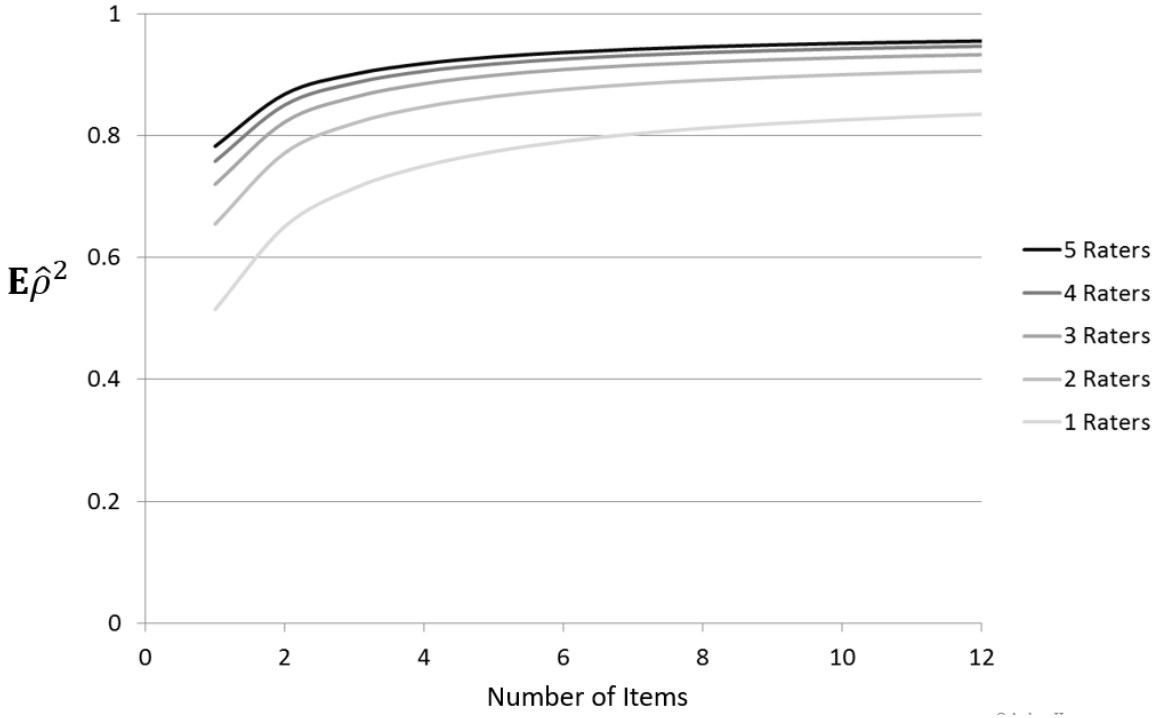
Statistical and psychometric methods for measurement: G Theory, DIF, & Linking



Andrew Ho, *Harvard Graduate School of Education*

The World Bank, Psychometrics Mini Course 2

Washington, DC. June 27, 2018

Learning Objectives from Part I

- How do we "develop and validate" a scale?
  - What is validation?
  - What is reliability?
  - What is factor analysis?
  - What is Item Response Theory?
  - How do we do all this in Stata, and interpret the output accurately?

Learning Objectives for Part II

1. **Generalizability Theory**:
   - *How can we describe and improve the precision of teacher evaluation scores?*

2. **Differential Item Functioning**:
   - *How can we detect items that are not functioning well across countries?*

3. **Linking:**
   - *How can we compare scores across different tests?*

3 hours, or 3 years?

So You Want to Be A Psychometrician?

Becoming a psychometrician, or testing expert, requires years of graduate work.

COURSE WORK AT U. OF IOWA

Ph.D. in Educational Measurement and Statistics

Applied Statistics

- Intermediate Statistical Methods
- Correlation and Regression
- Design of Experiments
- Nonparametric Statistical Methods
- Factor Analysis and Structural Equation Models
- Introduction to Multivariate Statistical Methods
- Topics in Educational Measurement and Statistics

Educational Measurement

- Construction and Use of Evaluation Instruments
- Educational Measurement and Evaluation Theory and Technique in Educational Measurement
- Scaling Methods
- Item Response Theory
- Seminar: Educational Measurement and Evaluation
- Equating and Scaling of Educational Tests
- Generalizability Theory
- Appraisal in Counseling

EDUCATION

The New York Times

As Test-Taking Grows, Test-Makers Grow Rarer

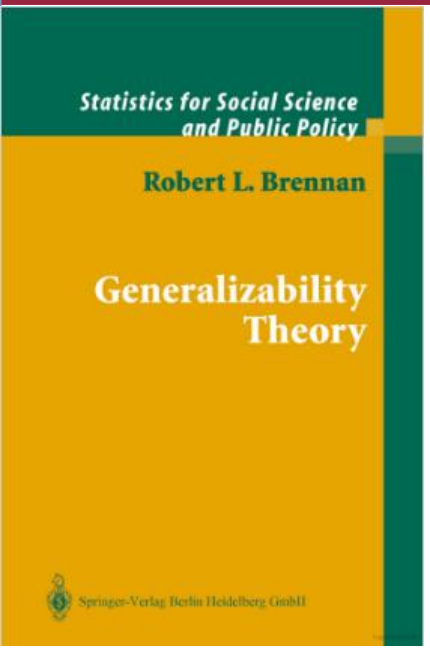
By DAVID M. HERSZENHORN MAY 5, 2006

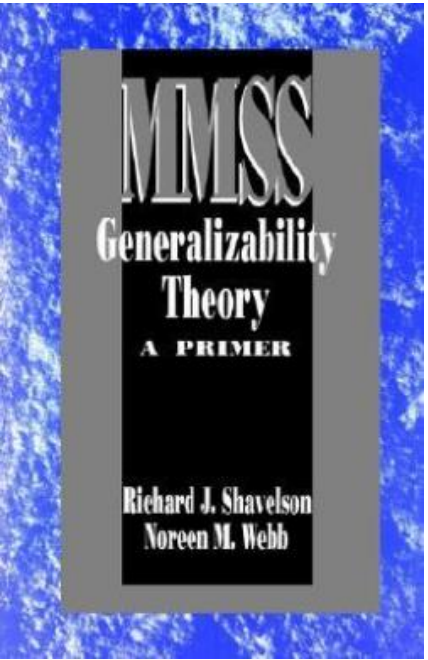
Sz-Shyan Wu is not a Cuban baseball star or a dissident musician. But in urging the United States government to grant him a work visa, the New York State Education Department is arguing that Mr. Wu, too, has talents so rare that bureaucracy must be cut and a red carpet rolled out.

Mr. Wu is a psychometrician or, in plain English, an expert on testing. And testing experts are in high demand.

With federal law requiring wider testing of schoolchildren, the nation faces a critical shortage of people like Mr. Wu with the mathematical, scientific, psychological and educational skills to create tests and analyze the results. The problem has sent states, testing companies and big school districts into a heated hiring competition, with test companies offering salaries as high as $200,000 a year or more plus perks.

G Theory: References & Examples

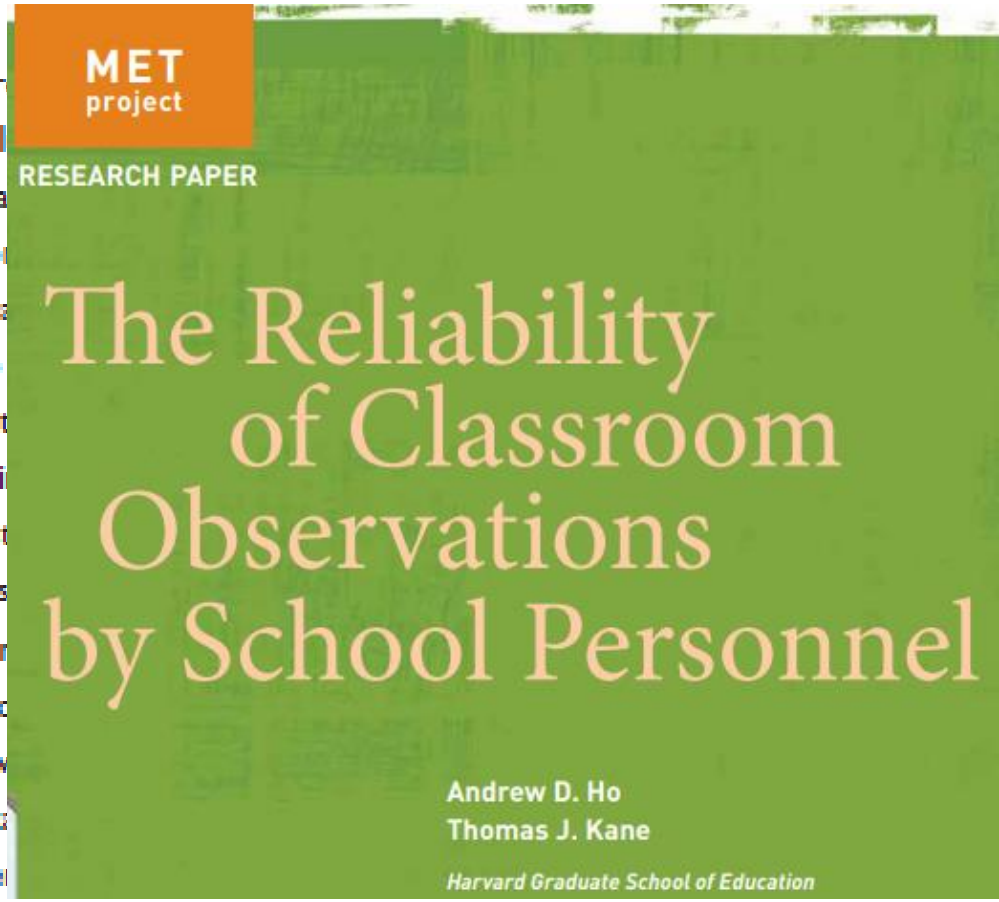




When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study

Heather C. Hill[1], Charalambos Y. Charalambous[2], and Matthew A. Kraft[1]

In recent years, interest has gr
as a means to several ends, incl
evaluation, and impact evalua
tions. Although education p
developed numerous observa
poses, many developers fail to
instrument use. In this article, t
observation to succeed in its ai
must be developed. These syst
vational instruments but also s
reliable and cost-efficient scor
ment, training, and certificatio
might be developed and improv
example that applies generaliz
matics observational instrume

