

Vertical Scaling in Value-Added Models for Student Learning

Derek Briggs

Jonathan Weeks

Ed Wiley

University of Colorado, Boulder

Presentation at the annual meeting of the National Conference on Value-Added Modeling. April 22-24, 2008. Madison, WI.

Overview

- Value-added models require some form of longitudinal data.
- Implicit assumption that test scores have a consistent interpretation over time.
- There are multiple technical decisions to make when creating a vertical score scale.
- Do these decisions have a sizeable impact on
 - student growth projections?
 - value-added school residuals?

Creating Vertical Scales

1. Linking Design

2. Choice of IRT Model

3. Calibration Approach

4. Estimating Scale Scores

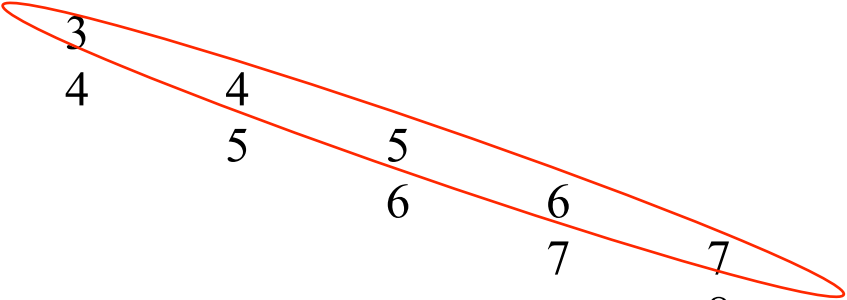
Data

Grade Cohorts	Year				
	2003	2004	2005	2006	2007
Grade 3 Reading	3				
Grade 4 Reading	4	4			
Grade 5 Reading		5	5		
Grade 6 Reading			6	6	
Grade 7 Reading				7	7
Grade 8 Reading					8

- Outcome measure is Colorado Student Assessment Program (CSAP) test scores in reading. [Items: ~70 MC, 14 CR]
- **Longitudinal item responses** for two cohorts of public and charter school students in the state of Colorado.
- Each grade by year cell combination contains roughly 56,000 students. 1,379 unique schools.
- Race/Ethnicity of Students: 64% White, 26% Hispanic, 6.3% Black

Data

Grade Cohorts	Year				
	2003	2004	2005	2006	2007
Grade 3 Reading	3				
Grade 4 Reading	4	4			
Grade 5 Reading		5	5		
Grade 6 Reading			6	6	
Grade 7 Reading				7	7
Grade 8 Reading					8



- Outcome measure is Colorado Student Assessment Program (CSAP) test scores in reading. [Items: ~70 MC, 14 CR]
- **Longitudinal item responses** for two cohorts of public and charter school students in the state of Colorado.
- Each grade by year cell combination contains roughly 56,000 students. 1,379 unique schools.
- Race/Ethnicity of Students: 64% White, 26% Hispanic, 6.3% Black

Data

Grade Cohorts	Year				
	2003	2004	2005	2006	2007
Grade 3 Reading	3				
Grade 4 Reading	4	4			
Grade 5 Reading		5	5		
Grade 6 Reading			6	6	
Grade 7 Reading				7	7
Grade 8 Reading					8

- Outcome measure is Colorado Student Assessment Program (CSAP) test scores in reading. [Items: ~70 MC, 14 CR]
- **Longitudinal item responses** for two cohorts of public and charter school students in the state of Colorado.
- Each grade by year cell combination contains roughly 56,000 students. 1,379 unique schools.
- Race/Ethnicity of Students: 64% White, 26% Hispanic, 6.3% Black

Linking Design

Year	Grade				
	3	4	5	6	7
2003	(34, 7)	(13, 3)	(56, 14)		
		(15, 3)			
2004		(56, 14)	(9, 3)	(56, 14)	
			(20, 2)		
2005			(58, 14)	(11, 4)	(57, 14)
				(15, 0)	
2006				(57, 14)	(10, 4)
					(58, 14)

- (MC items, CR items) Unique Items
- (MC items, CR items) Common Items

Note: No common linking items were available between 2006 and 2007.

Linking Design

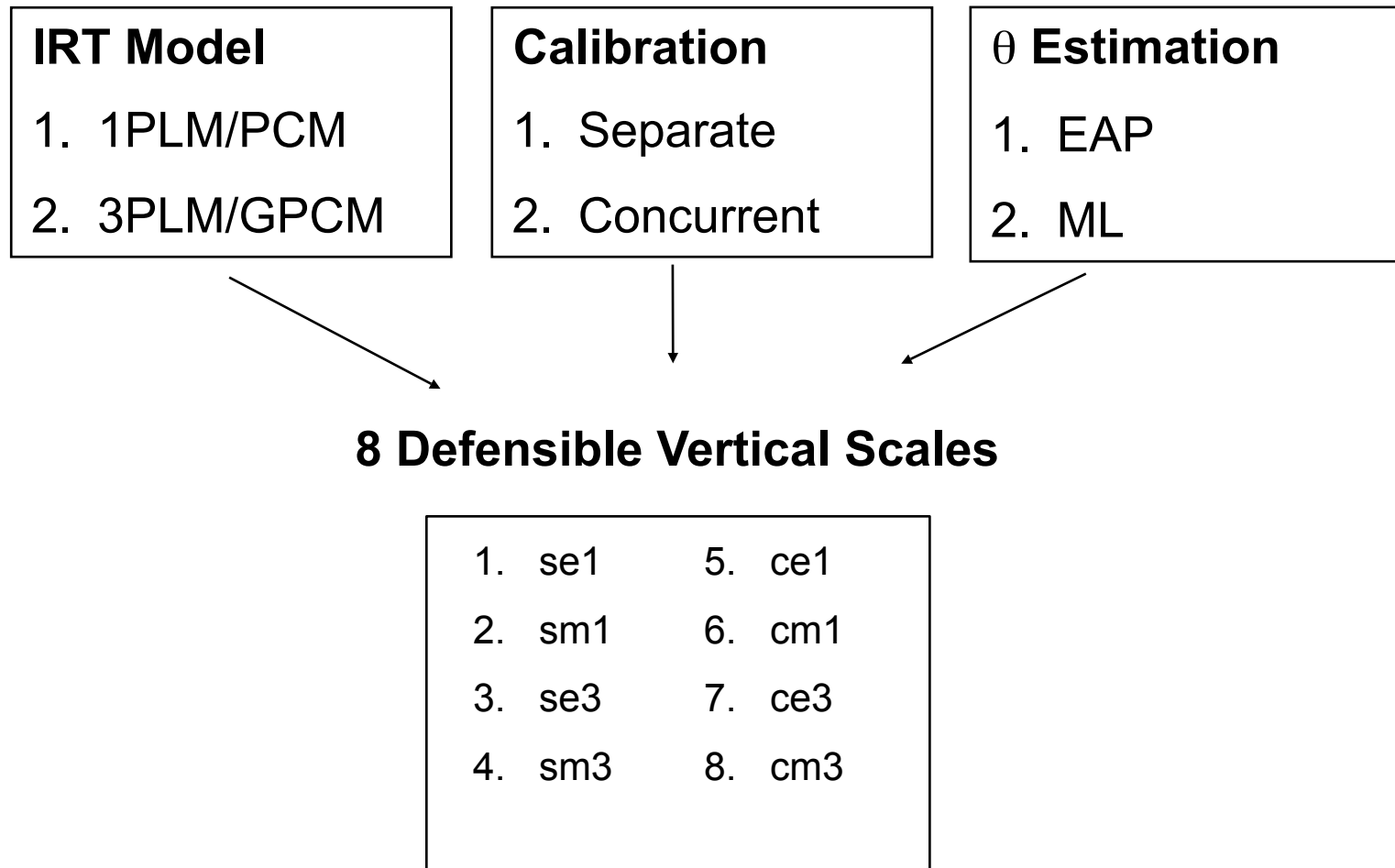
Year	Grade				
	3	4	5	6	7
2003	(34, 7)	(13, 3)	(56, 14)		
		(15, 3)			
2004		(56, 14)	(9, 3)	(56, 14)	
			(20, 2)		
2005			(58, 14)	(11, 4)	(57, 14)
				(15, 0)	
2006				(57, 14)	(10, 4) (58, 14)

- (MC items, CR items) Unique Items
- (MC items, CR items) Common Items**

Note: No common linking items were available between 2006 and 2007.

