

Measuring Effect Sizes, the Effect of Measurement Error

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

www.teacherpolicyresearch.org

National Conference on Value-Added Modeling
April 23, 2008

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

	Effect Sizes: Estimated effects relative to		
	S.D. of observed score		
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," 2007.)

How should effect sizes be measured?

How should effect sizes be measured?

We argue:

- Measure effects relative to the S.D. of gain scores, not the S.D. of scores.

How should effect sizes be measured?

We argue:

- Measure effects relative to the S.D. of gain scores, not the S.D. of scores.

How should effect sizes be measured?

We argue:

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

Notation:

Test score: $S_{i,g}$

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

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

Notation: Test score: $S_{i,g}$
 Universal score: $\tau_{i,g}$
 Measurement error: $\eta_{i,g}$

$$S_{i,g} = \tau_{i,g} + \eta_{i,g}$$

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

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

Notation: Test score: $S_{i,g}$
Universal score: $\tau_{i,g}$
Measurement error: $\eta_{i,g}$

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

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

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

Notation: Test score: $S_{i,g}$
Universal score: $\tau_{i,g}$
Measurement error: $\eta_{i,g}$

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

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

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

Information from technical reports: $\sigma_{\eta}^2 \approx 0.10 \Rightarrow \hat{\sigma}_{\Delta \tau}^2 < 0.439$

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

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

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: $\omega^0 = V(S_{i,g}) = \sigma_{S_g}^2$

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 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