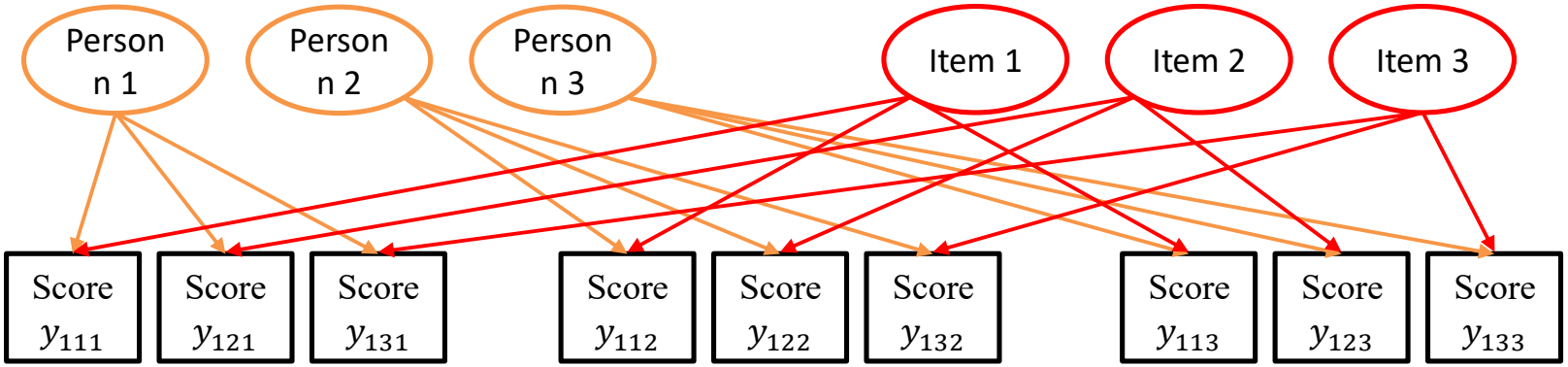


G Theory: 5 Essential Concepts

1. It's about SCORES.
   - Reliability and precision are properties of SCORES not TESTS.
2. What is reliability?
   - Reliability is the correlation between scores across a replication: $\rho_{XX'}$
3. The "Rule of 2"
   - If you don't have at least two replications (of items, lessons, raters, occasions), you cannot estimate relevant variance.
4. The "G Study"
   - A multilevel model with crossed, random effects gives us "variance components."
5. The "D Study"
   - Once we know variance components, we can increase score precision strategically, by averaging over replications.

Measurement as crossed random effects

- Canonical measurement data have a crossed-effects design:



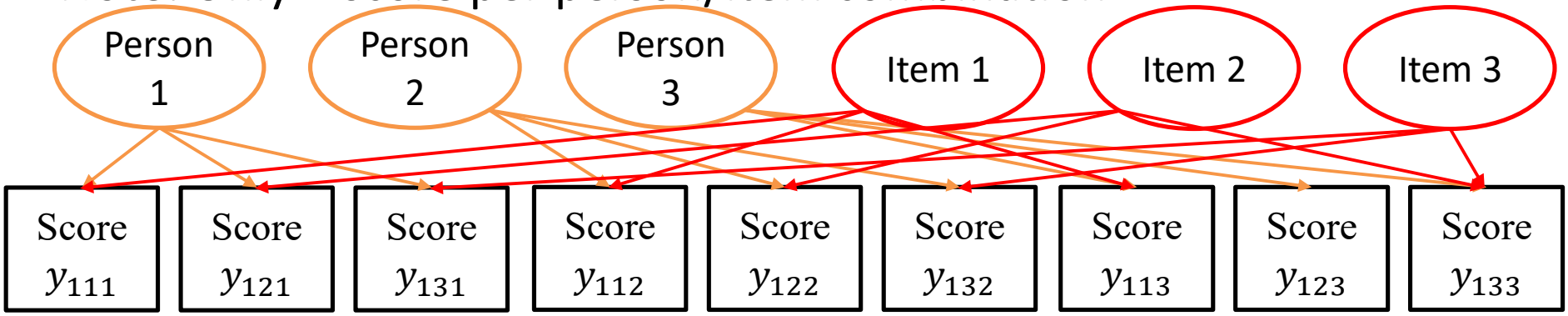
- y_{ijk} is a response replication i to item j by person k.
- We are used to seeing many replications i (students in schools). In measurement we generally have only 1 replication per person-item combination, so we can drop the subscript.

The Measurement Model

- A response i to item j by person k

$$y_{ijk} = \mu + \zeta_j + \zeta_k + \varepsilon_{ijk};$$
$$\zeta_j \sim N(0, \psi_1);$$
$$\zeta_k \sim N(0, \psi_2);$$
$$\varepsilon_{ijk} \sim N(0, \theta).$$

- Note: Only 1 score per person/item combination

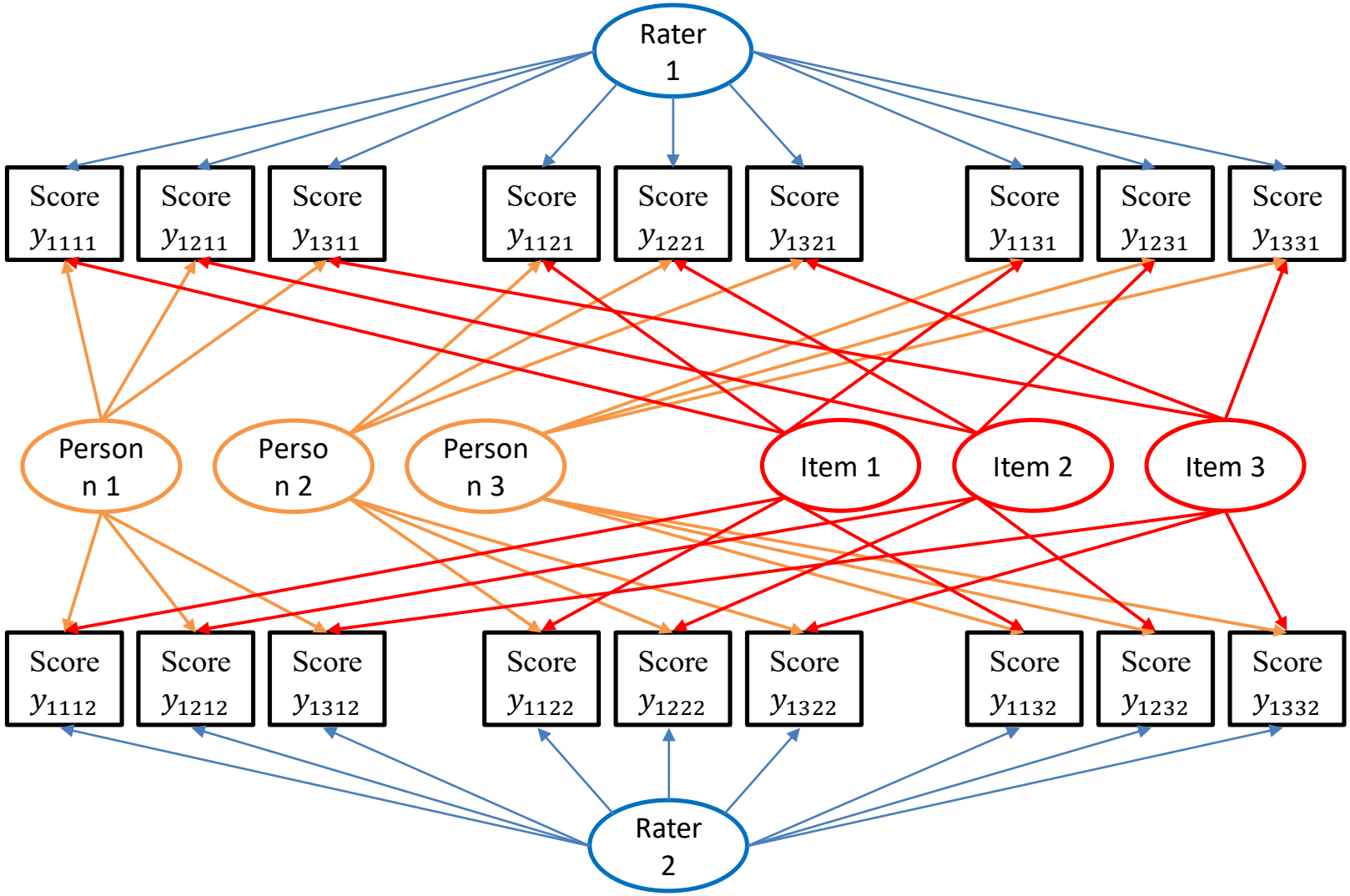


- μ – Overall average score
- ζ_j – Item location (easiness), ψ_1 –variance of item effects
- ζ_k –Person location (proficiency), ψ_2 –variance of person effects
- ε_{ijk} –Person-item interactions and other effects, θ –error variance

Two Intraclass Correlations (Reliabilities)

- A response i to item j by person k : $y_{ijk} = \mu + \zeta_j + \zeta_k + \varepsilon_{ijk}$;
$$\zeta_j \sim N(0, \psi_1);$$
$$\zeta_k \sim N(0, \psi_2);$$
$$\varepsilon_{ijk} \sim N(0, \theta).$$
- $\mu -$ Overall average score
- $\zeta_j -$ Item location (easiness). Variance: ψ_1
- $\zeta_k -$Person location (proficiency). Variance: ψ_2
- $\varepsilon_{ijk} -$Person-item interactions and other effects. Variance: θ
- Intraclass correlation: $\rho = \frac{\psi_2}{\psi_2 + \theta}$. The correlation between two item responses within persons. The proportion of *relative* response variation due to persons.
- Intraclass correlation: $\rho_\alpha = \frac{\psi_2}{\psi_2 + \frac{\theta}{n_j}}$. Cronbach's alpha: The correlation between two average (or sum) scores within persons. The proportion of *relative* score variance due to persons.

The Crossed Effects Model with Raters



G Theory Data Structure (Wide)

From Brennan (2002), 15 teachers, 3 raters, 5 items.

```
. table person item rater, c(mean score)
```

	rater and item														
	1					2					3				
person	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	1	3	1	1	1	1	3	2	2	1	1	2	1	1	1
2	1	2	2	1	1	3	3	2	3	2	1	2	2	2	2
3	3	4	3	3	3	3	4	3	3	3	4	3	3	3	2
4	2	3	2	2	2	2	4	4	3	3	1	2	3	2	1
5	2	2	2	2	1	1	3	3	2	3	1	2	2	1	1
6	3	4	3	3	2	4	4	4	4	4	2	4	4	3	3
7	4	4	2	3	4	4	5	5	5	4	3	4	3	4	3
8	2	3	4	3	2	3	4	3	4	2	3	3	4	3	2
9	2	3	3	3	3	2	3	4	4	2	1	1	2	2	1
10	2	3	2	3	1	3	2	3	3	2	2	3	2	2	1
11	2	3	3	3	2	2	3	3	4	3	1	2	3	3	2
12	2	2	3	3	2	2	3	3	3	2	2	3	4	2	2
13	1	2	1	2	2	1	2	3	2	2	1	1	2	1	1
14	2	2	3	2	2	2	3	4	4	2	1	2	2	2	1
15	2	3	2	2	1	3	3	4	3	4	2	2	3	1	2

G Theory Data Structure (Long)

	person	item	rater	score
1	1	1	1	1
2	1	2	1	3
3	1	3	1	1
4	1	4	1	1
5	1	5	1	1
6	1	1	2	1
7	1	2	2	3
8	1	3	2	2
9	1	4	2	2
10	1	5	2	
11	1	1	3	
12	1	2	3	
13	1	3	3	
14	1	4	3	
15	1	5	3	
16	2	1	1	
17	2	2	1	
18	2	3	1	
19	2	4	1	
20	2	5	1	
21	2	1	2	
22	2	2	2	
23	2	3	2	
24	2	4	2	3
25	2	5	2	2

From Brennan (2002), 15 teachers, 3 raters, 5 items.

