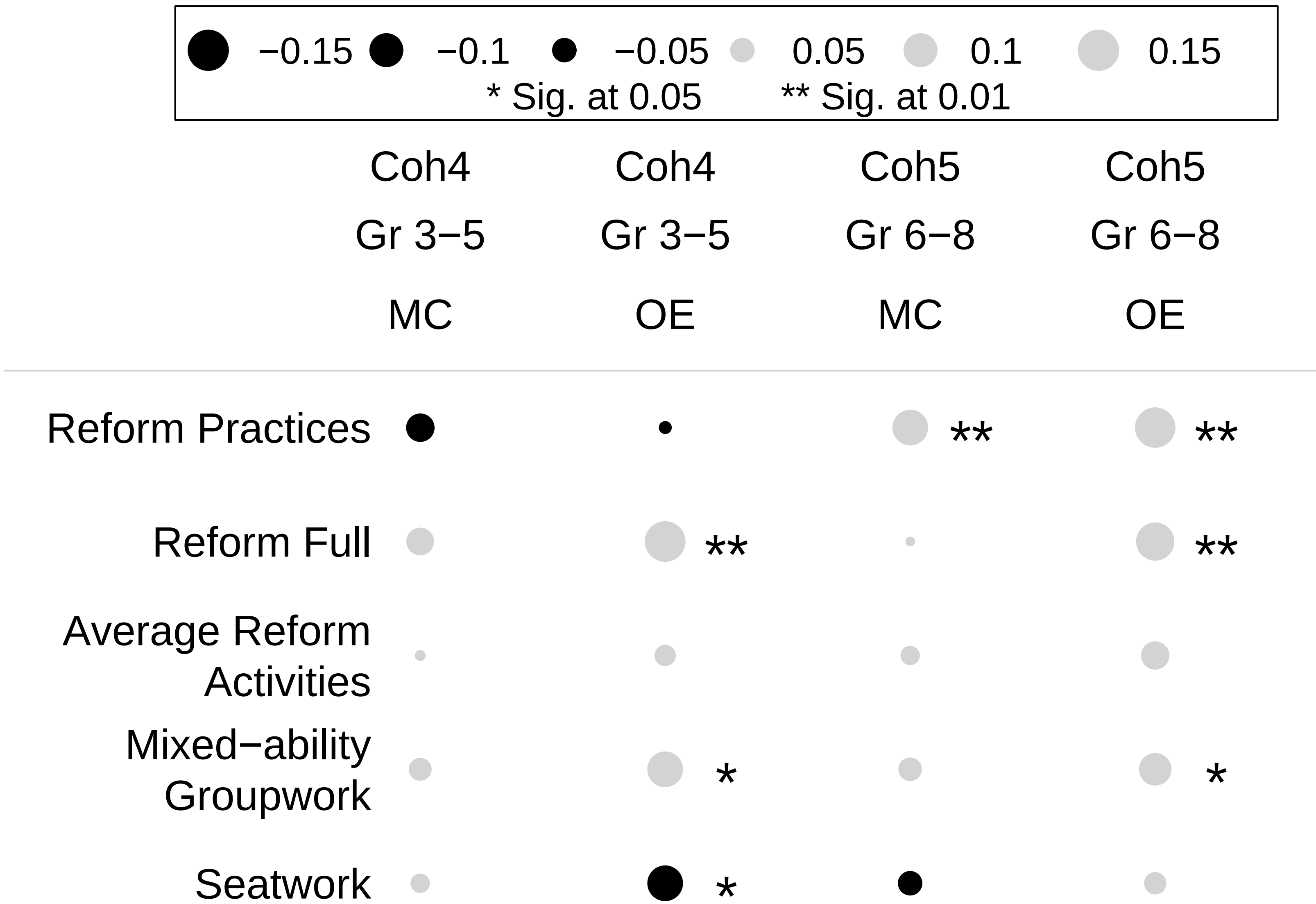
Complete Model Poses Estimation Challenges

- Off-the-shelf mixed-effects models routines generally not equipped to estimate the complex multiple-membership structure of the random effects with unknown persistence of past teacher effects
- Built specialized software capitalizing on Bayesian methods (Markov Chain Monte Carlo) for model estimation, which scales to very large datasets
- However, models with the size of datasets used here probably estimable in WinBugs/OpenBugs (free software for fitting Bayesian models) without much difficulty

Cumulative Effects: Mathematics



Cumulative Effects: Science



RAND

Conclusions I

- Increased testing, better data systems, and heightened demand for using data to better the education system are providing growing opportunities for longitudinal analyses
- However . . .
- (Variance partitioning)
  - Richness not a panacea for “needle in a haystack”
  - Student background remains strong predictor of achievement
- (Missing data)
  - Fractured records mount quickly in long and wide multilevel data series
  - Hard to recover power of (hypothetical) full data, but thoughtful multiple imputation can go a long way

Conclusions II

- (Cumulative effects)
  - Additive effects are simple and interpretable
  - But do we really believe additivity in such a complex system?
- (Nuisance correlations)
  - Accumulation of unmeasured inputs across changing contexts imparts messy structure to residuals
  - Even a first-order approximation can move the analysis into unfriendly computational territory