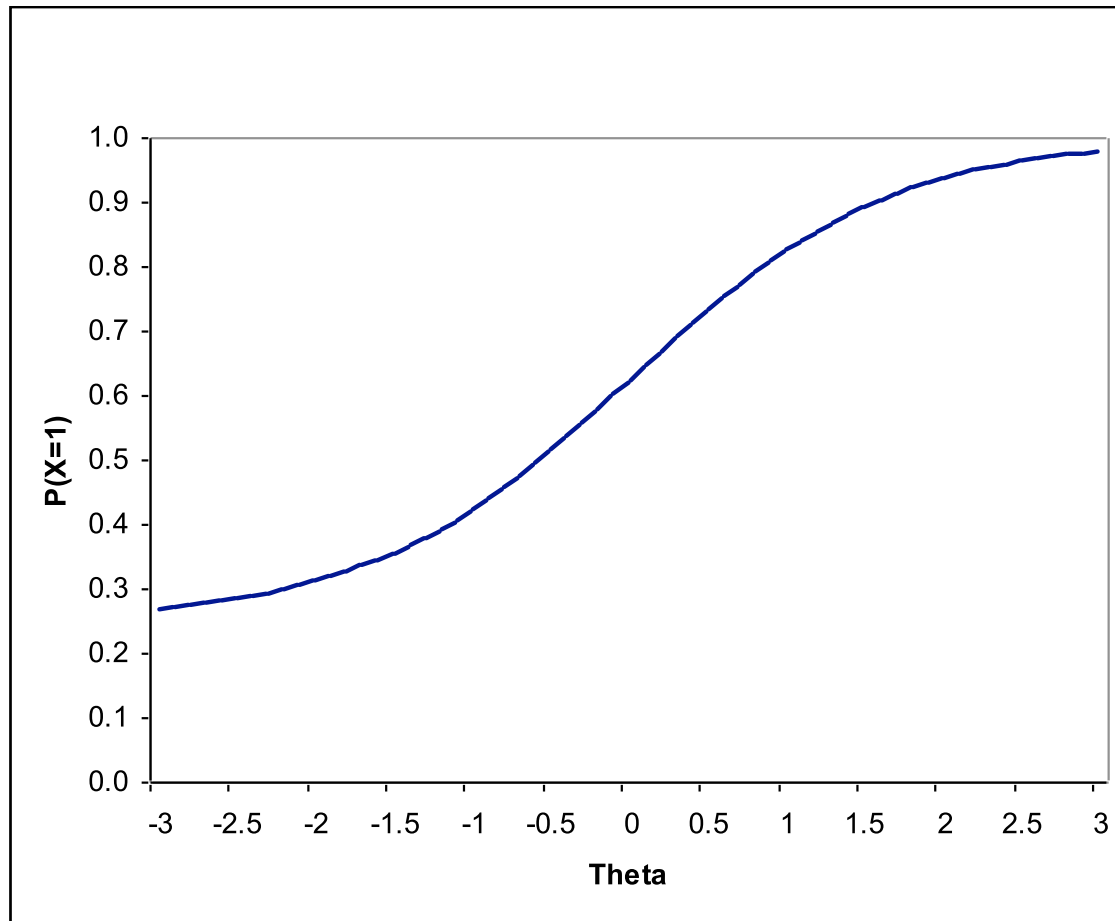
Creating a Vertical Scale

Technical Decisions Psychometricians Make



Item Response Theory Models

$$P(X_{is} = 1 | \theta_s) = \gamma_i + (1 - \gamma_i) \frac{\exp(\alpha_i(\theta_s - \beta_i))}{1 + \exp(\alpha_i(\theta_s - \beta_i))}$$



The eqn above is the 3 Parameter Logistic (3PL) IRT model for binary test items.

The 1PL model results from imposing the constraints

$$\gamma_i = 0$$

$$\alpha_i = 1$$

Reasons for choosing particular IRT model specification: statistical, pragmatic, philosophical.

IRT Assumptions and Properties

Assumptions

- Unidimensionality: The test only measures one latent construct.
- Local Independence: Conditional on this latent construct, item responses are independent.

Properties

- Scale Indeterminacy: The scale of a test is only identified up to a linear transformation.
- Parameter Invariance: If the model fits, item & person parameters should be the same regardless of the group of persons & items used to estimate them.