See here and the .do file for reshaping data from double/triple-wide formats.

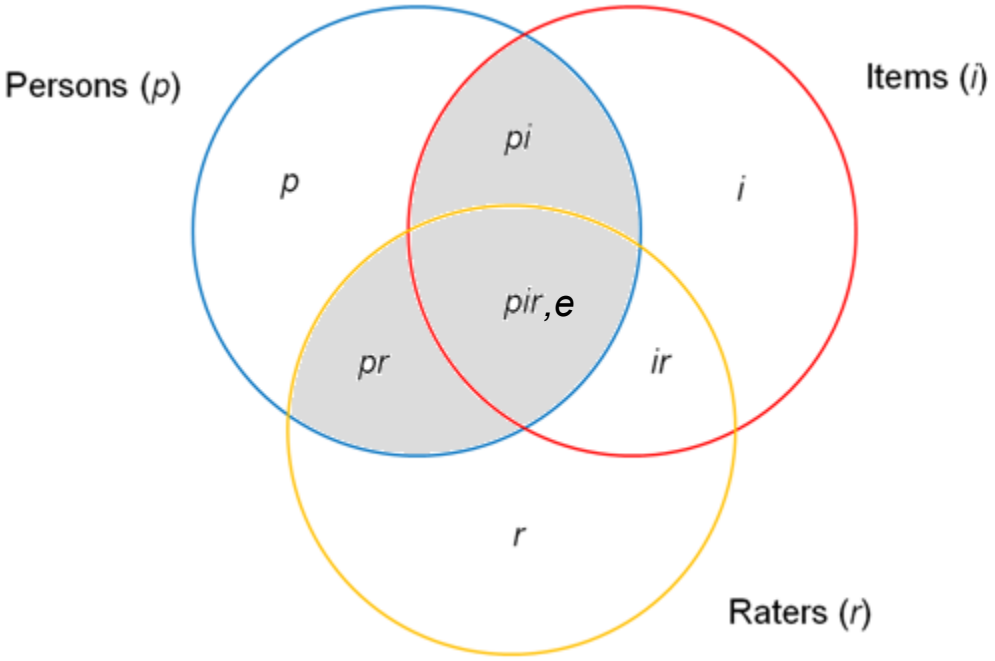
```
*-------------------------------------------------------------------------------
* Reshape from double-wide to double-long
* See:  http://www.ats.ucla.edu/stat/stata/faq/doublewide.htm
*-------------------------------------------------------------------------------

    local i = 1
    foreach var of varlist r1item1-r3item5 {
        rename `var' score`i'
        local i = `i'+1
    }

    reshape long score, i(person) j(seq)
    recode seq (1 6 11=1) (2 7 12=2) (3 8 13=3) (4 9 14=4) (5 10 15=5), gen(item)
    recode seq (1/5=1) (6/10=2) (11/15=3), gen(rater)
    drop seq
    order person item rater score
    label variable item ""
    label variable rater ""
    table person item rater, c(mean score)

    save pxixr_long.dta, replace
```

The "$p \times i \times r$" Measurement Model

- A response by person p to item i, rated by rater r:

$$y_{pir} = \mu + \zeta_p + \zeta_i + \zeta_r + \zeta_{pi} + \zeta_{pr} + \zeta_{ir} + \varepsilon_{pir,e};$$
$$\zeta_p \sim N(0, \psi_p); \zeta_{pi} \sim N(0, \psi_{pi}); \zeta_{pr} \sim N(0, \psi_{pr})$$
$$\zeta_i \sim N(0, \psi_i); \zeta_r \sim N(0, \psi_r); \zeta_{ir} \sim N(0, \psi_{ir})$$
$$\varepsilon_{pir,e} \sim N(0, \theta).$$

- The Stata "mixed" command will estimate 7 variance components.



Stata Output

```
. mixed score || _all: R.person || _all: R.item || _all: R.rater || _all: R.pXi
  || _all: R.pXr || _all: R.iXr, variance reml nolog
```

score	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	2.475556	.344058	7.20	0.000	1.801214	3.149897

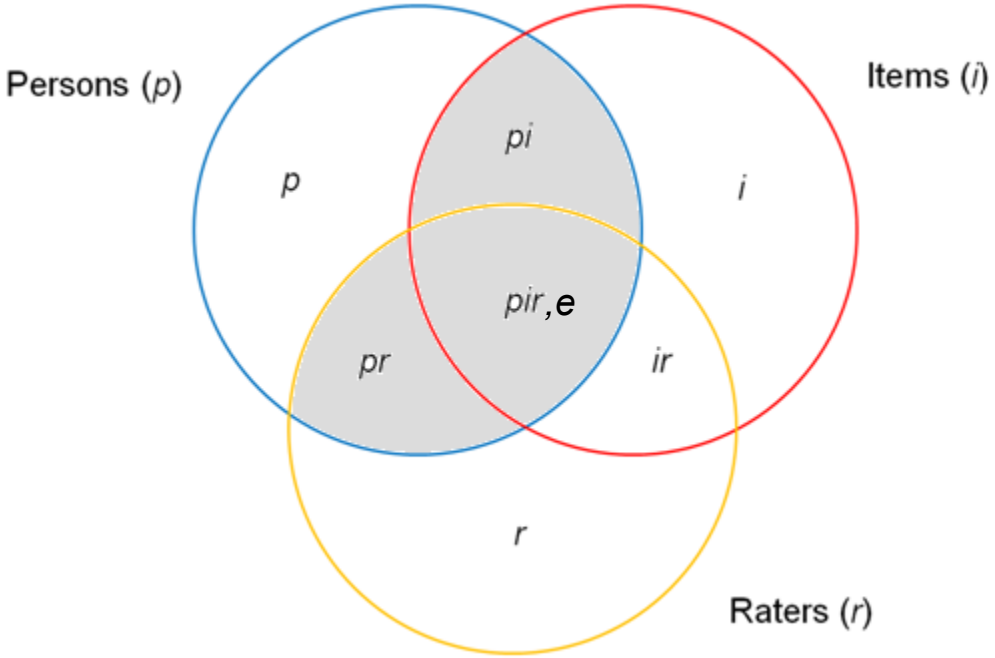
Random-effects Parameters		Estimate	Std. Err.	[95% Conf.	Interval]
_all: Identity	var(R.person)	.3777783	.1591872	.1654093	.8628078
_all: Identity	var(R.item)	.1344455	.1046619	.0292355	.6182755
_all: Identity	var(R.rater)	.1876201	.1972506	.0238993	1.472899
_all: Identity	var(R.pXi)	.0422223	.0273449	.011865	.1502505
_all: Identity	var(R.pXr)	.0490476	.0281733	.0159104	.1512009
_all: Identity	var(R.iXr)	.0139683	.0159899	.0014816	.1316866
	var(Residual)	.2649205	.0354015	.2038774	.3442406

How do we interpret variance components?

Source	var	sd	percentage
p	0.3778	0.6146	35.31%
i	0.1344	0.3667	12.56%
r	0.1876	0.4332	17.53%
pi	0.0422	0.2055	2.04%
pr	0.0490	0.2215	2.37%
ir	0.0140	0.1182	0.58%
pir,e	0.2649	0.5147	10.08%

Source	Description
p	Variance in teacher proficiency, persistent across items and raters
i	Item variance. Some item scores are higher across teachers and raters
r	Rater variance. Some rater scores are higher across teachers and items
pi	Some teachers score higher on some items than others, across raters.
pr	Some raters score some teachers higher than others, across items.
ir	Some raters score some items higher than others, across teachers.
pir,e	Some raters score some teachers higher on some items, and random error.

The p x i x r D Study for Relative Error

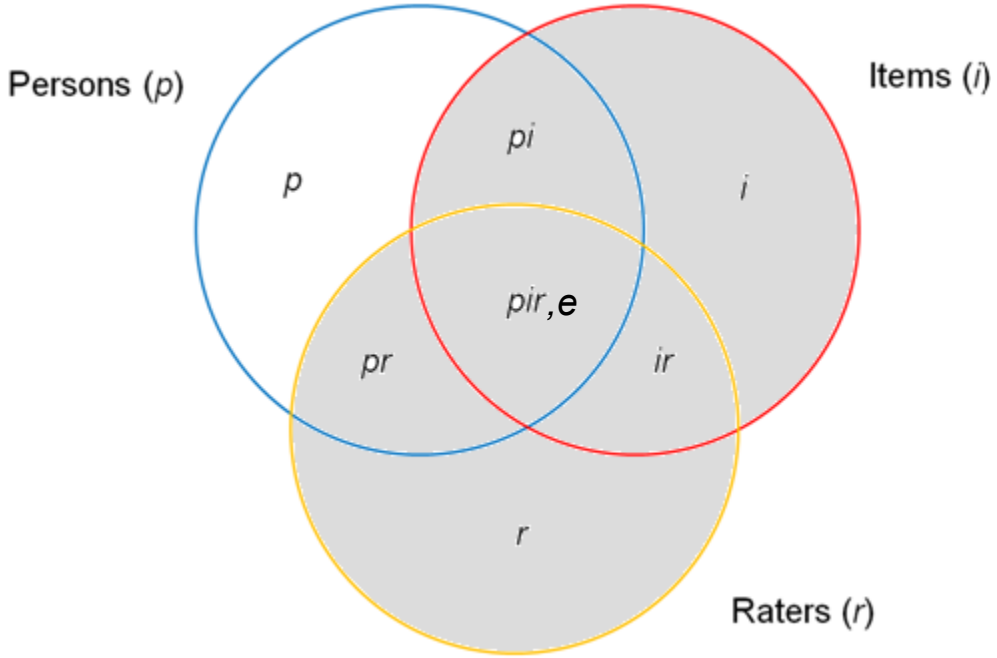


- Identify the variance components that lead to changes in relative position.
- Includes all variance components that intersect with p.
- Why not σ_i^2? Items that are more or less difficult… for everybody.
- Why not σ_r^2?
- Why not σ_{ir}^2?
- Variance components refer to single-unit replications, so we must divide by relevant numbers of items/raters to obtain error for average scores (over items/raters).
- Mnemonic: Divide by the n's in subscript in the numerator (besides p).

