

Measuring Effect Sizes: the Effect of Measurement Error

Don Boyd, Pam Grossman, Hamp Lankford,
Susanna Loeb, and Jim Wyckoff

CALDER

November 21, 2008

www.teacherpolicyresearch.org

Estimated Effect Sizes for Teacher Attributes Math Grades 4 & 5, NYC 2000-2005

	Effect Sizes:			
	Estimated effects relative to			
	S.D. of test scores			
First year of experience	0.065**			
Not certified	-0.042**			
Attended competitive college	0.014*			
One S.D. increase in math SAT score	0.041**			

** 1% statistical significance * 5% statistical significance.

(from Boyd, Lankford, Loeb, Rockoff and Wyckoff, "The Narrowing Gap in New York City Teacher Qualifications and its Implications for Student Achievement in High Poverty Schools," JPAM 2008.)

How should effect sizes be measured?

- Measure effects relative to the S.D. of gain scores, not the S.D. of scores.
- Account for test measurement error when computing effect sizes.

⇒ Effect sizes are four times as large.

Measuring effect sizes:

S.D. in gain scores or S.D. in scores?

- Depends on the question.
- Achievement tests typically measure cumulative learning.
 - Any short-run intervention likely will have only a modest effect on achievement.
 - Thus, effect sizes based on achievement levels are likely to be modest.
- Gain scores reflect students' learning during the year in which the intervention occurred.
 - Effect sizes based on gains then provide a better perspective for many teacher-based interventions.

Estimated Effect Sizes for Teacher Attributes Math Grades 4 & 5, NYC 2000-2005

	Effect Sizes:			
	Estimated effects relative to			
	S.D. of test scores	S.D. of test score gains		
First year of experience	0.065**	0.103		
Not certified	-0.042**	-0.067		
Attended competitive college	0.014*	0.022		
One S.D. increase in math SAT score	0.041**	0.065		

** 1% statistical significance * 5% statistical significance.

(from Boyd, Lankford, Loeb, Rockoff and Wyckoff, "The Narrowing Gap in New York City Teacher Qualifications and its Implications for Student Achievement in High Poverty Schools," JPAM 2008.)



How should effect sizes be measured?

- Measure effects relative to the S.D. of gain scores, not the S.D. of scores.
- **Account for test measurement error when computing effect sizes.**

Notation:

Test score: $S_{i,g}$

Universe score: $\tau_{i,g}$

Measurement error: $\eta_{i,g}$

$$S_{i,g} = \tau_{i,g} + \eta_{i,g} \quad \Delta S_{i,g} = \Delta \tau_{i,g} + \eta_{i,g} - \eta_{i,g-1}$$

$$V(\eta_{i,g}) = \sigma_{\eta_i}^2$$

$$\sigma_{S_g}^2 = \sigma_{\tau_g}^2 + \sigma_{\eta}^2$$

$$\sigma_{\Delta S}^2 = \sigma_{\Delta \tau}^2 + 2\sigma_{\eta}^2$$

$$\sigma_{\tau_g}^2 = \sigma_{S_g}^2 - \sigma_{\eta}^2$$

$$\sigma_{\Delta \tau}^2 = \sigma_{\Delta S}^2 - 2\sigma_{\eta}^2$$

Upper bound estimate of universe-score-gain S.D.

$$\sigma_{\Delta\tau}^2 = \sigma_{\Delta S}^2 - 2\sigma_{\eta\bullet}^2$$

$$\sigma_S^2 \approx 1, \quad \hat{\sigma}_{\Delta S}^2 = 0.398$$

Information from technical reports:

$$\sigma_{\eta\bullet}^2 \approx 0.10 \quad \Rightarrow \quad \hat{\sigma}_{\Delta\tau} < 0.439$$

Effect sizes measured relative to $\sigma_{\Delta\tau}$ are more than twice as large as those based on $\sigma_S = 1$.

Reported reliability coefficients ...

- ... frequently present a biased picture
- ... tend to overstate the trustworthiness of educational measurement
- ... standard errors understate within-person variability [resulting from the]
- ... random variation within each individual in health, motivation, mental efficiency, concentration, forgetfulness, carelessness, ...

L.S. Feldt & R.L. Brennan, “Reliability” chapter in *Educational Measurement*, 3rd edition

Estimating test measurement error

We estimate test measurement error from all sources employing the observed student-level covariance structure of test scores in New York City from 1999 to 2007.