Separate Calibration

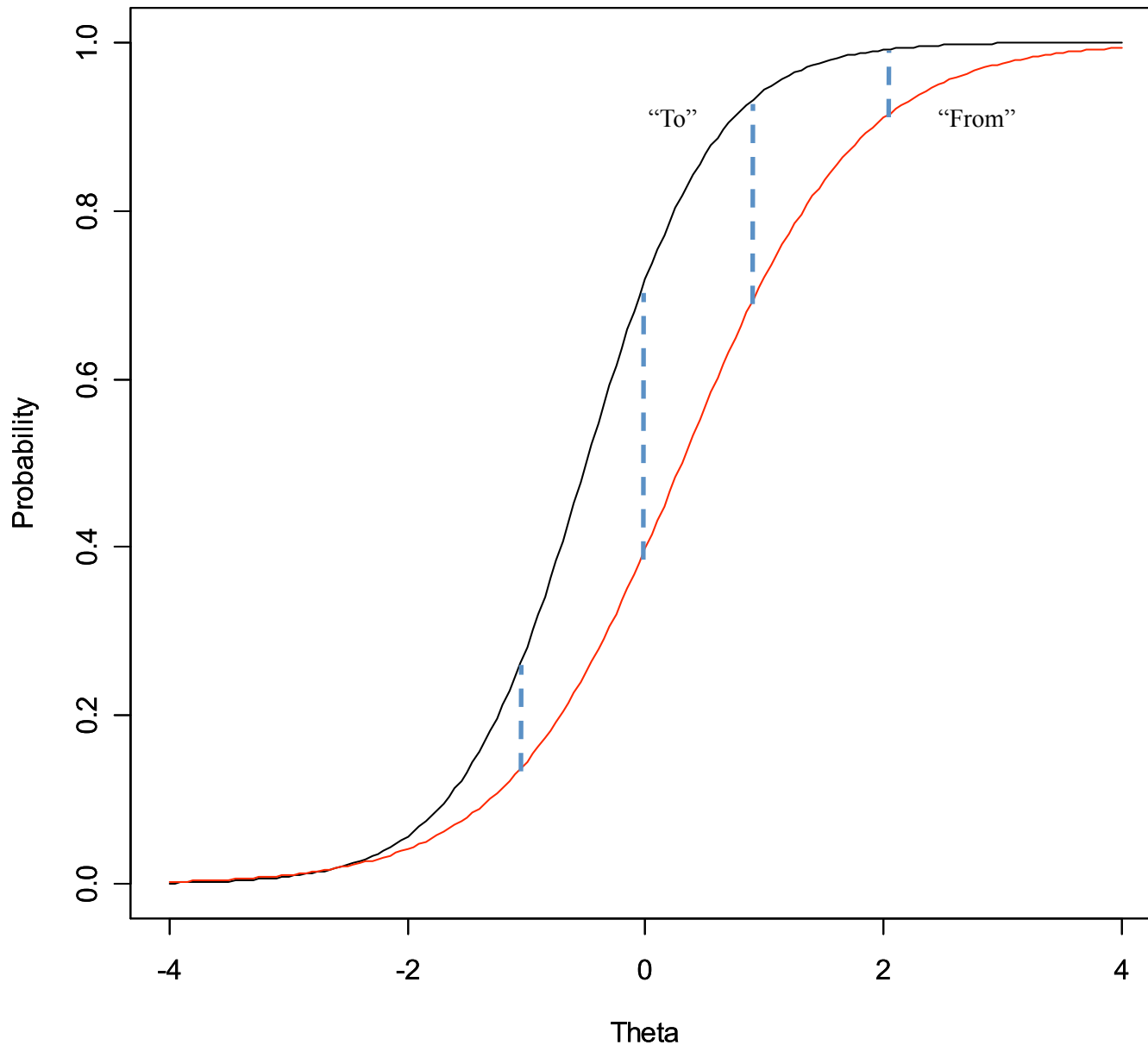
1. Item and person parameters are estimated separately for each grade by year combination.
2. A linear transformation is used to place the parameters from one test—the “From” scale—onto the scale of the other—the “To” scale.

- Ability Estimates $\theta_T = A\theta_F + B$

- Item Parameters $\alpha_T = \frac{\alpha_F}{A}$ $\beta_T = A\beta_F + B$

A and B represent “linking constants”

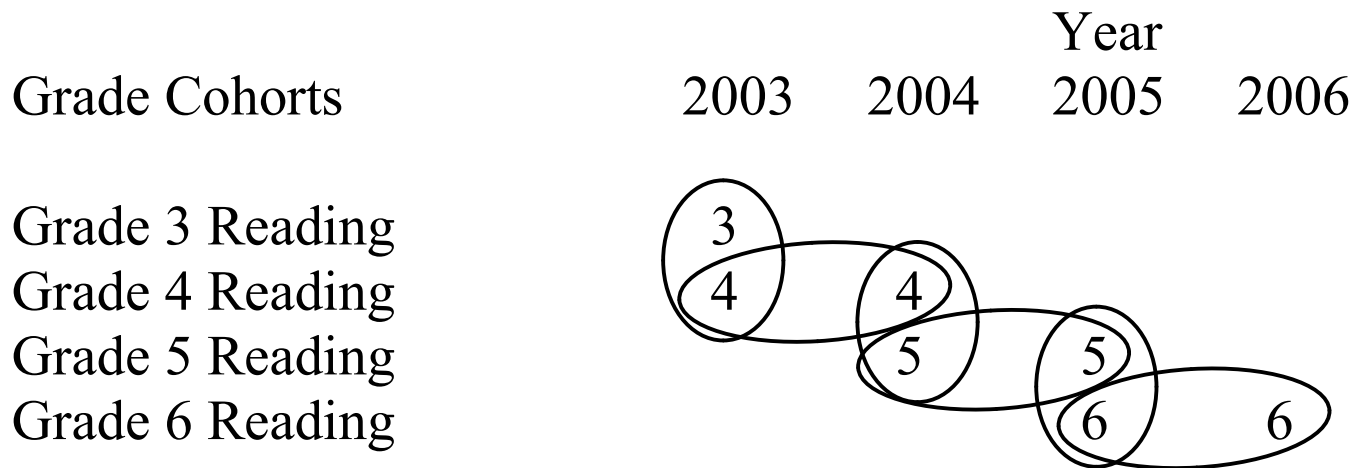
Estimating Linking Constants



Stocking & Lord Approach

1. Compute a test characteristic curve for each test as the sum of item characteristic curves.
2. Sum the squared differences between the curves.
3. Find the linking constants A and B that minimizes the criterion in 2.

