Reducing Bias in Observational Analyses of Education Data by Accounting for Test Measurement Error

J.R. Lockwood
SREE
March 8, 2012

This material is based on work supported by the Department of Education Institute of Education Sciences under Grant No. R305D090011. 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 these organizations. This talk has not been formally reviewed and should not be cited, quoted, reproduced, or retransmitted without RAND’s permission.
Quantitative Analysis of Student Test Scores Is At An All-Time High

- Explosion of standardized testing and accountability pressure at all levels of public education system

- Use scores from standardized assessments to study: policies, interventions, curricula, teaching practices, individual teachers (“value-added” modeling), etc.

- Even with funding environment moving toward experimental evaluations, observational data analysis using student test scores is still the norm

- Test scores are typically both the main outcome variables, and main covariates used for adjustment
Big Points Up Front

- In most observational studies of education data, “treatment” and “control” groups will differ on attributes related to achievement, and past achievement is typically the best control.

- Adjusting for past achievement is tricky due to:
  - Ambiguous structural relationships among test scores
  - Large, heteroskedastic test score measurement error whose variance depends on unobserved achievement

- Develop a general latent regression modeling approach that addresses these challenges, and demonstrate its value for bias reduction through a case study of estimating teacher effects.
How are Test Scores Constructed?

1. Construct items designed to measure an “achievement” construct \( \theta \) relative to state/grade/subject content
   - IRT model: \( \Pr(\text{item i correct } | \theta) \equiv P_i(\theta) = c_i + \frac{(1-c_i)}{1+e^{-1.7a_i(\theta-b_i)}} \)
   - Estimate “item parameters”: discrimination \((a_i)\), difficulty \((b_i)\), guessing \((c_i)\)

2. Construct a test of 30-60 multiple choice items

3. Estimate an individual student’s value of \( \theta \) by likelihood methods applied to sequence of correct/incorrect answers

4. Apply special sauce to get scale scores
   - Often multiple stages of transformation applied to \( \hat{\theta} \) to ease interpretation, put scores from different grades or subjects on a common numerical scale, and make scale scores a one-to-one function of raw scores
Implications for Modeling Test Scores [I]

- Test scores are not like height or stock values
- Even when designed to be on a “vertical scale” - which is no longer true for most state assessments - test scores across grades probably do not measure the same construct
  - Multi-dimensionality
  - Shifting content to match grade-specific standards
- Implication: sequences of test scores within student across grades and subjects have ambiguous structural relationships
  - Depend on test developer, scoring and scaling methods, content development, alignment with curricula, etc.
Implications for Modeling Test Scores [II]

- There is little theoretical justification for our usual hammers for correcting for student heterogeneity in observational analyses, particularly without a vertical scale:
  - Gain scores
  - Fixed effects

- More appropriate to think of scores as just correlated attributes
  - Prior achievement scores are proxy variables for unobservables related to current achievement scores
  - Motivates “kitchen-sink” type regressions
Measurement Error Further Complicates Matters

- Under IRT model, a student’s raw score (number correct) is the sum of conditionally independent Bernoulli($P_i(\theta)$) variables:

$$\text{Var(raw score)} = \sum_i P_i(\theta)(1 - P_i(\theta))$$

- Basic concept leads to standard error of measure ($SEM$) of individual student’s score

- Depends on IRT model, number and attributes of items, and $\theta$

- Dependence on $\theta$ leads to:
  1. Heteroskedasticity
  2. Uncertainty about each student’s true $SEM$ because $\theta$ is unknown
Framework for Using Score and SEM Data

- For a grade/subject/year: \((\hat{\theta}, SEM(\hat{\theta}))\) pair for each student
  - \(\hat{\theta}\) is approximate MLE of \(\theta\) given item responses ⇒
  \[
  \begin{align*}
  \hat{\theta} | \theta & \approx N (\theta, SEM^2(\theta))
  \end{align*}
  \]
  - Note error variance is not \(SEM^2(\hat{\theta})\)

- Challenging class of models because variance depends on mean: no standard software, ugly likelihood function

- A relatively minor challenge is that we do not have \(SEM(\theta)\) but rather its evaluation at discrete points
  - Use polynomial approximations: 4th or 5th degree sufficient
Measurement Error Hinders Covariate Adjustment

- Back to big picture:
  - Want to use past test scores to control for differences among students in observational studies
  - Want to use kitchen-sink regressions
  - Want to account for test score measurement error

- Focus on regression adjustment but similar issues apply to other observational approaches (e.g. propensity scoring) which have their own challenges that we are working on
  - E.g. we have developed theorems and estimation methods for propensity score weighting with error-prone covariates
General Model For Treatment Effects on Achievement

\[ Y = \alpha + T' \beta + \omega + \epsilon \]

- **Y**: target test score (outcome variable)
- **T**: treatment indicators with effects \( \beta \)
  - Teachers, interventions, policies, etc.
- **\( \omega \)**: unobserved component of target test score that IS NOT orthogonal to \( T \)
- **\( \epsilon \)**: unobserved component of target test score that IS orthogonal to \( T \) (and \( \omega \))
General Model For Treatment Effects on Achievement

\[ Y = \alpha + T' \beta + \omega + \epsilon \]

- Suppose

\[ \omega = A' \lambda + Z' \gamma + \xi \]

where \( A \) are error-free past test scores and \( Z \) are observed student characteristics other than test scores.

- Could relax linearity in \( A \) and/or allow for interactions of \( A \) and \( Z \) but have not tried this.