$$\sigma_\delta^2 = \frac{\zeta_{pi}}{n_i} + \frac{\zeta_{pr}}{n_r} + \frac{\zeta_{pir,e}}{n_i n_r}$$

$$\mathbf{E}\rho^2 = \frac{\zeta_p}{\zeta_p + \sigma_\delta^2}$$

The p x i x r D Study for Absolute Error

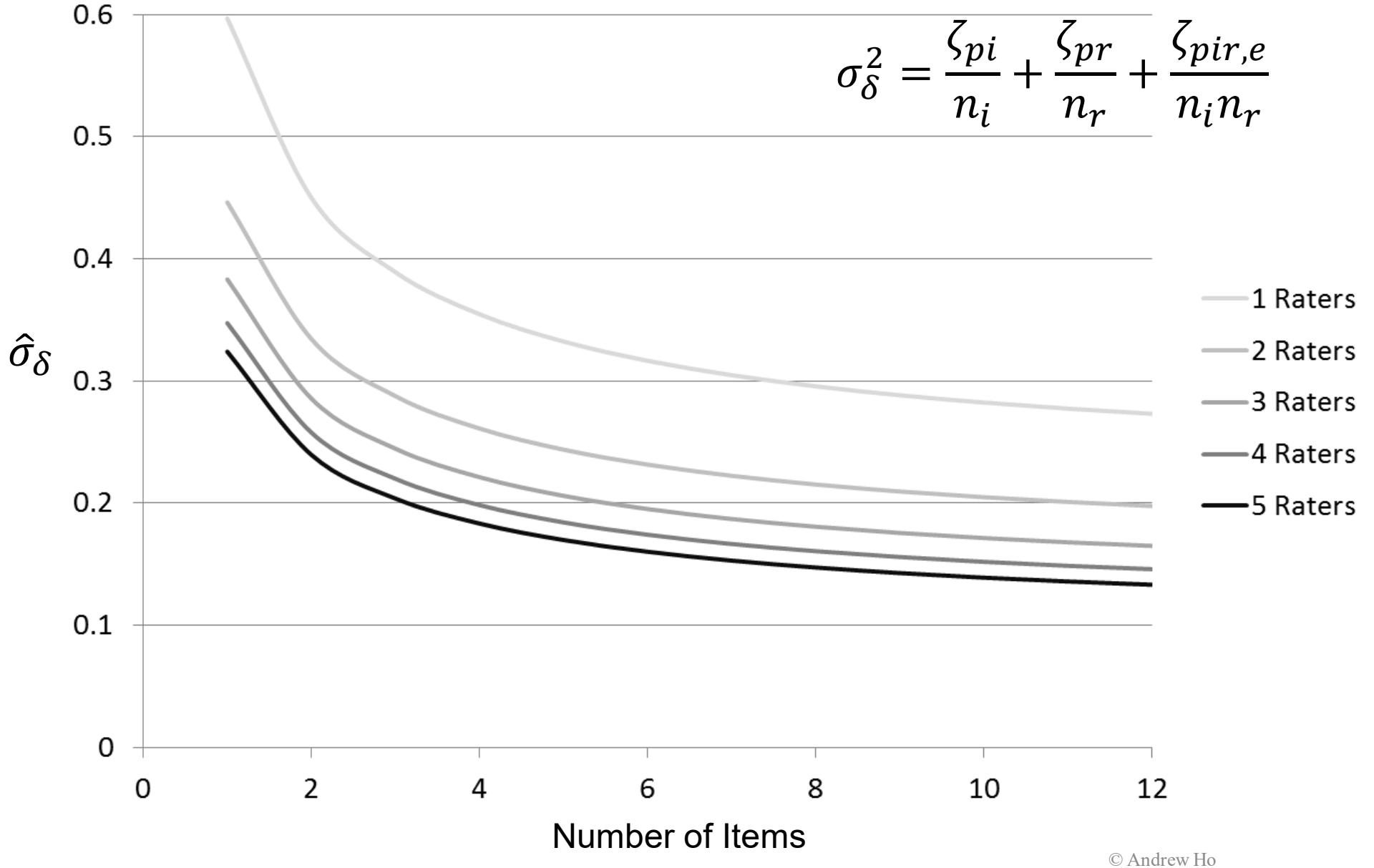


- Identify the variance components that lead to changes in absolute position.
- Includes all variance components besides σ_p^2.
- Why include σ_i^2? Items that are more or less difficult for everybody will lead to changes in absolute position.
- Why include σ_r^2?
- Why include σ_{ir}^2?

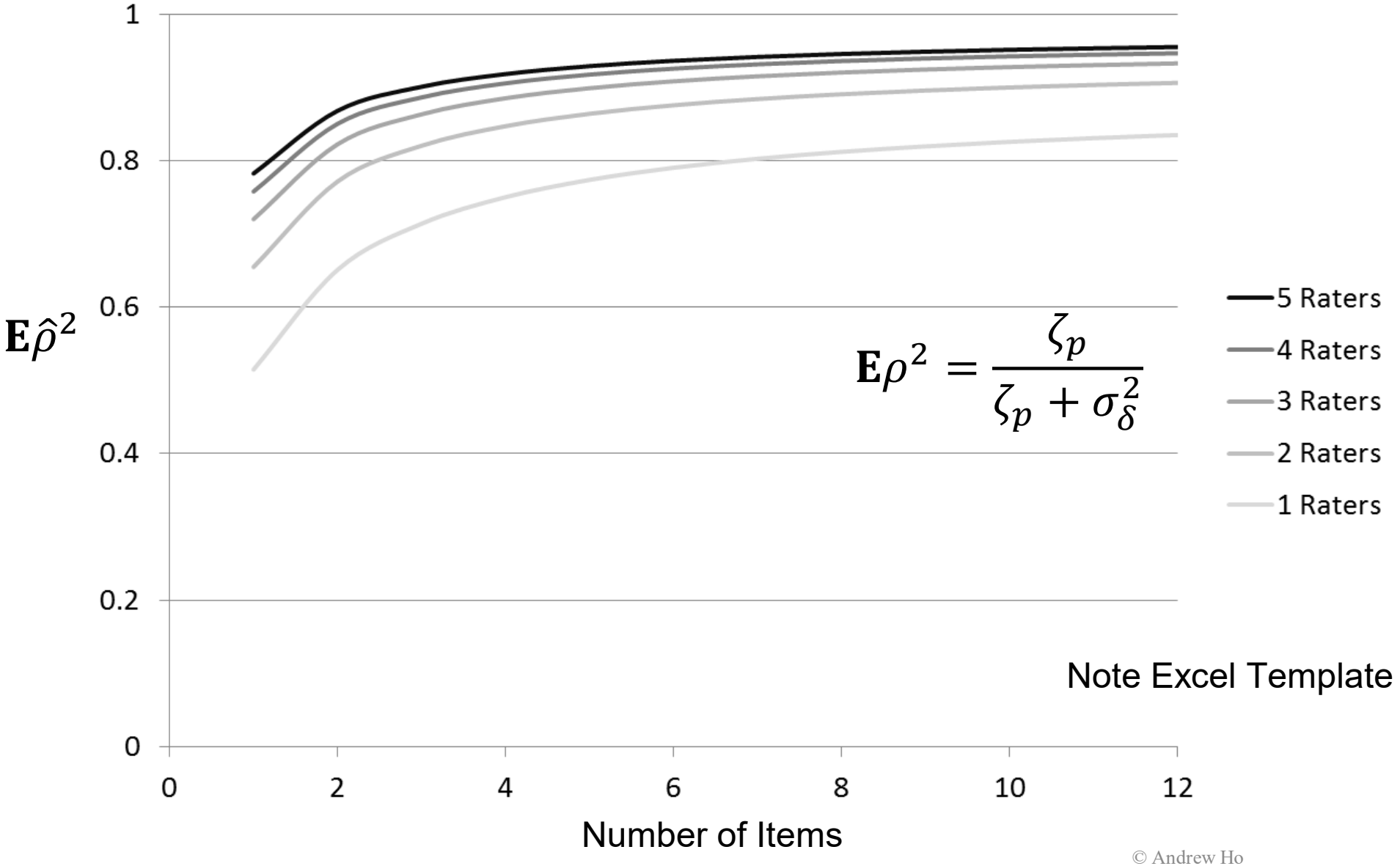
$$\sigma_\Delta^2 = \frac{\zeta_i + \zeta_{pi}}{n_i} + \frac{\zeta_r + \zeta_{pr}}{n_r} + \frac{\zeta_{ir} + \zeta_{pir,e}}{n_i n_r}$$

$$\Phi = \frac{\zeta_p}{\zeta_p + \sigma_\Delta^2}$$

D Study for the Relative Standard Error of Measurement, $\hat{\sigma}_\delta$



D Study for the Generalizability Coefficient for Relative Error, $\mathbf{E}\hat{\rho}^2$



Ho & Kane (2012) Table 10 D Study

Table 10

RELIABILITY ACHIEVED WITH DIFFERENT ALLOCATIONS OF CLASSROOM OBSERVATIONS

SCENARIO	Rater Combination	Lesson 1	Lesson 2	Lesson 3	Lesson 4	Lesson 5	Lesson 6	Total Full Observations	Implied Reliability
SCENARIO 1	Own Administrator	X						2	.59
	Same-Cert Peer (Full Obs)		X						
SCENARIO 2	Own Administrator	X	X					4	.66
	Same-Cert Peer (Full Obs)			X	X				
SCENARIO 3	Own Administrator	X	X					4	.69
	Same-Cert Peer 1 (Full Obs)			X					
	Same-Cert Peer 2 (Full Obs)				X				
SCENARIO 4	Own Administrator	X	X					4	.72
	Same-Cert Peer 1 (Full Obs)			X					
	Same-Cert Peer 2 (15 Min)				X				
	Same-Cert Peer 3 (15 Min)					X			
	Same-Cert Peer 4 (15 Min)						X		

Learning Objectives for Part II

1. **Generalizability Theory**:
   - *How can we describe and improve the precision of teacher evaluation scores?*

2. **Differential Item Functioning**:
   - *How can we detect items that are not functioning well across countries?*