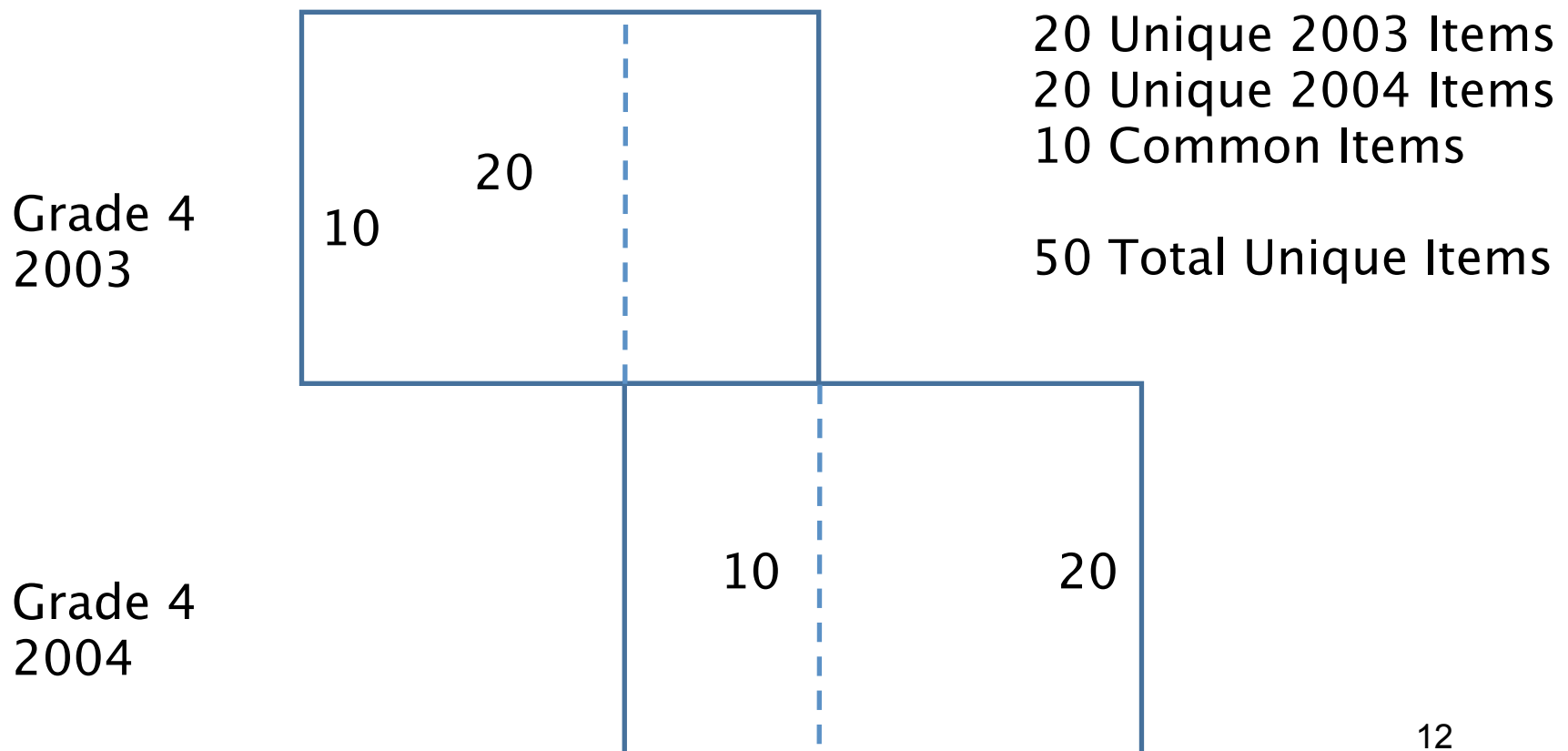
Separate Calibration w/ CO Data



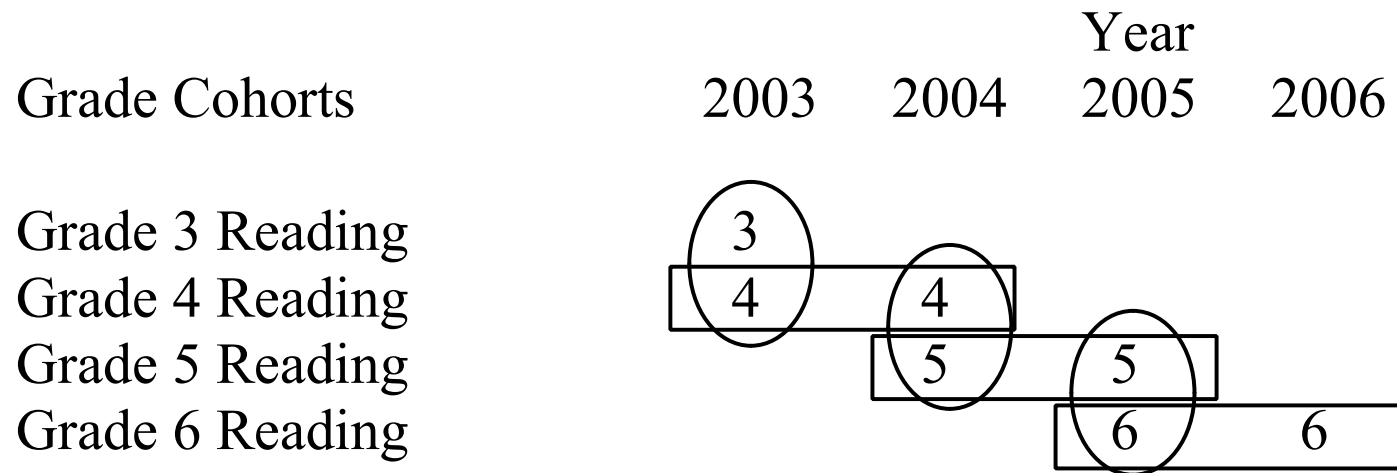
Each oval represents a the linking of two separate item calibrations using the Stocking & Lord approach.

Concurrent Calibration

The item parameters from multiple groups of test-takers are estimated simultaneously



Hybrid Calibration w/ CO Data



- Each oval represents a the linking of two separate item calibrations using the Stocking & Lord ICC approach.
- Each rectangle represents the concurrent, multigroup calibration of the same grade level across two years.

Estimating Student Scale Scores

In IRT, estimation of student-level scale scores happens after item parameters have been estimated. Two key options:

1. Maximum Likelihood estimates (ML)
2. Expected a Posteriori estimates (EAP)

Tradeoffs:

- ML estimates are asymptotically unbiased.
- EAP estimates minimize measurement error.

Value-Added Models

- 1. Parametric Growth (HLM)**
- 2. Non-Parametric Growth
(Layered Model)**

Parametric Growth Model

- Linear Mixed Effects Model (3 Level HLM)
- Given 3 years of test score data for a student (grades 3-5), project a scale score 3 years later (grade 8) [Model proposed by OR, HI]
- Score projection is a function of
 - two fixed effects (intercept & slope)
 - two student level random effects (level 2 intercept & slope)
 - two school level random effects (level 3 intercept & slope)