- Observed test scores \( X \) are error-prone measures of \( A \). Assume \( X = A + e \) where \( A \) and \( e \) are uncorrelated and \( e \sim N(0, \text{diag}(SEM^2(A))) \)
Putting Pieces Together

\[ Y = \alpha + T' \beta + \omega + \epsilon \]

\[ \omega = A' \lambda + Z' \gamma + \xi \]

\[ X = A + e \]

- Observe \((Y, T, Z, X)\) and know distribution \(f(e \mid A)\)
- Hope is that \(\text{Var}(\xi)\) is small - i.e. effective proxy
- Intuition: If tests were infinitely long \((SEM \equiv 0)\) this reduces to standard ANCOVA of \(Y\) on \((T, Z, A)\)
  - Use \(X\) and \(SEM\) information to estimate parameters of this model - “latent regression”
Implementation

❑ Three pieces of model:

1. Outcomes model for $Y$

2. Latent variable model capturing relationships of components of $A$ to one another and to other covariates

3. Measurement model specifying how $X$ relates to $A$, including allowing $SEM$ to depend on $A$

❑ If true $SEM$ of each score were known, could be specified with \texttt{gllamm} in Stata or other latent modeling packages

❑ To allow $SEM$ to be a function of latent scores, implemented model in Bayesian framework using WinBUGS, OpenBUGS and JAGS
Application to Teacher Effect Estimation

- Math achievement in single grade 6 cohort from one district - about 4000 students and 110 math teachers
- Typical set of student demographic and program participation variables (race, gender, FRL, ELL, special education, gifted)
- Up to 12 prior achievement scores available for adjustment (3 years, 4 subjects) along with associated $SEM$ for each individual score
- Compare standard ANCOVA (regression on $(T, Z, X)$) to latent regression (regression on $(T, Z, A)$) in sequence of models including increasing numbers of past test scores
- Examine estimated teacher effects for evidence of bias due to omitted student variables
Controlling for More Scores Helps

![Graph showing variance in VA across teachers with controls for prior scores. The graph compares ANCOVA and LR methods. The x-axis represents controls for prior scores (M1, +R1, +L1, +S1, +M2, +M3, KS, M123), and the y-axis represents variance in VA across teachers. The line indicates a decreasing trend in variance as more scores are controlled for.]
LR Has Less Spurious Teacher Variance

Variance in VA Across Teachers

Controls for Prior Scores

M1  +R1  +L1  +S1  +M2  +M3  KS  M123

ANCOVA
LR
LR Has Weaker Correlation With Average Prior Achievement

Corr(VA, Prior Year M1 Mean)

Controls for Prior Scores
LR Has Weaker Correlation With %FRL

Controls for Prior Scores

Corr(VA, Prior Year %FRL)
LR Is Modestly More Efficient Too

controls for prior scores

average SE of teacher effects

M1 +R1 +L1 +S1 +M2 +M3 KS M123

ANCOVA
LR
Conclusions

- Standardized test scores are messy
  - Ambiguous structural relationships
  - Large, heteroskedastic measurement error whose variance depends on unobserved achievement

- Accounting for measurement error when using test scores for covariate adjustment in observational studies can reduce bias
  - Developed improved kitchen-sink regression that recovers parameters of model we could fit if tests were infinitely long
  - Dominates traditional ANCOVA on both bias and variance

- Research community should be requesting SEM information from administrative data so we can improve what we do
Future Research

- How does latent regression compare to other approaches to measurement error correction (IV, regression calibration, moment corrections to cross-product matrix) in terms of bias reduction and efficiency in the settings we typically encounter?

- A lot of problems with using test scores as covariates go away if we can use item-level data directly in modeling (e.g. MIMIC modeling) - what are the tradeoffs?

- Build contextual/aggregate variables into latent regression and revisit analyses of peer effects, which can be badly biased if test measurement error is ignored

- Consider methods for controlling for other sources of test measurement error (e.g. “bad day”) not accounted by SEM (Lankford et al.)
SUPPLEMENTAL MATERIAL
Example of SEM Function Approximation

![Graph showing a curve for SEM function approximation with score values ranging from -3 to 3 on the x-axis and SEM values ranging from 0.5 to 2.0 on the y-axis. The graph includes a blue curve and black dots representing data points.]
Differences Correlated With Students

Corr(LR – ANCOVA, M1 Mean)

Controls for Prior Scores

M1 +R1 +L1 +S1 +M2 +M3 KS M123
LR Model for Teacher Effects

\[ Y_i \sim \text{ind } \mathcal{N} \left( \gamma^{(1 \times k)} Z_i^{(k \times 1)} + \lambda^{(1 \times p)} U_i^{(p \times 1)} + \theta_j(i), \nu^2 + \nu(E[Y_i]) \right) \quad \forall i \]

\[ \bar{U}_j^{(p \times 1)} \sim \text{iid } \mathcal{N} \left( 0^{(p \times 1)}, T^{(p \times p)} \right) \text{ for each teacher } j \]

\[ U_i^{(p \times 1)} \sim \text{ind } \mathcal{N} \left( \Gamma^{(p \times k)} Z_i^{(k \times 1)} + \bar{U}_{j(i)}^{(p \times 1)}, \Sigma^{(p \times p)} \right) \text{ for each student } i \]

\[ X_i^{(p \times 1)} \sim \text{ind } \mathcal{N} \left( U_i^{(p \times 1)}, \text{diag}(\nu(U_i))^{(p \times p)} \right) \text{ for each student } i \]