Auto-covariance matrix

$$\Omega(i) = E(S_i S_i') = E(\tau_i \tau_i') + E(\eta_i \eta_i')$$

Auto-Covariance Matrix of Test Scores

	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Grade 4	1.004	0.7975	0.7675	0.7574	0.7189
Grade 5		0.9933	0.7813	0.7639	0.7218
Grade 6			0.9899	0.7958	0.7579
Grade 7				0.9820	0.7884
Grade 8					0.9826

Auto-Covariance Matrix of Test Scores

	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Grade 4	1.004	0.7975	0.7675	0.7574	0.7189
Grade 5		0.9933	0.7813	0.7639	0.7218
Grade 6			0.9899	0.7958	0.7579
Grade 7				0.9820	0.7884
Grade 8					0.9826

Stationarity: $\omega^0 = V(S_{i,g}) = \sigma_{S_g}^2 \omega^s \equiv Cov(S_{i,g}, S_{i,g+s})$

Auto-Covariance Estimates Assuming Stationarity

parameters	estimates	S.D.
$\hat{\omega}^0$	0.9924	0.0022
$\hat{\omega}^1$	0.7907	0.0018
$\hat{\omega}^2$	0.7631	0.0018
$\hat{\omega}^3$	0.7396	0.0018
$\hat{\omega}^4$	0.7189	0.0017

A Structural Model of Test-Score Auto-Covariance

$$\left. \begin{aligned} S_{i,g} &= \tau_{i,g} + \eta_{i,g} \\ \tau_{i,g} &= \beta \tau_{i,g-1} + \theta_{i,g} \\ \theta_{i,g} &= \mu_i + \varepsilon_{i,g} \end{aligned} \right\} \Rightarrow \begin{aligned} \omega^0 &= \gamma^0 + \sigma_{\eta}^2 \\ \omega^1 &= \beta \gamma^0 + \lambda \\ \omega^2 &= \beta^2 \gamma^0 + (\beta + 1) \lambda \\ \omega^3 &= \beta^3 \gamma^0 + (\beta^2 + \beta + 1) \lambda \\ \omega^4 &= \beta^4 \gamma^0 + (\beta^3 + \beta^2 + \beta + 1) \lambda \end{aligned}$$

Estimating the Structural Parameters

$$\hat{\omega}^0 = 0.9924$$

$$\hat{\omega}^1 = 0.7907$$

$$\hat{\omega}^2 = 0.7631$$

$$\hat{\omega}^3 = 0.7396$$

$$\hat{\omega}^4 = 0.7189$$

$$Q = \sum_j \left(\hat{\omega}^j - \omega^j(\chi) \right)^2$$

$$\omega^0 = \gamma^0 + \sigma_{\eta_\bullet}^2$$

$$\omega^1 = \beta \gamma^0 + \lambda$$

$$\omega^2 = \beta^2 \gamma^0 + (\beta + 1) \lambda$$

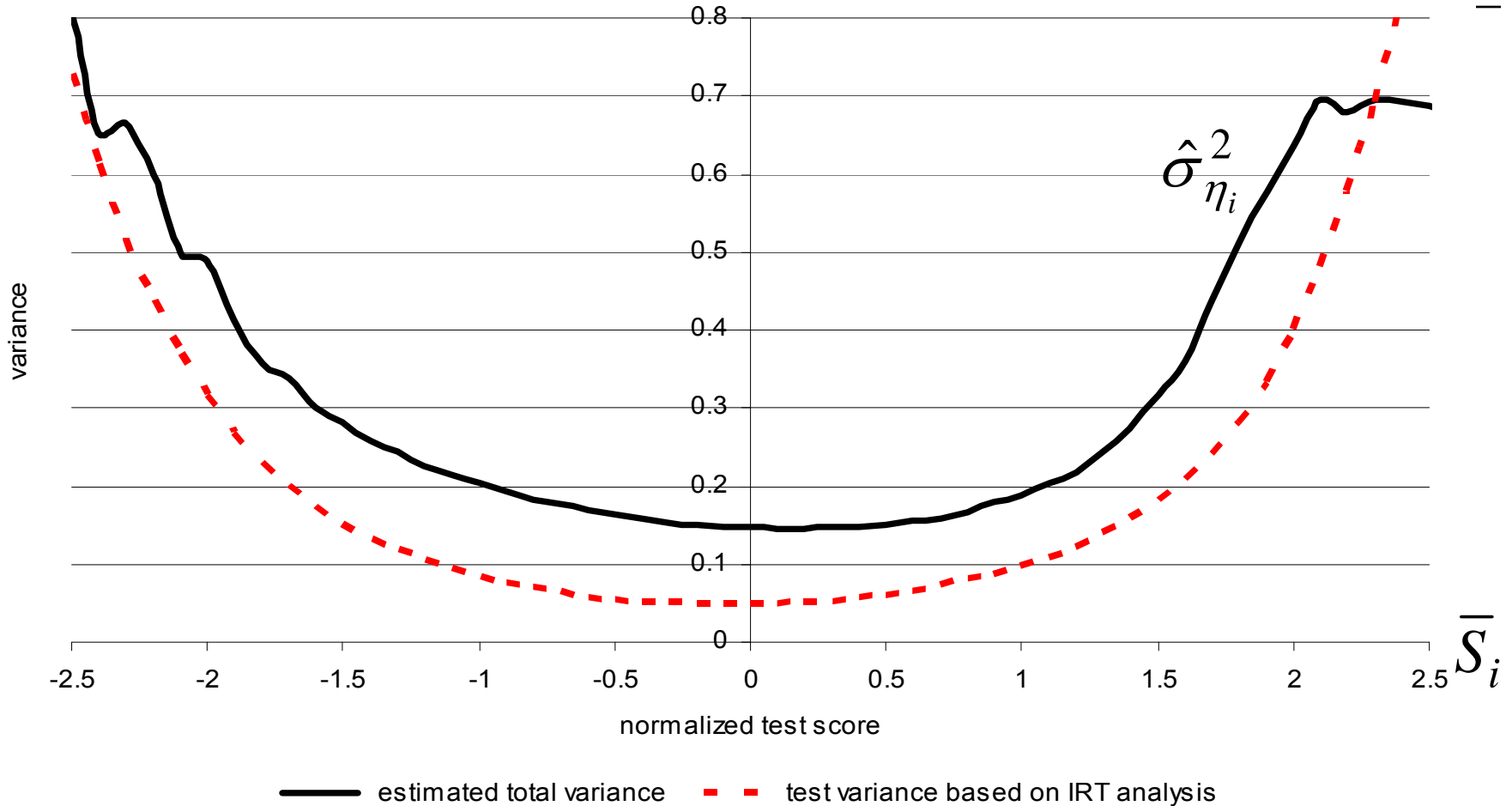
$$\omega^3 = \beta^3 \gamma^0 + (\beta^2 + \beta + 1) \lambda$$