3. **Linking:**
   - *How can we compare scores across different tests?*

Operationalizing Differential Item Functioning

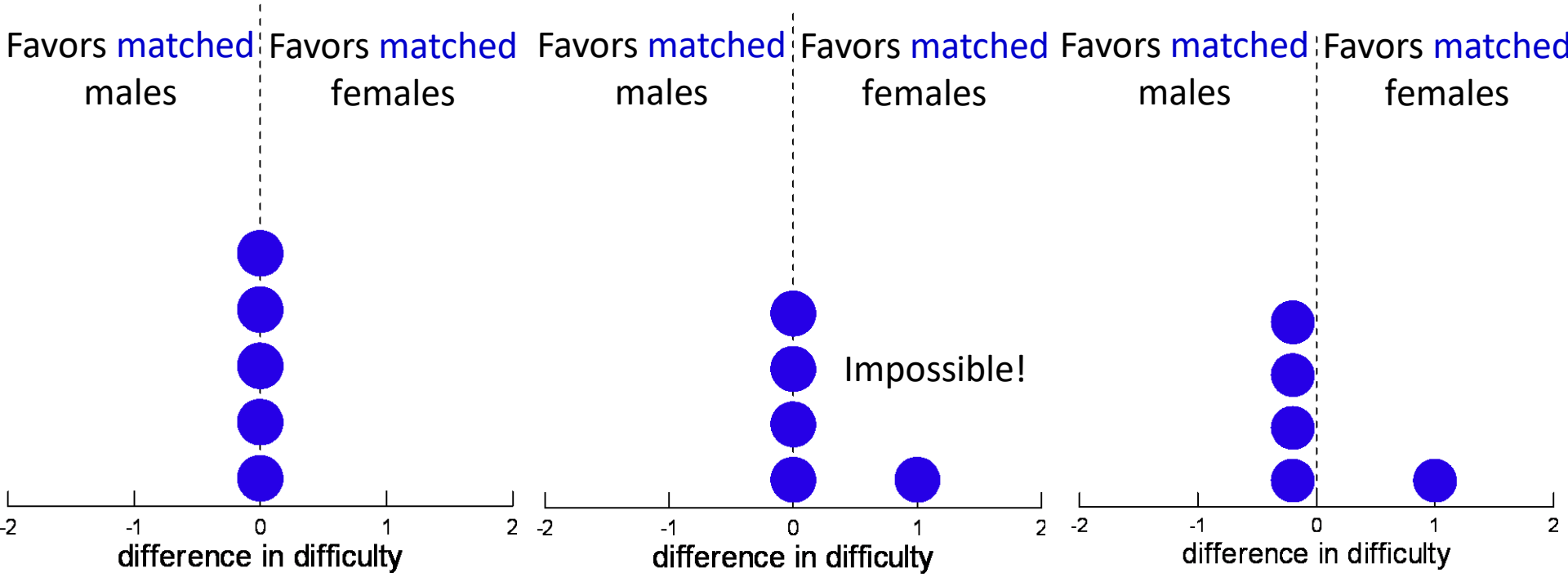
1. Define a "matching criterion" internal or external to the item or test whose differential functioning you wish to assess.
   - Generally a total test score (or θ), an external test score, a test score minus the target item ("rest" score), or a chosen subset of items that serve as a reference.

2. Examine group differences conditional on (at identical values of) the matching criterion.
   - Finding a trustworthy matching criterion is a challenge.
   - It must be free of differential functioning itself, lest it distort the estimation of differential functioning.

3. Flag items with DIF for review by a content committee.
   - Even if you find DIF, it may be construct-relevant!
   - As in Mini Course I, content is king.

Internal Matching Criteria

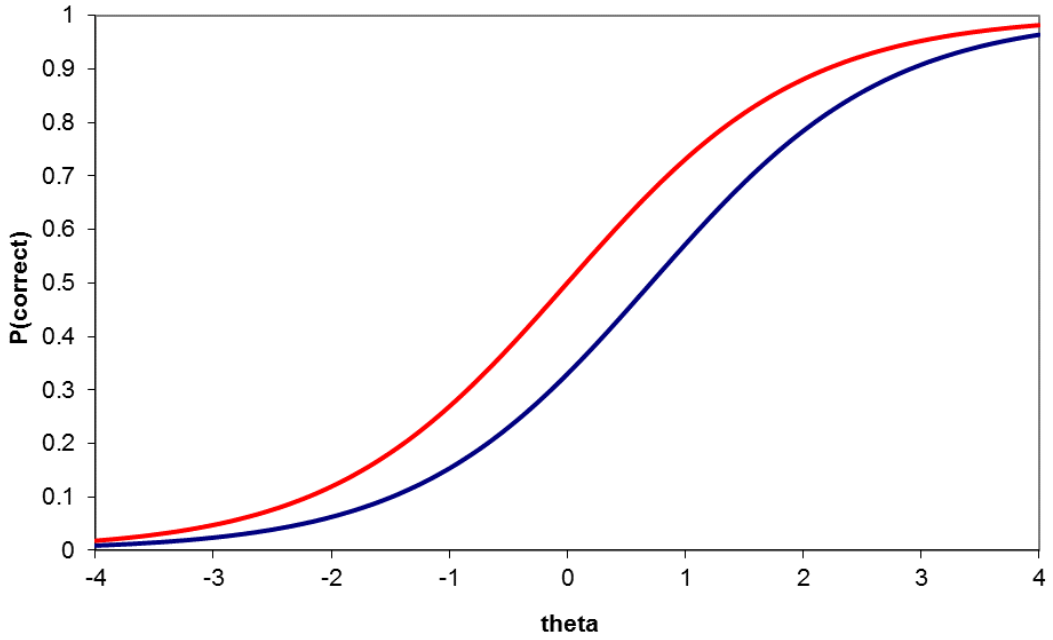
If you use a total score (or θ) to match, DIF ends up being relative. DIF across all items will average to (approximately) 0.



Uniform vs. Nonuniform DIF

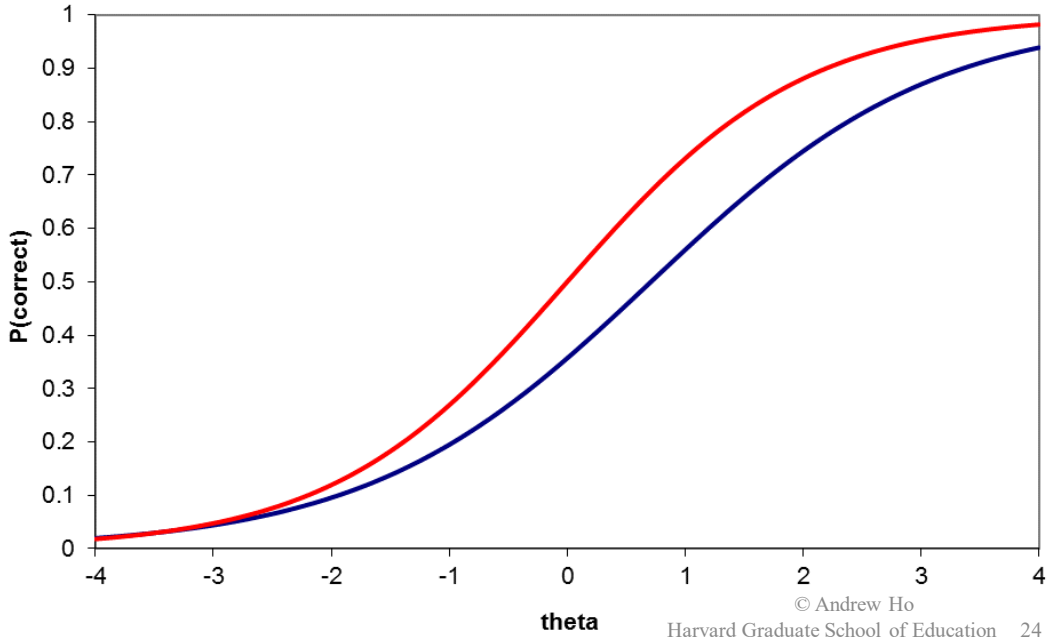
Uniform DIF

Conditional on theta, this item is more difficult for the blue group than the red group.



Nonuniform DIF

Conditional on theta, the item is more difficult for the blue group than the red group, particularly for higher scoring examinees.



The Cochran-Mantel-Haenszel Test

- Stratify people by total score (the matching criterion)
- Within each stratum, tabulate 2x2 right-wrong by group and calculate the odds ratio within each matched value.

	Correct	Incorrect	Total
Reference	n_{11k}	n_{12k}	$n_{1.k}$
Focal	n_{21k}	n_{22k}	$n_{2.k}$
Total	$n_{.1k}$	$n_{.2k}$	n_k

- Combine the ratios (weighted average odds ratio over k)
  - Some report the common odds ratio (1 under null).
  - Others report the log of this odds ratio (0 under null).

$$\text{OR}_{\text{MH}} = \frac{\sum_{k=1}^{K-1} n_{11k} n_{22k} / n_k}{\sum_{k=1}^{K-1} n_{12k} n_{21k} / n_k}$$

- This test statistic is chi-square distributed with 1 *df*

MH Test on a Math/Science test, by gender

```
. tab(female), summarize(sumscore)
```

1=female, 0=male	Summary of sumscore Mean	Std. Dev.	Freq.
0	4.8344284	1.8915108	761
1	4.4925575	1.8140931	739
Total	4.666	1.8610265	1,500

```
. difmh q1-q9, group(female) noyates // Group var
```

Mantel-Haenszel DIF Analysis

Item	Chi2	Prob.	Odds Ratio
q1	12.9350	0.0003	1.6053
q2	1.9489	0.1627	1.1809
q3	6.9466	0.0084	1.4543
q4	4.0939	0.0430	0.7879
q5	5.3504	0.0207	0.7011
q6	0.5855	0.4442	1.1046
q7	1.9423	0.1634	0.8359
q8	2.3383	0.1262	0.7761
q9	2.2098	0.1371	0.8294

- Large χ^2 values (above 1.96^2) indicate statistically significant differences.
- Odds ratios over 1 favor females.
- Odds of a female getting q1 correct is 60% $[OR - 1]$ higher than males conditional on total score.
- If P_m were 50%, P_f would be $[\frac{OR}{1+OR}]$ =62%

Odds Ratio	Logits log(OR)	50% to… (OR/(1+OR))
1.0	0.0	50%
1.5	0.4	60%
2.0	0.7	67%
2.5	0.9	71%
3.0	1.1	75%
3.5	1.3	78%
4.0	1.4	80%
4.5	1.5	82%
5.0	1.6	83%
5.5	1.7	85%
6.0	1.8	86%
6.5	1.9	87%
7.0	1.9	88%
7.5	2.0	88%
8.0	2.1	89%
8.5	2.1	89%