$$S_{i,g} = \tau_{i,g} + \eta_{i,g}$$

A Structural Model of Test-Score Auto-Covariance

$$S_{i,g} = \tau_{i,g} + \eta_{i,g}$$

$$\tau_{i,g} = \beta \tau_{i,g-1} + \theta_{i,g}$$

A Structural Model of Test-Score Auto-Covariance

$$S_{i,g} = \tau_{i,g} + \eta_{i,g}$$

$$\tau_{i,g} = \beta \tau_{i,g-1} + \theta_{i,g}$$

$$\theta_{i,g} = \mu_i + \varepsilon_{i,g}$$

A Structural Model of Test-Score Auto-Covariance

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

Estimating the Structural Parameters

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

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

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

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

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

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

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

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

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

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

Estimating the Structural Parameters

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

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

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

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

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

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

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

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

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

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

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

Estimating the Structural Parameters

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

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

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

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

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

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

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

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

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

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

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

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

Estimating the Structural Parameters

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

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

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

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

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

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

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

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

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

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

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

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

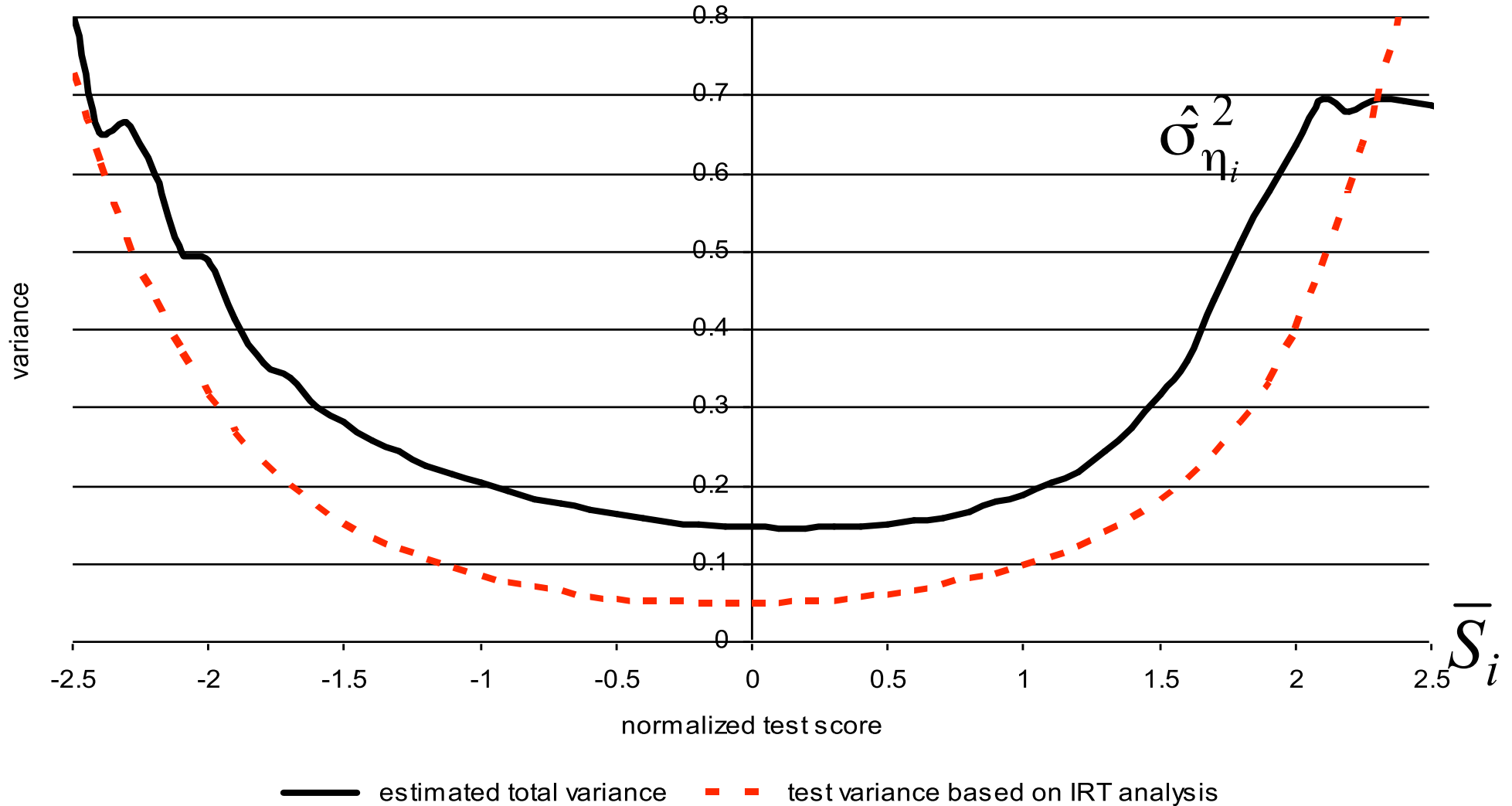
- Key Assumptions:**
1. $\beta > 0$; at least some persistence in achievement.
 2. Measurement error is completely transitory.

Estimates of the Measurement Error Variance and Standard Deviation of the Universal Gain Scores

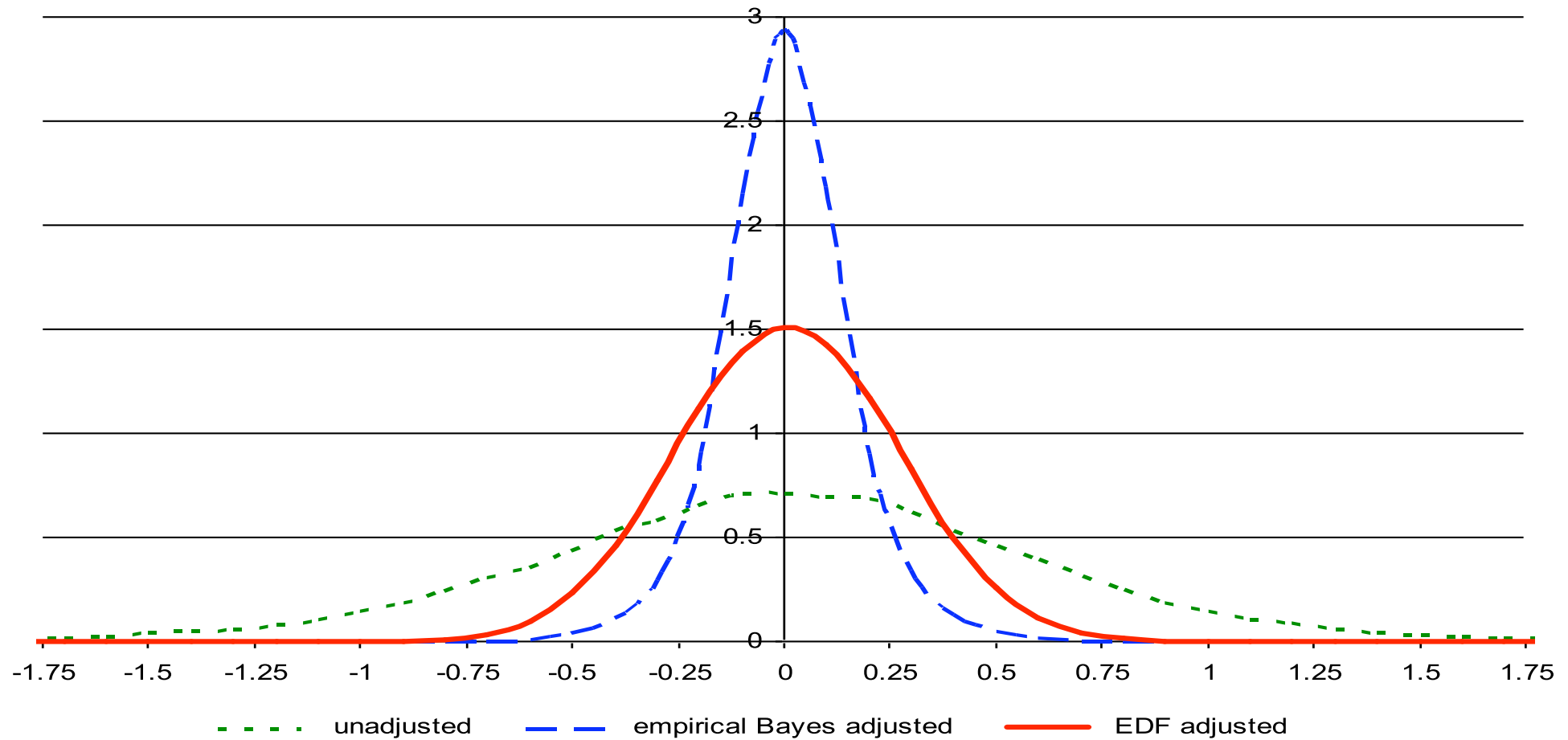
Model 1: $\hat{\sigma}_{\eta\bullet}^2 = 0.170 \quad \Rightarrow \quad \hat{\sigma}_{\Delta\tau} = 0.241$