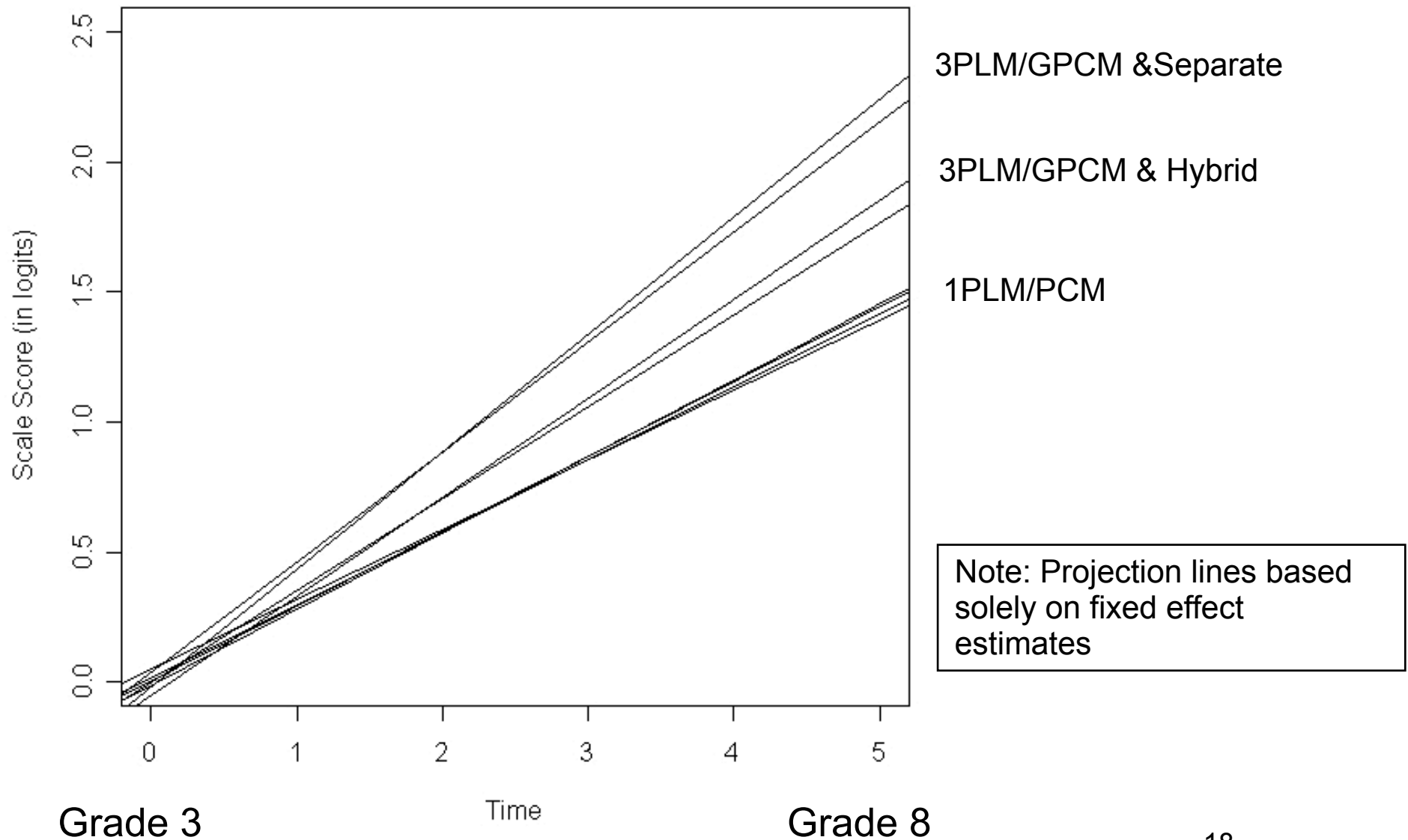
Fixed Effect Estimates

	Separate Calibration			
	1PL		3PL	
	EAP	ML	EAP	ML
Intercept	-0.0144	0.0147	-0.0194	0.0343
Slope	0.2940	0.2804	0.4524	0.4241

	Hybrid Calibration			
	1PL		3PL	
	EAP	ML	EAP	ML
Intercept	0.0053	0.0471	-0.0504	-0.0033
Slope	0.2876	0.2688	0.3807	0.3532

Note: Scale Score Outcome is in Logit Units, Base Year = Grade 3

Comparing Growth Projections



Estimated Correlation Between Random Effect Terms in HLM

	Separate Calibration			
	1PL		3PL	
	EAP	ML	EAP	ML
Student-Level Cor(int, slope)	-0.839	-0.835	-0.212	-0.601
School-Level Cor(int, slope)	-0.674	-0.753	-0.256	-0.473

	Hybrid Calibration			
	1PL		3PL	
	EAP	ML	EAP	ML
Student-Level Cor(int, slope)	-0.764	-0.818	0.037	-0.456
School-Level Cor(int, slope)	-0.625	-0.737	-0.047	-0.280

Correlations of Student and School Slope Estimates by Vertical Scale

	Min	Max	Median	Mean
Student-level	0.525	0.999	0.900	0.850
School-level	0.538	0.999	0.905	0.852

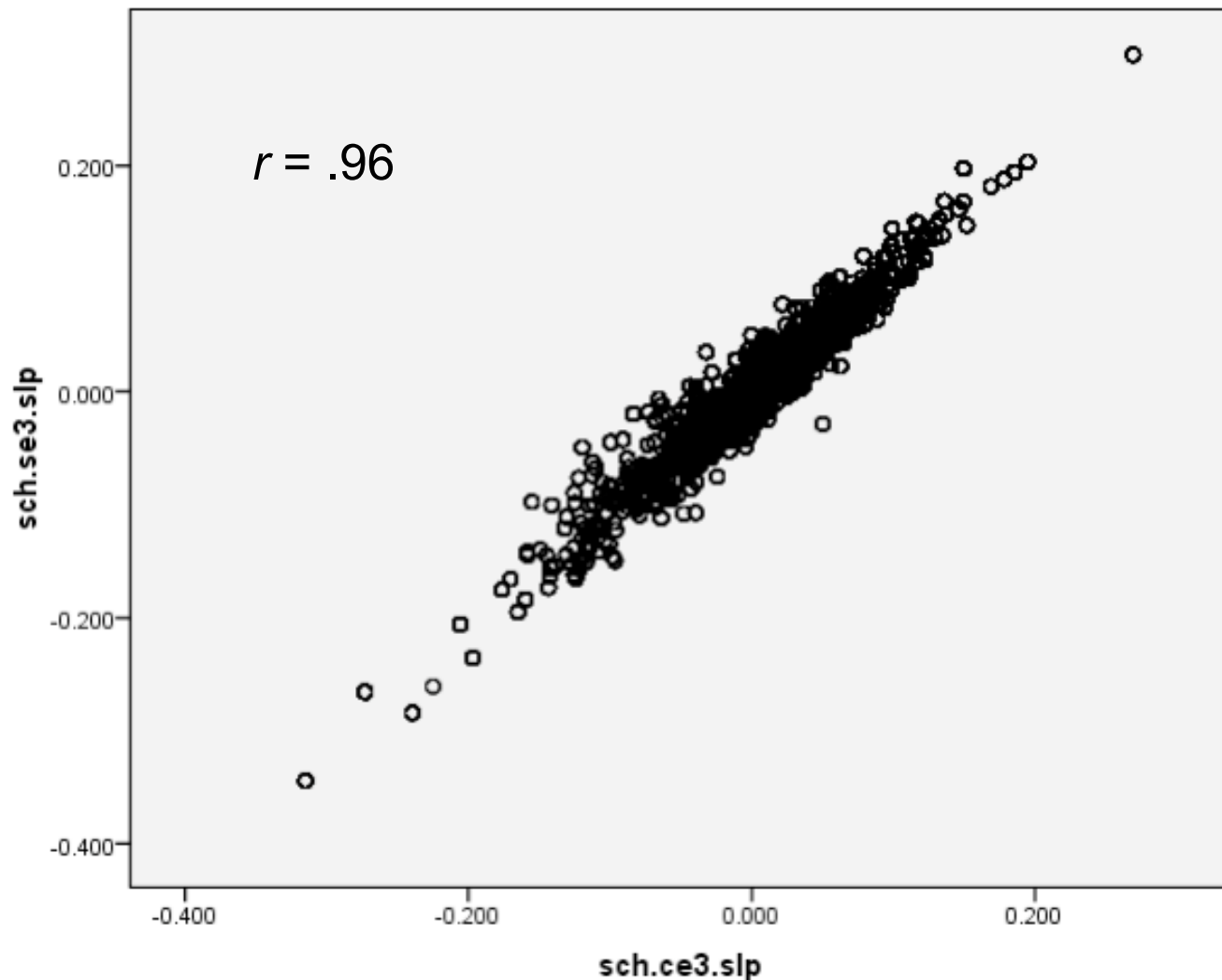
	sep1.eap	sep1.mle	sep3.eap	sep3.mle	hyb1.eap	hyb1.mle	hyb3.eap	hyb3.mle
sep1.eap		0.962	0.843	0.917	0.995	0.963	0.661	0.789
sep1.mle	0.959		0.732	0.901	0.936	0.999	0.525	0.756
sep3.eap	0.851	0.733		0.921	0.877	0.747	0.955	0.934
sep3.mle	0.923	0.904	0.922		0.917	0.912	0.806	0.953
hyb1.eap	0.995	0.932	0.884	0.921		0.939	0.715	0.807
hyb1.mle	0.961	0.999	0.748	0.914	0.936		0.548	0.775
hyb3.eap	0.685	0.538	0.960	0.811	0.736	0.560		0.903
hyb3.mle	0.809	0.767	0.939	0.957	0.824	0.785	0.905	

Note: Value above diagonal = EB slopes student-level; values below = EB slopes school-level

Empirical Bayes Estimates of School-Level Growth

Standard Approach in Colorado:

- Separate
- 3PLM/GPCM
- EAP

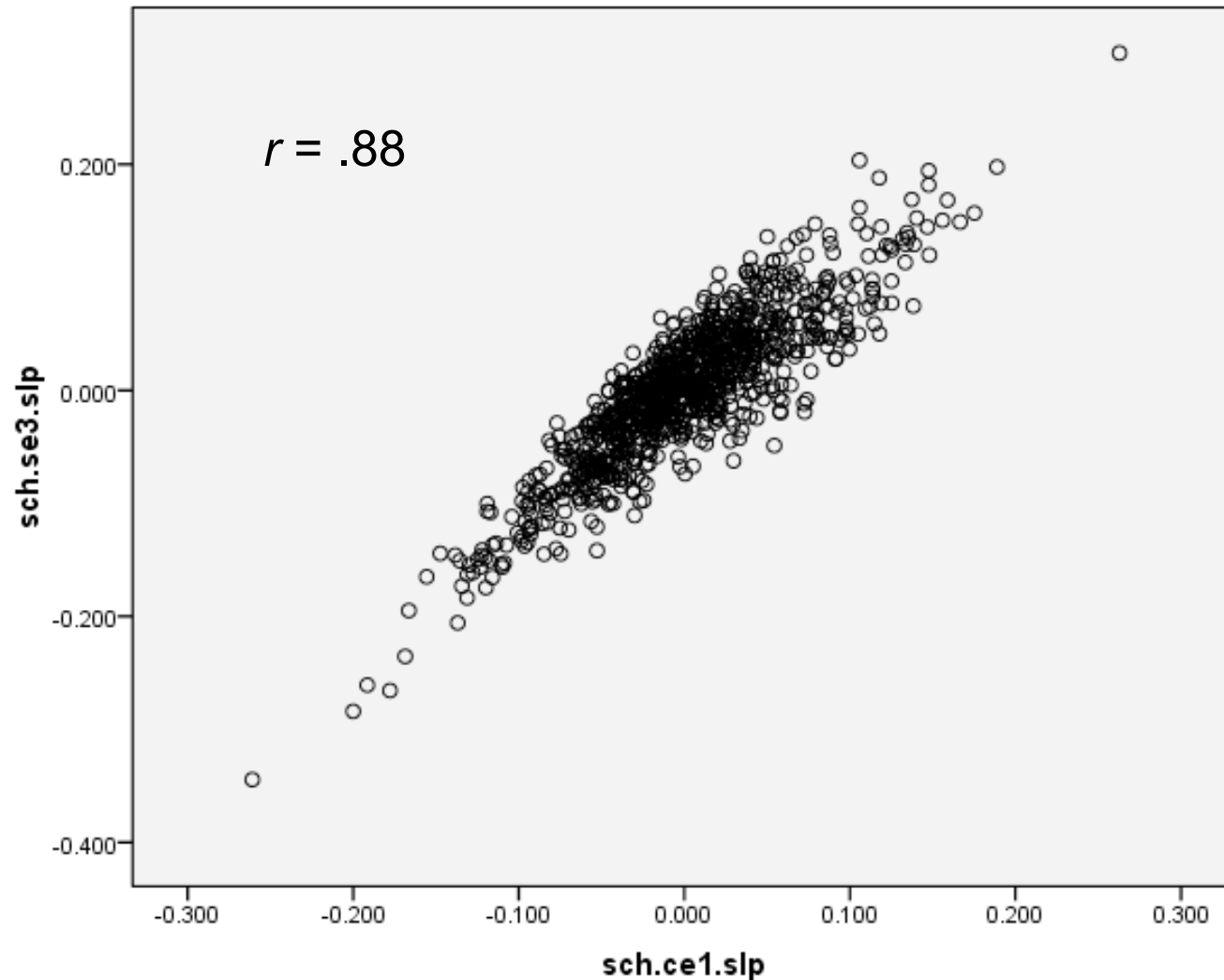


Switch to Hybrid calibration

Empirical Bayes Estimates of School-Level Growth

Standard Approach in Colorado:

- Separate
- 3PLM/GPCM
- EAP

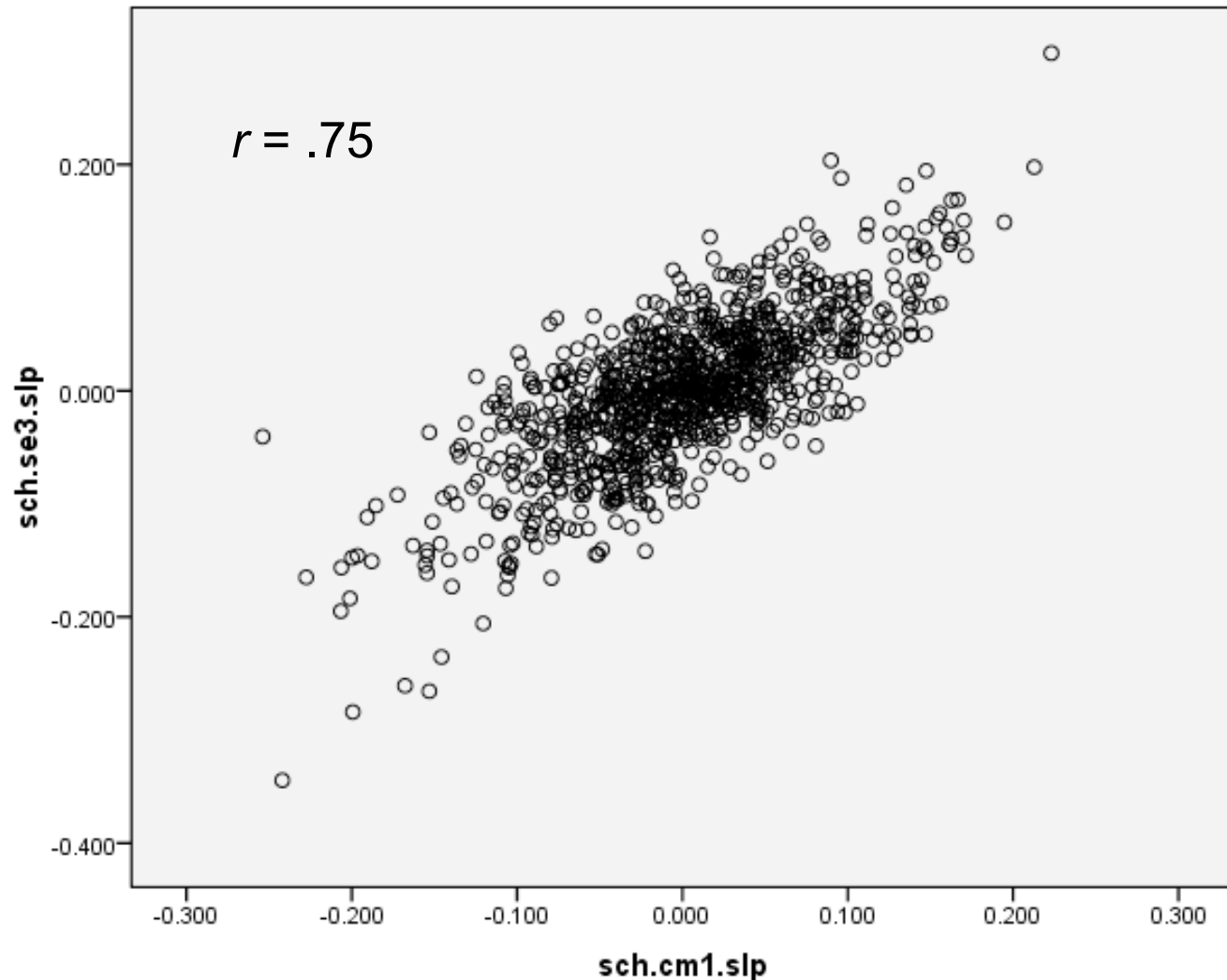


Switch to Hybrid calibration & 1PLM/GPCM²²

Empirical Bayes Estimates of School-Level Growth

Standard Approach in Colorado:

- Separate
- 3PLM/GPCM
- EAP



Switch to Hybrid calibration, 1PLM/GPCM, MLE

Layered Model

$$Y_{i03} = \mu_{03} + \hat{e}_{03} + \varepsilon_{i03}$$

$$Y_{i04} = \mu_{04} + \hat{e}_{03} + \hat{e}_{04} + \varepsilon_{i04}$$

$$Y_{i05} = \mu_{05} + \hat{e}_{03} + \hat{e}_{04} + \hat{e}_{05} + \varepsilon_{i05}$$

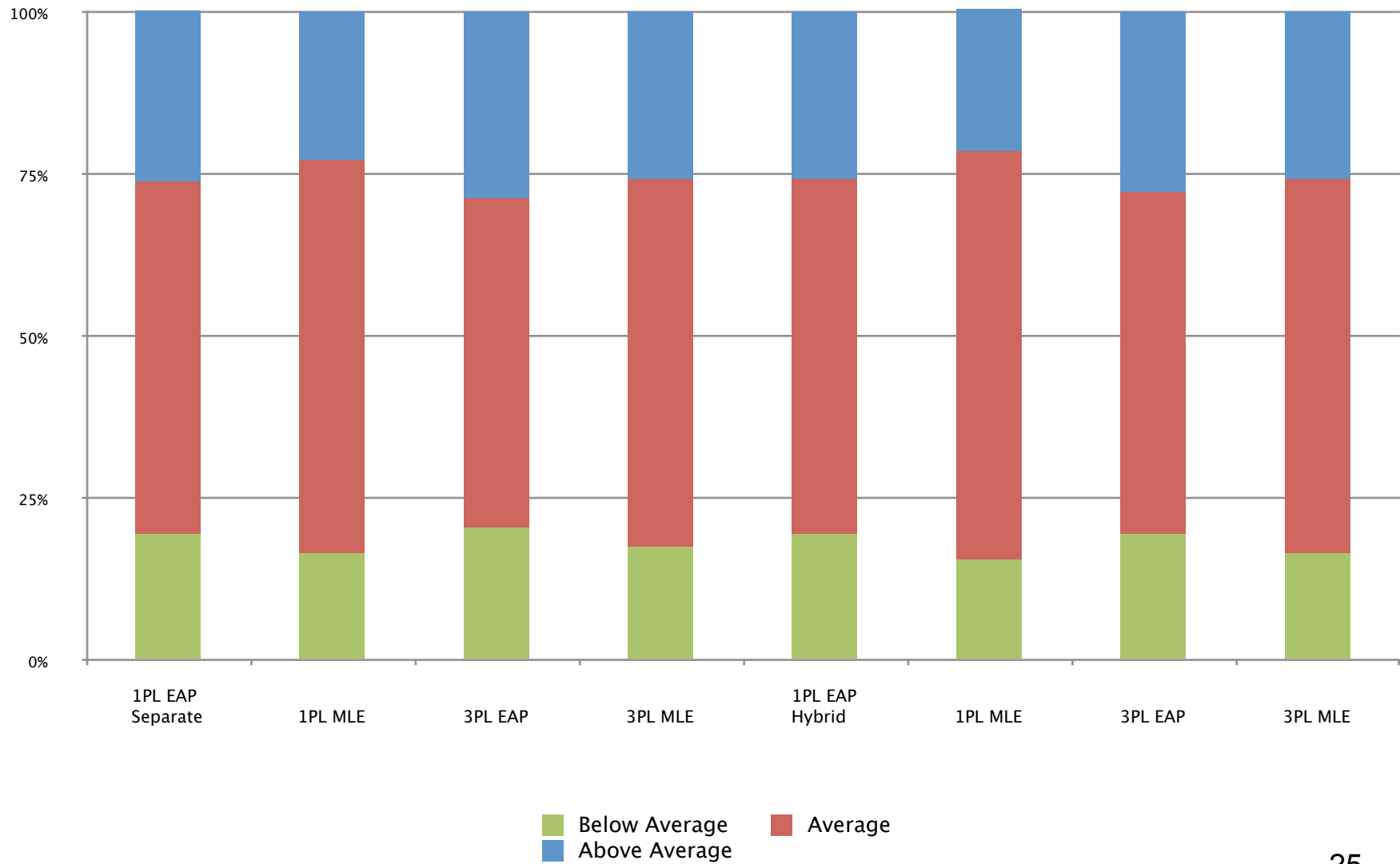
$$Y_{i06} = \mu_{06} + \hat{e}_{03} + \hat{e}_{04} + \hat{e}_{05} + \hat{e}_{06} + \varepsilon_{i06}.$$

Value-Added Parameters of Interest: $\{\hat{e}_4, \hat{e}_5, \hat{e}_6\}$

Notes: Model above assumes complete persistence. Bayesian estimation using non-informative priors.