Nonuniform DIF in a MH context

- The `mhodds` command gives a test of **nonuniform DIF** for any particular item, as well.

```
Mantel-Haenszel estimate controlling for sumscore
------------------------------------------------------------------------------
 Odds Ratio     chi2(1)         P>chi2          [95% Conf. Interval]
------------------------------------------------------------------------------
   1.605273       12.94         0.0003           1.237327    2.082636
------------------------------------------------------------------------------
Test of homogeneity of ORs (approx): chi2(7)    =     4.69
                                     Pr>chi2    =   0.6973
```

$(k - 2\ df)$

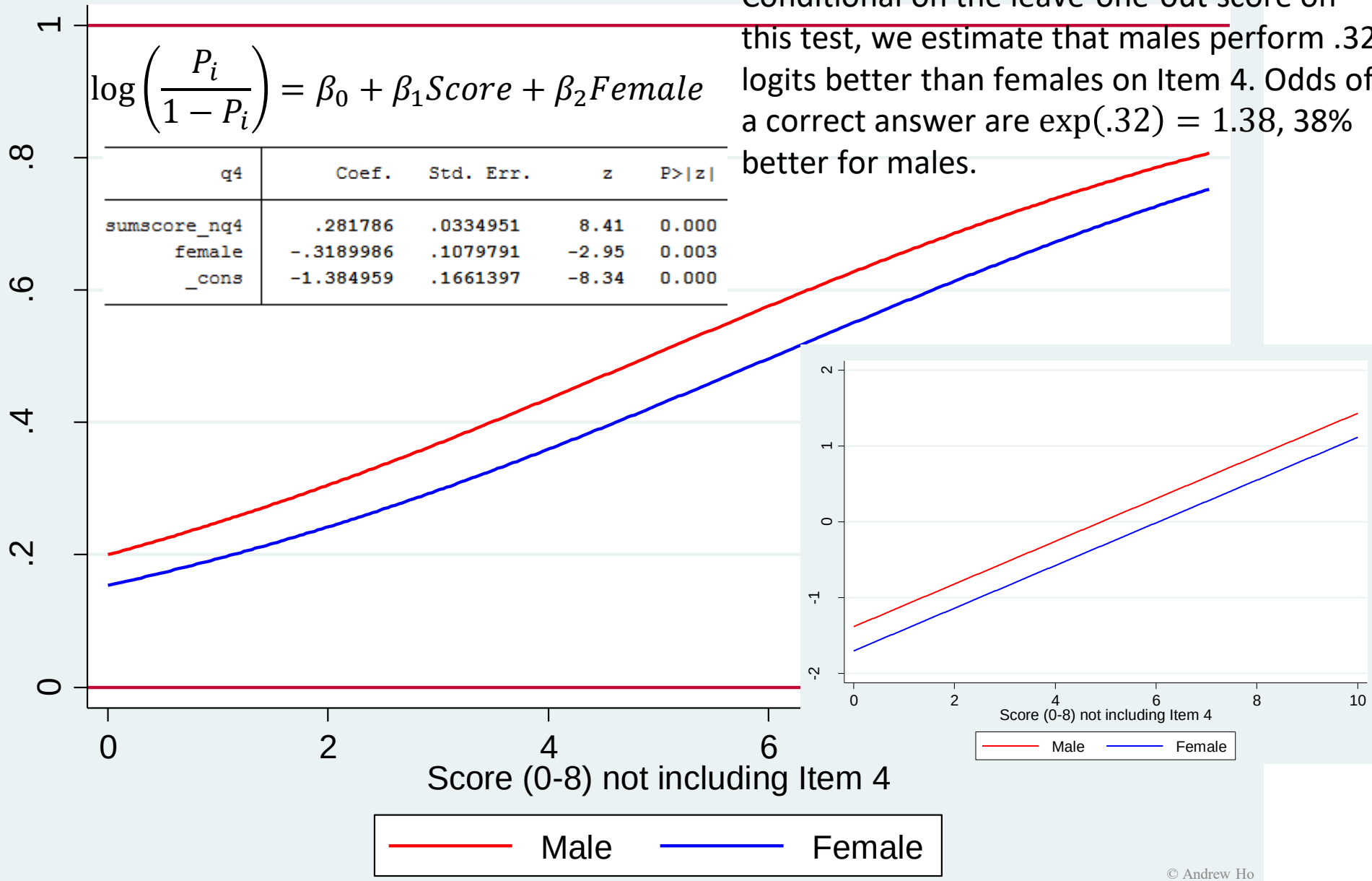
- We conclude that DIF is uniform in favor of females for Item 1, conditional on the sum score on the test.
- If it were nonuniform, use graphs to describe the direction.

Logistic Regression Approaches (Item 4) - Uniform

$$\log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 Score + \beta_2 Female$$

q4	Coef.	Std. Err.	z	P>|z|
sumscore_nq4	.281786	.0334951	8.41	0.000
female	-.3189986	.1079791	-2.95	0.003
_cons	-1.384959	.1661397	-8.34	0.000

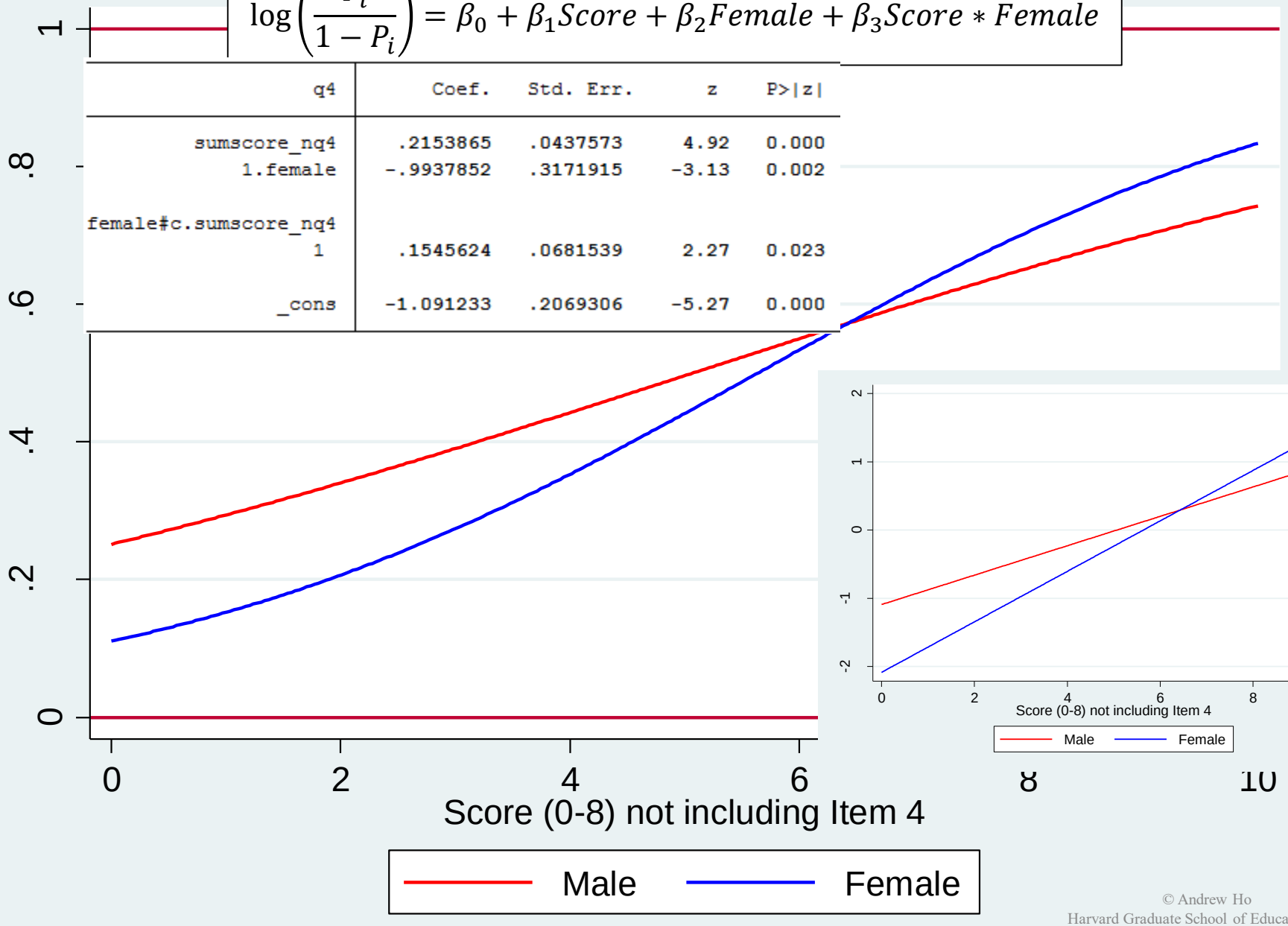
Conditional on the leave-one-out score on this test, we estimate that males perform .32 logits better than females on Item 4. Odds of a correct answer are $\exp(.32) = 1.38$, 38% better for males.



Logistic Regression Approaches (Item 4) - Nonuniform

$$\log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 Score + \beta_2 Female + \beta_3 Score * Female$$

q4	Coef.	Std. Err.	z	P>|z|
sumscore_nq4	.2153865	.0437573	4.92	0.000
1.female	-.9937852	.3171915	-3.13	0.002
female#c.sumscore_nq4				
1	.1545624	.0681539	2.27	0.023
_cons	-1.091233	.2069306	-5.27	0.000



IRT Approaches to DIF (See Camilli, 2006)

- Fit a model with different b parameters for groups
  - Conduct a z-test for differences in b parameter estimates.
  - Likelihood ratio tests constraining group parameters to be equal (vs. not).
  - Also: Calculate areas between item characteristic curves.

- There are also Structural Equation Modeling approaches to DIF. See Stata examples 22 & 23 for continuous (not binary) responses, here.

From DIF to Fairness (See Camilli, 2006)

- Fairness in treatment; fairness as lack of measurement bias, fairness in access to the construct, fairness as validity of score interpretations for intended uses.

- Sensitivity Review - From Bond, Moss, and Carr (1996), "a generic term for a set of procedures for ensuring 1) that stimulus materials used in assessment reflect the diversity in our society and the diversity of contributions to our culture, and 2) that the assessment stimuli are free of wording and/or situations that are sexist, ethnically insensitive, stereotypic, or otherwise offensive to subgroups of the population."

- AERA/APA/NCME Standard 3.2, "Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics."