$$\omega^4 = \beta^4 \gamma^0 + (\beta^3 + \beta^2 + \beta + 1) \lambda$$

$$\chi \equiv \left[\sigma_{\eta_\bullet}^2 \quad \gamma^0 \quad \beta \quad \lambda \right]$$

- Key Assumptions:**
1. $\beta > 0$ at least some persistence in achievement that is constant across grades.
 2. Measurement error is uncorrelated across grades

Estimated Total Measurement Error Variance and Average Variance of Measurement Error Associated with Test Construction



Estimates of the Measurement Error Variance and the Reliability of Test Score Gains

Model 1: $\hat{\sigma}_{\eta_{\bullet}}^2 = 0.170 \quad \Rightarrow \quad \frac{\hat{\sigma}_{\Delta\tau}^2}{\hat{\sigma}_{\Delta S}^2} = 0.146$

Model 2: $\hat{\sigma}_{\eta_{\bullet}}^2 = 0.165 \quad \Rightarrow \quad \frac{\hat{\sigma}_{\Delta\tau}^2}{\hat{\sigma}_{\Delta S}^2} = 0.171$

Estimates of the Standard Deviation of the Universe Score Gains

Model 1: $\hat{\sigma}_{\eta_{\bullet}}^2 = 0.170 \quad \Rightarrow \quad \hat{\sigma}_{\Delta\tau} = 0.241$

Model 2: $\hat{\sigma}_{\eta_{\bullet}}^2 = 0.165 \quad \Rightarrow \quad \hat{\sigma}_{\Delta\tau} = 0.261$

Effect sizes measured relative to $\sigma_{\Delta\tau}$ are four times as large as those based on $\sigma_S = 1$.

Estimated Effect Sizes for Teacher Attributes Math Grades 4 & 5, NYC 2000-2005

	Effect Sizes:			
	Estimated effects relative to			
	S.D. of test scores	S.D. of test score gains	S.D. of universe scores	
First year of experience	0.065**	0.103	0.072	
Not certified	-0.042**	-0.067	-0.046	
Attended competitive college	0.014*	0.022	0.016	
One S.D. increase in math SAT score	0.041**	0.065	0.045	

** 1% statistical significance * 5% statistical significance.

(from Boyd, Lankford, Loeb, Rockoff and Wyckoff, "The Narrowing Gap in New York City Teacher Qualifications and its Implications for Student Achievement in High Poverty Schools," JPAM 2008.)

Estimated Effect Sizes for Teacher Attributes Math Grades 4 & 5, NYC 2000-2005

	Effect Sizes:			
	Estimated effects relative to			
	S.D. of test scores	S.D. of test score gains	S.D. of universe scores	S.D. of universe score gains
First year of experience	0.065**	0.103	0.072	0.253
Not certified	-0.042**	-0.067	-0.046	0.162
Attended competitive college	0.014*	0.022	0.016	0.054
One S.D. increase in math SAT score	0.041**	0.065	0.045	0.158

** 1% statistical significance * 5% statistical significance.

(from Boyd, Lankford, Loeb, Rockoff and Wyckoff, "The Narrowing Gap in New York City Teacher Qualifications and its Implications for Student Achievement in High Poverty Schools," JPAM 2008.)

Conclusion

- It is important to account for the test measurement error from all sources when measuring effect sizes and the dispersion in student achievement more generally.
- The overall extent of test measurement error can be inferred in a relatively straightforward manner.
- Observed teacher attributes are seen to be linked to important gains in student achievement once we account for test measurement error.

This and other papers available at:

www.teacherpolicyresearch.org