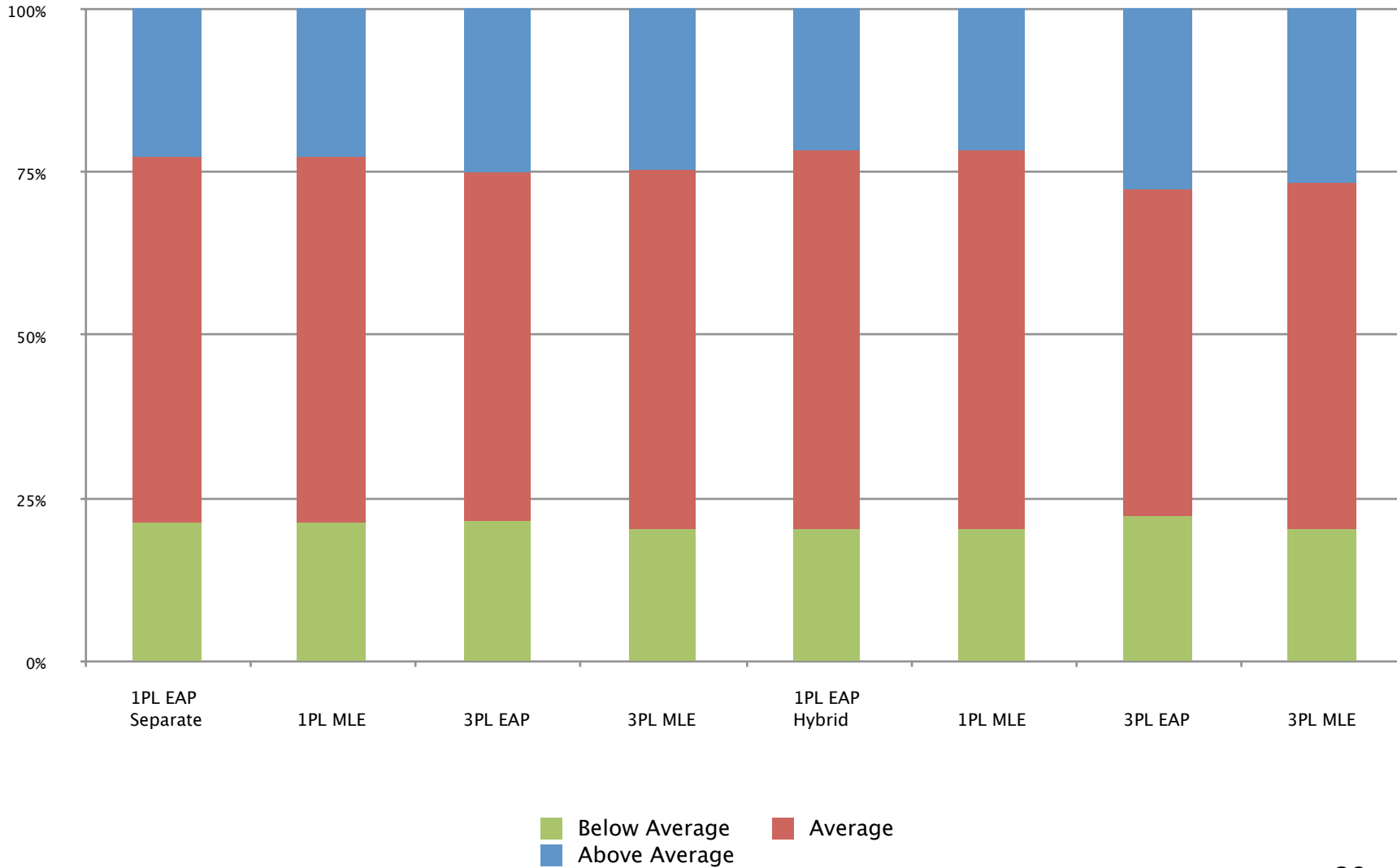
Differences in Schools Identified

Grade 4 Percent of Percent of Schools Identified (N=941)



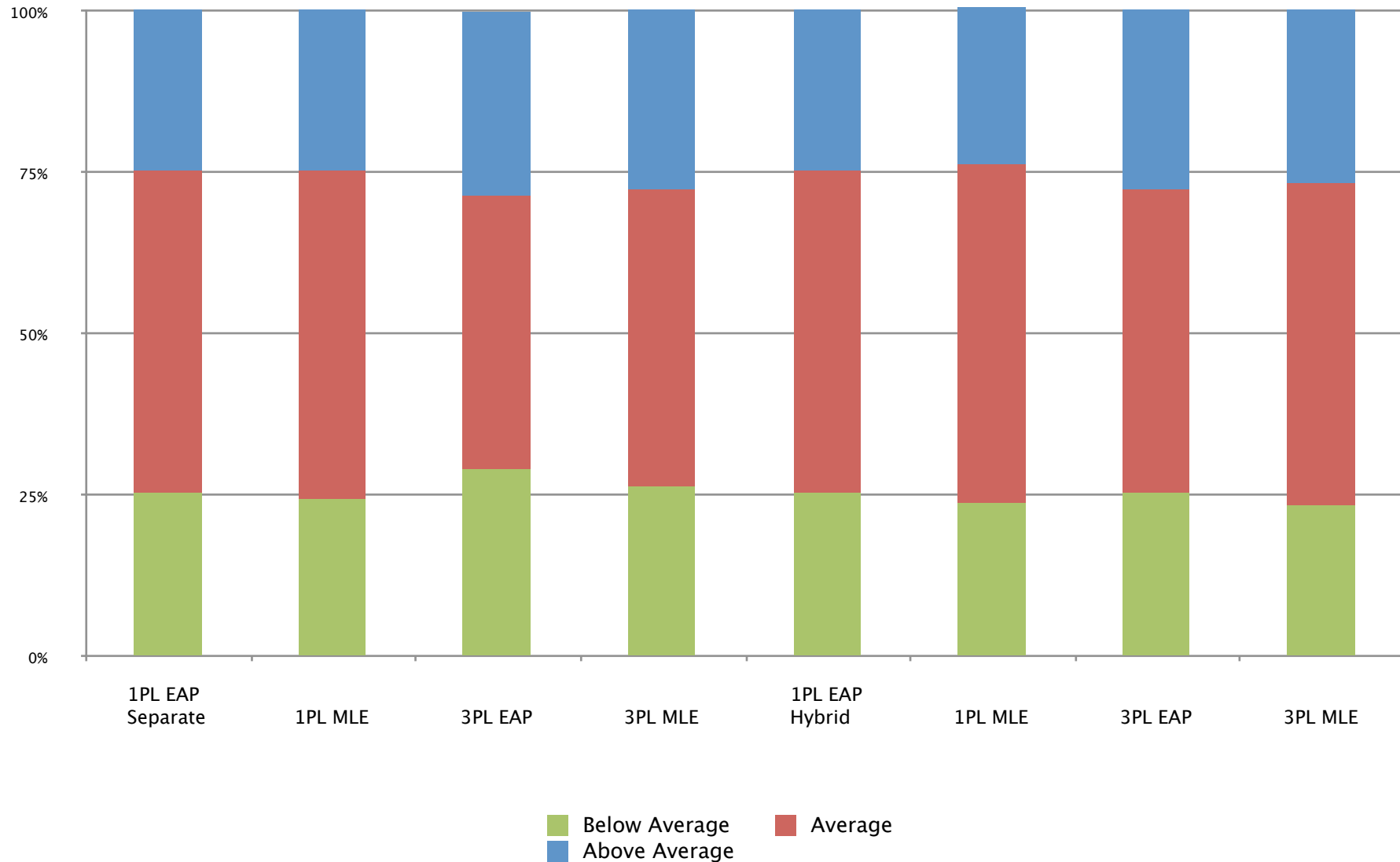
Grade 5

Grade 5 Percent of Percent of Schools Identified (N=950)



Grade 6

Grade 6 Percent of Percent of Schools Identified (N=640)



(Separate, 3PLM, EAP) vs (Hybrid, 1PLM, ML)

Grade 4 - Number of Identified Schools

N = 941		Hybrid 1PL MLE		
		+	0	-
Separate 3PL EAP	+	188	82	1
	0	19	422	38
	-	0	87	104

Grade 5 - Number of Identified Schools

N = 950		Hybrid 1PL MLE		
		+	0	-
Separate 3PL EAP	+	191	51	0
	0	22	451	34
	-	0	48	153

Grade 6 - Number of Identified Schools

N = 640		Hybrid 1PL MLE		
		+	0	-
Separate 3PL EAP	+	147	36	0
	0	8	259	7
	-	0	40	143

Conclusion

- Vertical scales have (largely) arbitrary metrics.
- Absolute interpretations of parametric growth can deceive.
 - Students might appear to grow “faster” solely because of the scaling approach.
 - Can criterion-referencing (i.e., standard-setting) reliably take this into account?
- A better approach might focus on changes in norm-referenced interpretations (but this conflicts with the NCLB perspective on growth).
- The layered model was relatively insensitive to the choice of scale, but there are still some noteworthy differences in numbers of schools identified.

Future Directions

- Full concurrent calibration.
- Running analysis with math tests.
- Joint analysis with math and reading tests.
- Acquiring full panel data.
- Developing a multidimensional vertical scale.

derek.briggs@colorado.edu