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

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

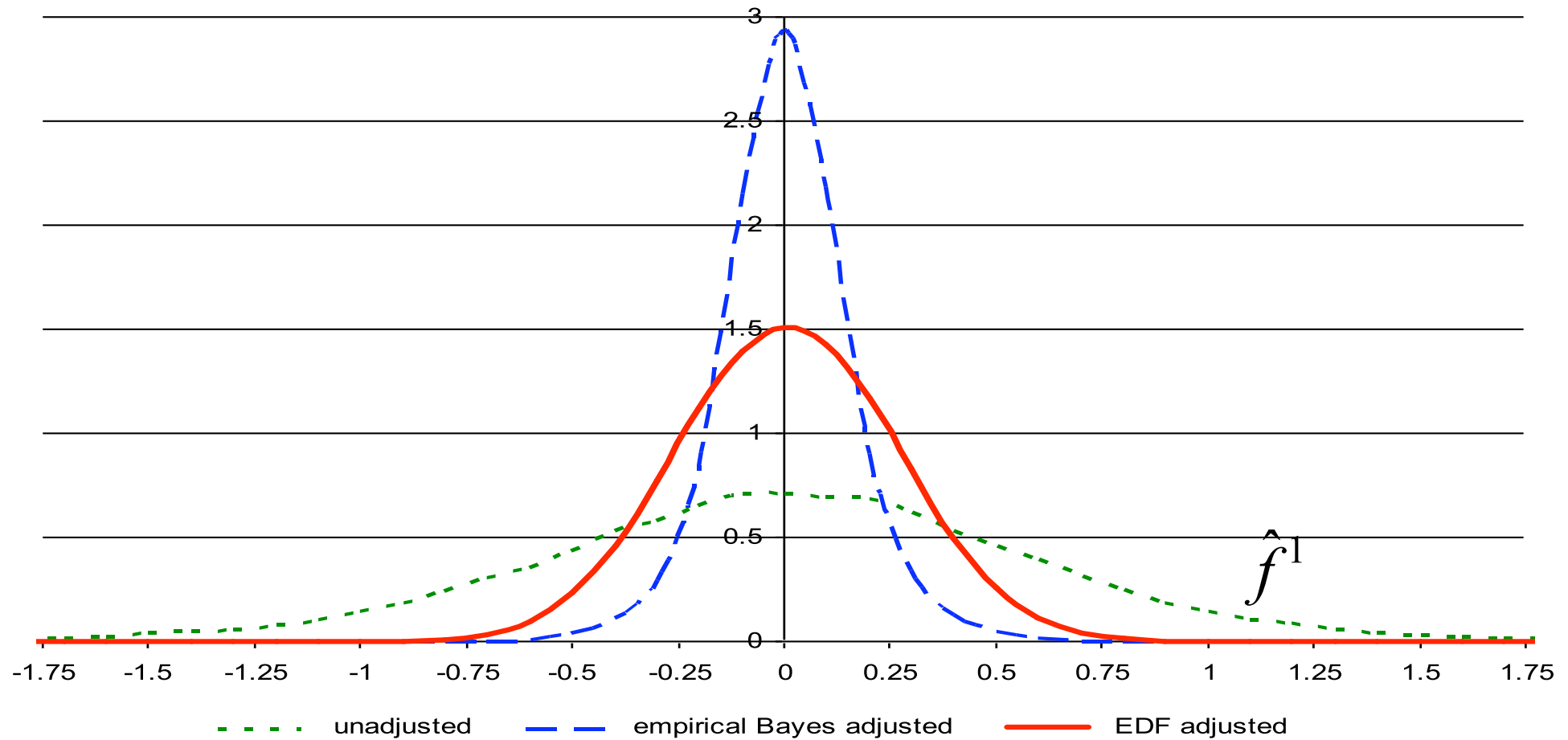


Estimated Empirical Distribution of Universal Gain Scores and Distributions of Gain Scores and Empirical Bayes Estimates