You don't need fancy DIF methods (Wainer, 2010)

IV. Enough, for Now

Earlier, I suggested that we already had enough psychometrics for current purposes, and efforts in other directions would be more likely to bear fruit. This does not mean that no one should work in these areas, but only that the primary focus of the field should, in my view, shift in other directions. In addition, I most expressly do not mean that we should not apply these methods to current problems, only that further research into their expansion and continued development should be of lower priority.[11] Some areas that could profit from some neglect are:

11. Differential item functioning (DIF)—There are, fundamentally, two approaches developed for studying DIF; observed score methods, of which the Mantel-Haenszel statistic (Holland & Thayer, 1988) is both the best known and the best performing, and model-based methods, in which likelihood-ratio tests provide the probability of DIF (Thissen, Steinberg, & Wainer, 1988, 1993). These two approaches are more than enough to suit virtually any occasion. Journal editors I have spoken with admitted to the same feeling I had as an editor when a new submission arrived on yet another simulation showing the sensitivity of some DIF method to one variation or another of parameter distributions. Enough already!

Learning Objectives for Part II

1. **Generalizability Theory**:
   - *How can we describe and improve the precision of teacher evaluation scores?*

2. **Differential Item Functioning**:
   - *How can we detect items that are not functioning well across countries?*

3. **Linking:**
   - *How can we compare scores across different tests?*

A linking example (Reardon, Kalogrides, Ho, 2017)

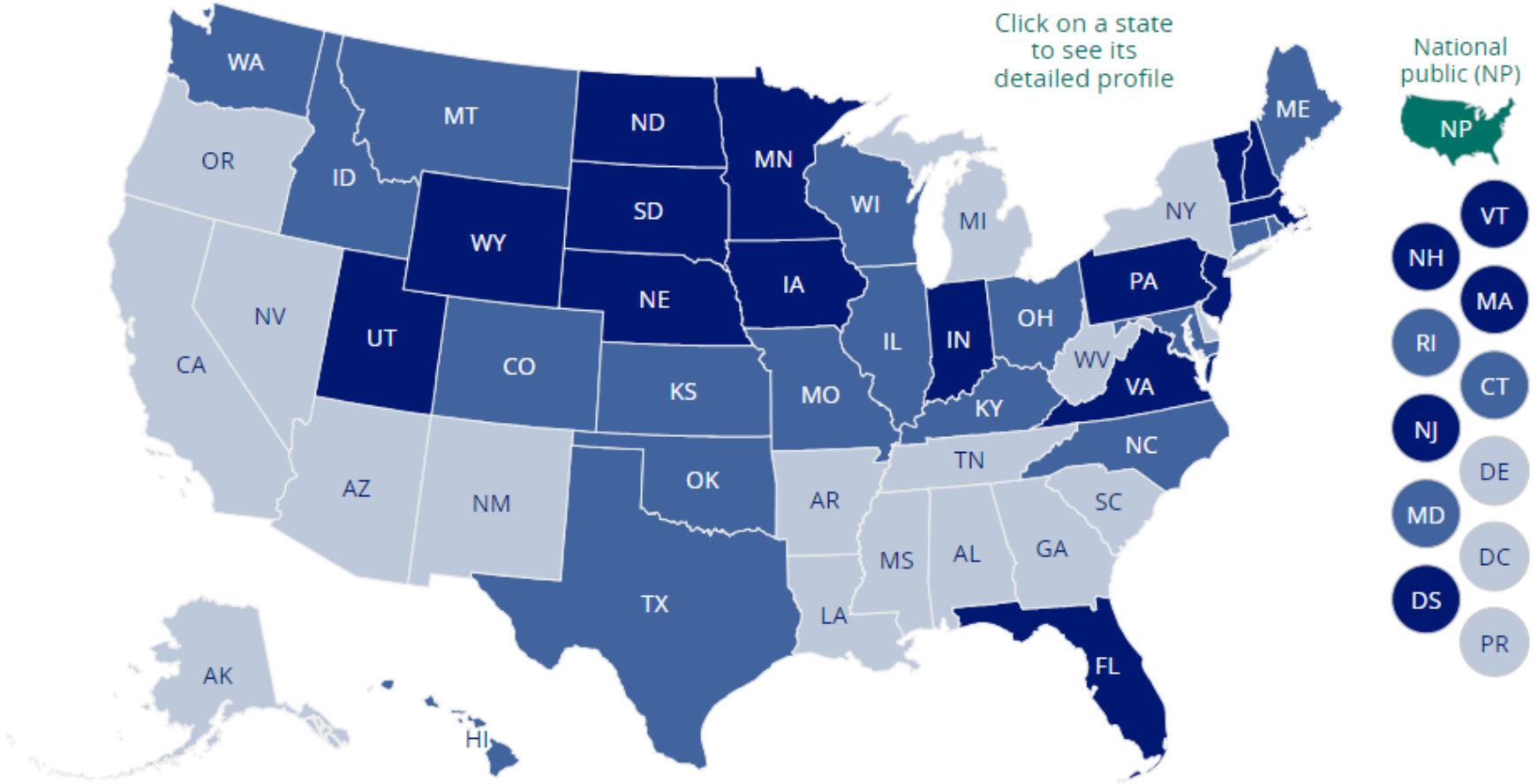
- How can we compare the **average academic performance** of **districts in different states**?

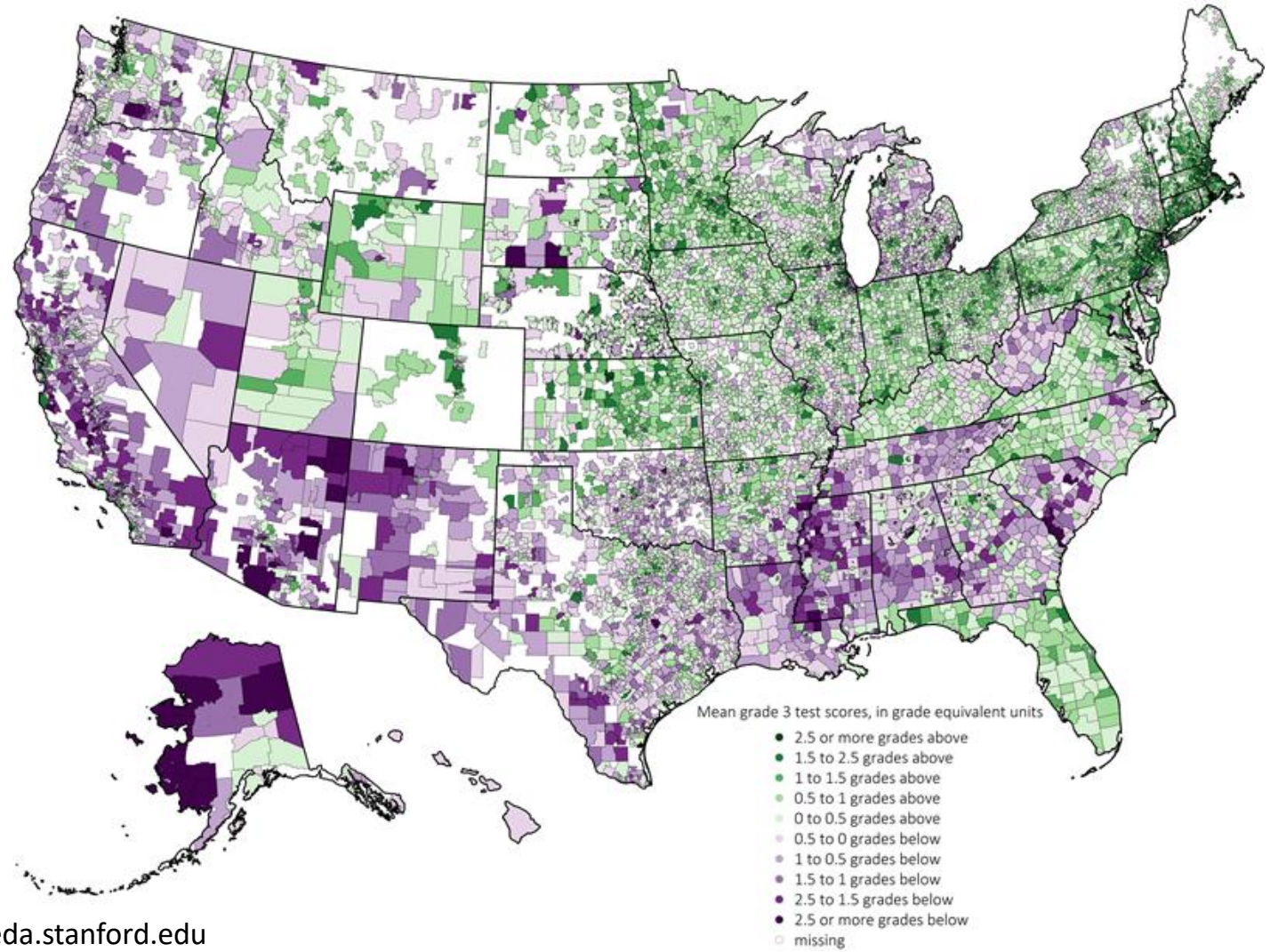
- How can we compare the **average academic progress** of **districts in different states**?

National Assessment of Educational Progress (NAEP)