$$F(z) = \sum_i I(\Delta\tau_{i,g} \leq z) / N$$

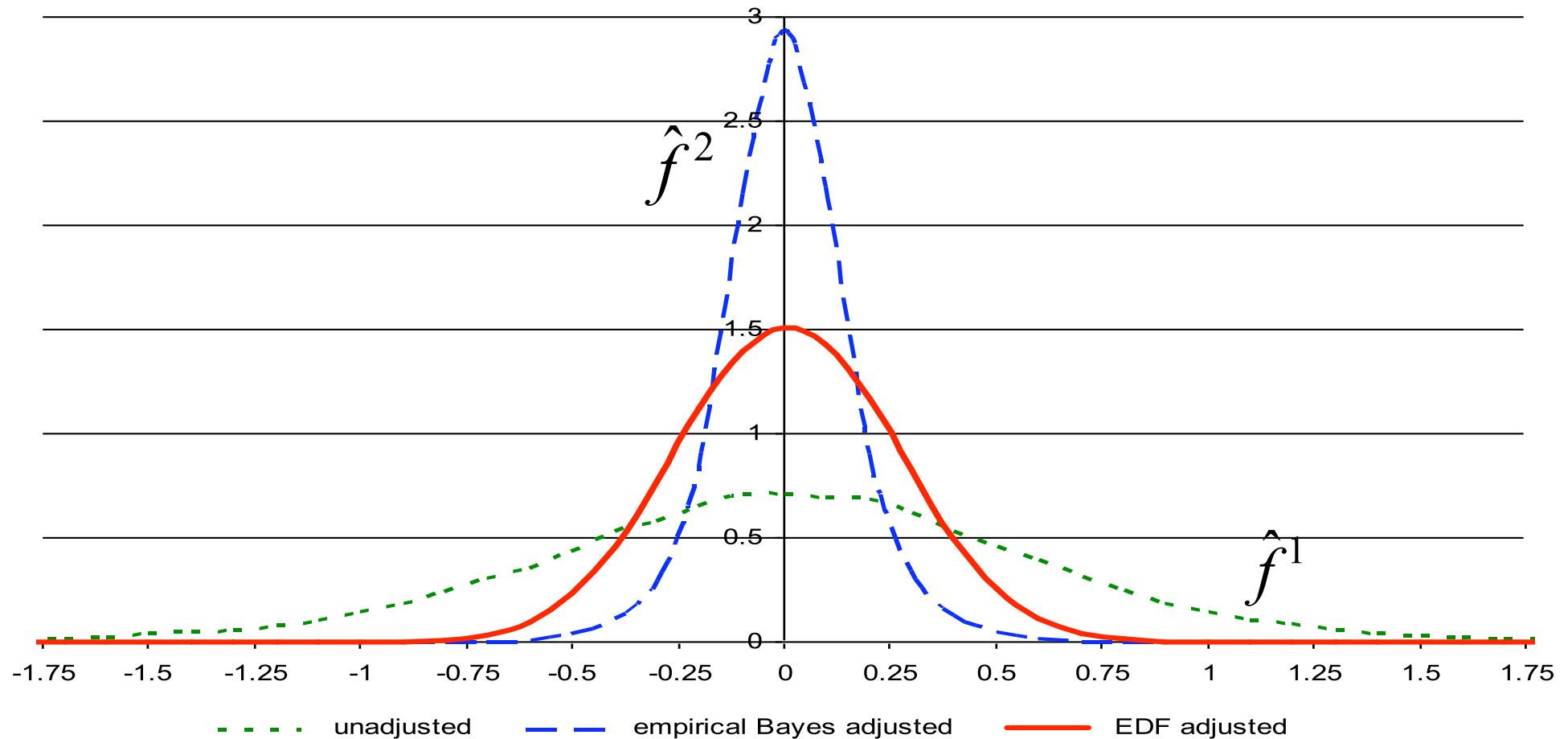
Estimated Empirical Distribution of Universal Gain Scores and Distributions of Gain Scores and Empirical Bayes Estimates



$$F(z) = \sum_i I(\Delta\tau_{i,g} \leq z) / N$$

$$\hat{F}^1(z) = \sum_i I(\Delta S_{i,5} \leq z) / N$$

Estimated Empirical Distribution of Universal Gain Scores and Distributions of Gain Scores and Empirical Bayes Estimates

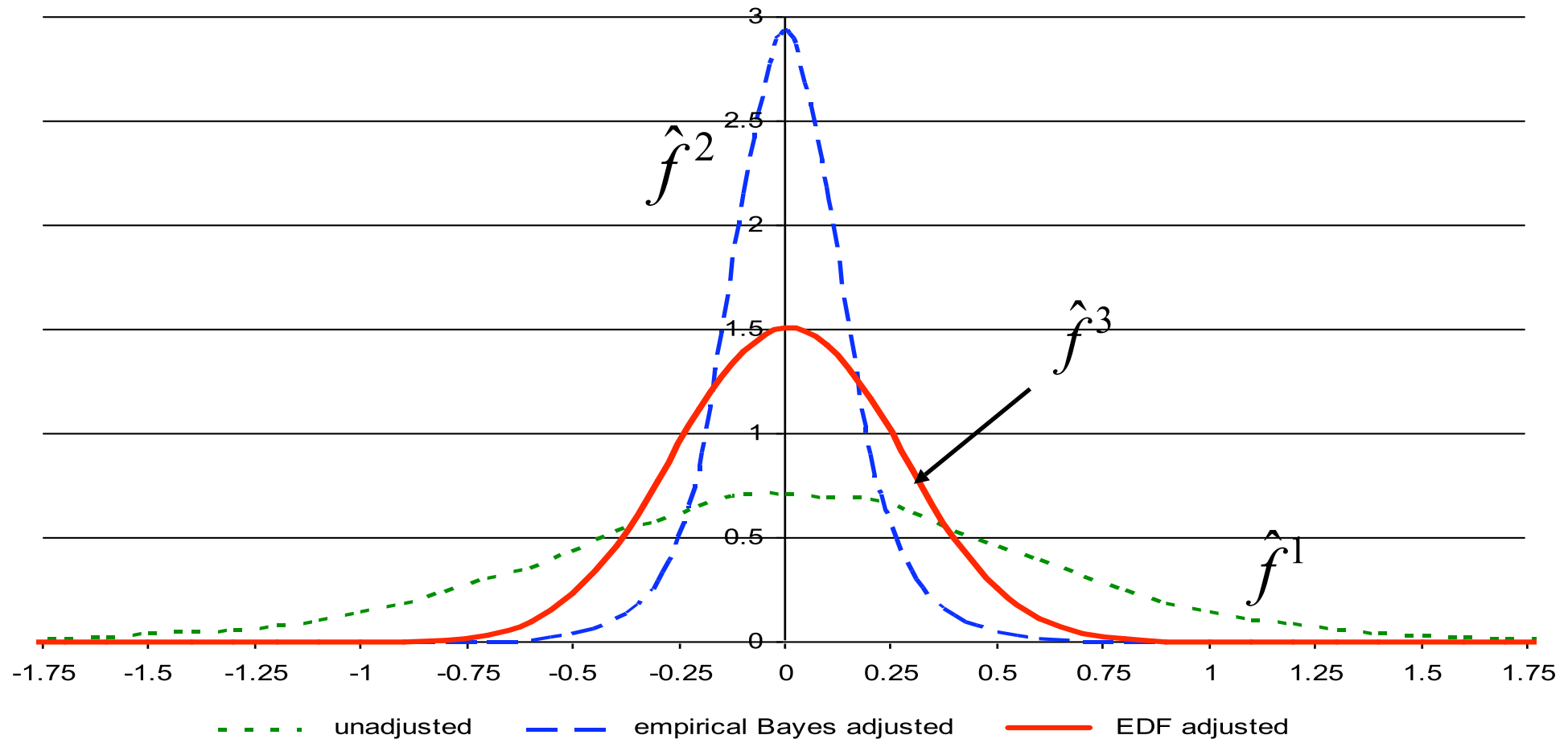


$$F(z) = \sum_i I(\Delta\tau_{i,g} \leq z) / N$$

$$\hat{F}^1(z) = \sum_i I(\Delta S_{i,5} \leq z) / N$$

$$\hat{F}^2(z) = \sum_i I(\Delta S_{i,5}^{EB} \leq z) / N$$

Estimated Empirical Distribution of Universal Gain Scores and Distributions of Gain Scores and Empirical Bayes Estimates



$$F(z) = \sum_i I(\Delta\tau_{i,g} \leq z) / N$$

$$\hat{F}^1(z) = \sum_i I(\Delta S_{i,5} \leq z) / N$$

$$\hat{F}^2(z) = \sum_i I(\Delta S_{i,5}^{EB} \leq z) / N$$

$$\hat{F}^3(z) | S = \sum_i \Phi\left(\frac{z - \Delta S_{i,5}^{EB}}{\hat{\sigma}_{\eta \bullet} \sqrt{\hat{G}_i^\Delta}}\right) / N$$

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

	Effect Sizes: Estimated effects relative to		
	S.D. of observed score		
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," 2007.)

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

	Effect Sizes: Estimated effects relative to		
	S.D. of observed score	S.D. of observed gain score	
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," 2007.)

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

	Effect Sizes: Estimated effects relative to		
	S.D. of observed score	S.D. of observed gain score	S.D. of universal score gain
First year of experience	0.065**	0.103	0.253
Not certified	-0.042**	-0.067	0.162
Attended competitive college	0.014*	0.022	0.054
One S.D. increase in math SAT score	0.041**	0.065	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," 2007.)

Average Qualifications of Teachers in Poorest Quartile of Schools by Math Achievement Quintiles Predicted Solely Based on Teacher Qualifications (excluding experience), 2000-20005

VA Quintile	Mean VA	Not Certified	LAST Pass First	LAST Score	Math SAT	Verbal SAT	College Ranking Competitive or Higher
1	-0.068						
2	-0.032						
3	-0.01						
4	0.01						
5	0.045						
Range	0.113						

Average Qualifications of Teachers in Poorest Quartile of Schools by Math Achievement Quintiles Predicted Solely Based on Teacher Qualifications (excluding experience), 2000-20005

VA Quintile	Mean VA	Not Certified	LAST Pass First	LAST Score	Math SAT	Verbal SAT	College Ranking Competitive or Higher
1	-0.068	0.731	0.46	227	355	440	0.101
2	-0.032	0.141	0.656	239	414	467	0.121
3	-0.01	0.076	0.779	245	423	462	0.224
4	0.01	0.031	0.851	252	450	470	0.352
5	0.045	0.013	0.908	254	512	474	0.494
Range	0.113	-0.718	0.448	27	157	34	0.393

Average Qualifications of Teachers in Poorest Quartile of Schools by Math Achievement Quintiles Predicted Solely Based on Teacher Qualifications (excluding experience), 2000-20005

VA Quintile	Mean VA	Not Certified	LAST Pass First	LAST Score	Math SAT	Verbal SAT	College Ranking Competitive or Higher
1	-0.068	0.731	0.46	227	355	440	0.101
2	-0.032	0.141	0.656	239	414	467	0.121
3	-0.01	0.076	0.779	245	423	462	0.224
4	0.01	0.031	0.851	252	450	470	0.352
5	0.045	0.013	0.908	254	512	474	0.494
Range	0.113	-0.718	0.448	27	157	34	0.393

The 0.11 average difference is 0.43
of a S.D. in universal gain scores.

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.
- Accounting for test measurement error, we see that observed teacher attributes are linked to important gains in student achievement.