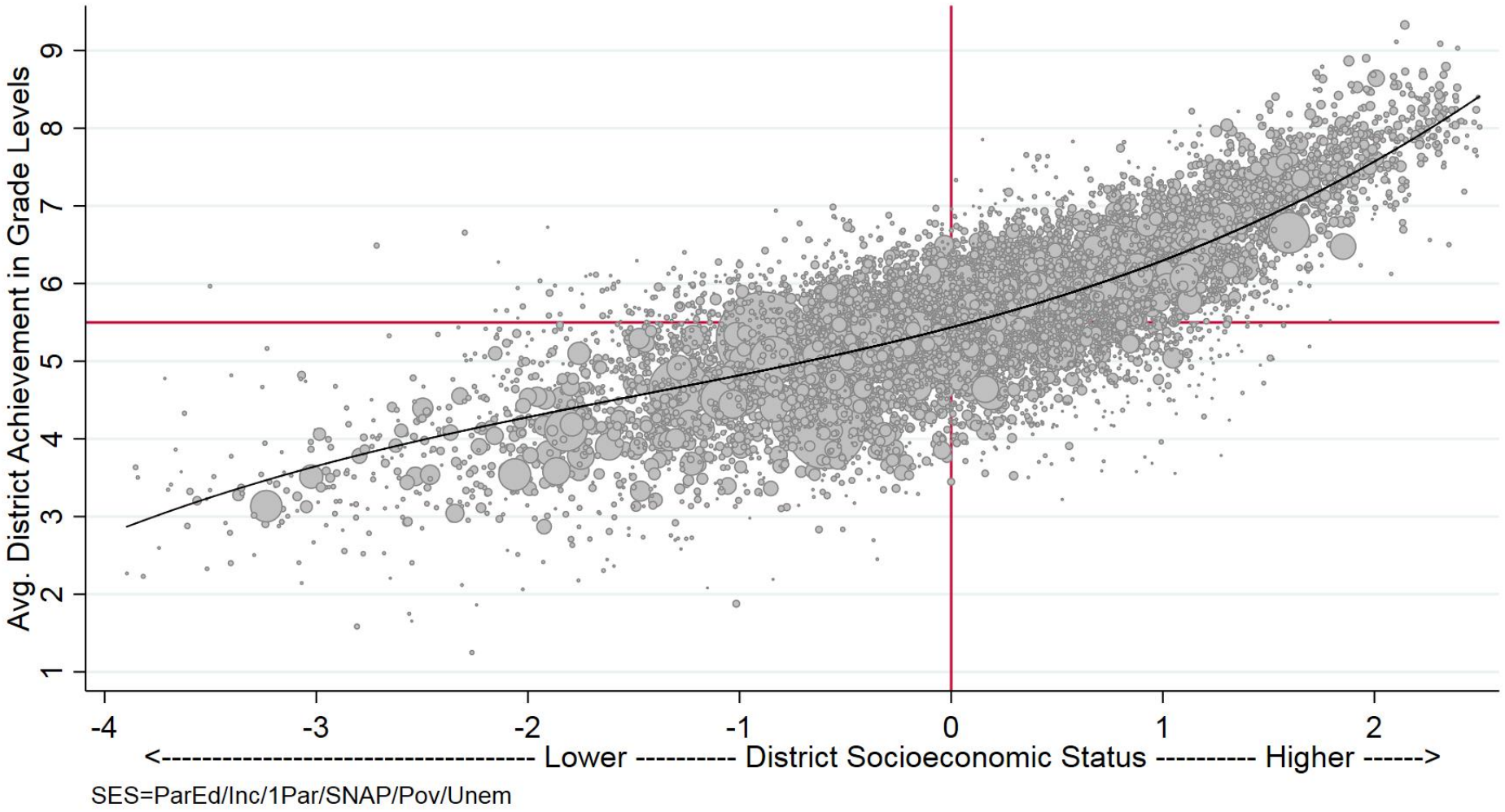
Mathematics, grade 4

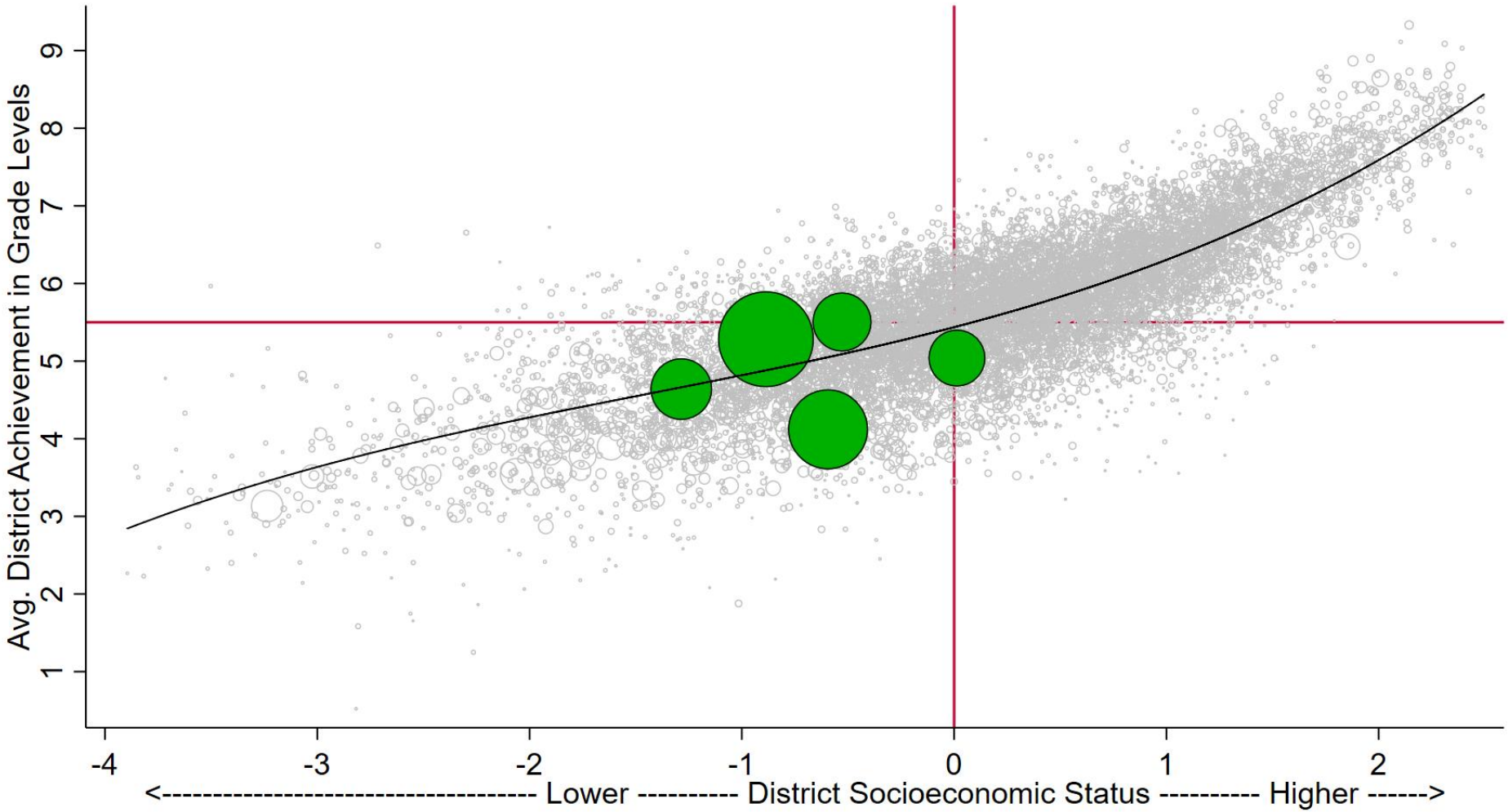
Difference in average scale scores between all jurisdictions and National public, for All students [TOTAL] , 2017





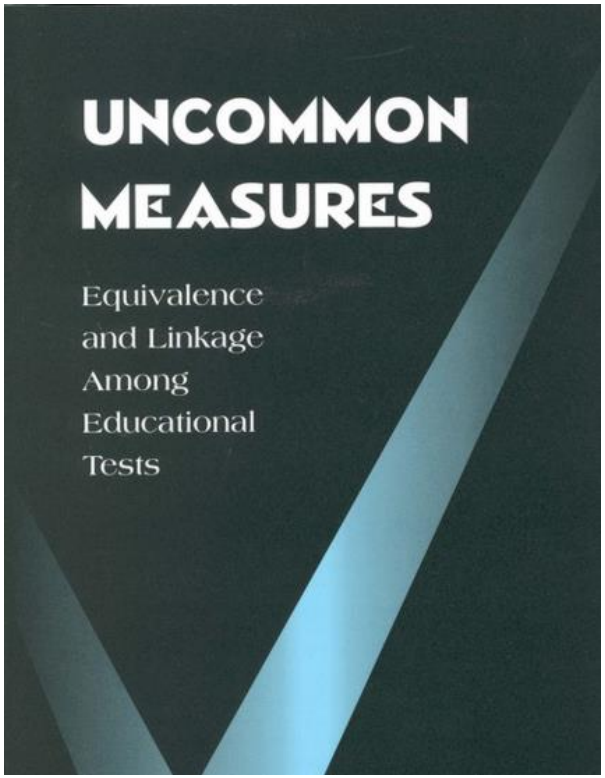
Source: seda.stanford.edu


Avg. District Achievement in Grade Levels
1
2
3
4
5
6
7
8
9
-4
-3
-2
-1
0
1
2
<---------- Lower ---------- District Socioeconomic Status ---------- Higher ------>
SES=ParEd/Inc/1Par/SNAP/Pov/Unem


Avg. District Achievement in Grade Levels
1
2
3
4
5
6
7
8
9
-4
-3
-2
-1
0
1
2
<---------- Lower ---------- District Socioeconomic Status ---------- Higher ------>

How are these cross-state "linkages" valid?!

- Hanushek and Woessman (2012) link international assessments to the NAEP scale.
- Bandeira de Mello et al. (2011, 2013, 2015) map state proficiency standards to the NAEP scale.
- Kolen and Brennan (2014) review linking methods.
- In the NRC Report, *Uncommon Measures*, Feuer et al. (1999) recommend against a NAEP as a common national measure for student-level assessment.
- We link district distributions to the NAEP scale to support district-level policy analysis.
- We will treat the issue empirically, using a series of validation checks. How?



When the essential counterfactual question can be answered for at least some of the target units.

What do we need to link test score scales?

Common persons

Population	Sample	X	Y
P	1	✓	✓

Common populations

Population	Sample	X	Y
P	1	✓	
P	2		✓

Common items

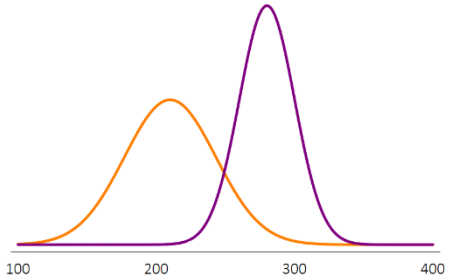
Population	Sample	X	A	Y
P	1	✓	✓	
Q	2		✓	✓

Holland and Dorans (2006)

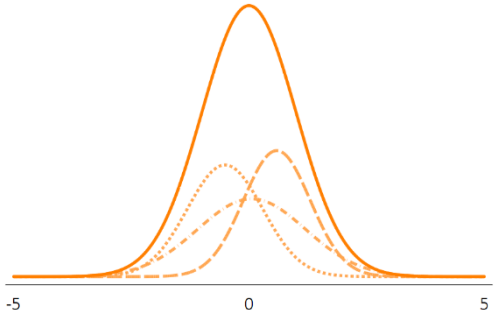
We use common-population linking: NAEP provides state estimates

Notation

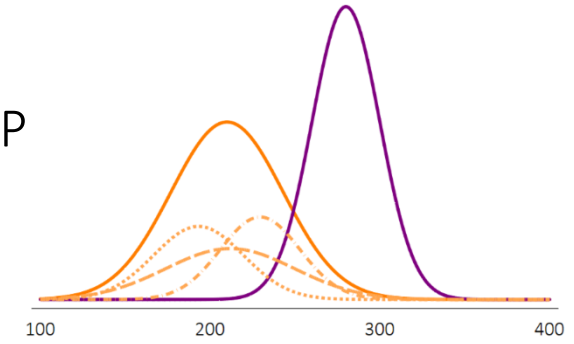
Given each state's NAEP mean and SD, $\hat{\mu}^{\mathbf{naep}}_{sygb}$ and $\hat{\sigma}^{\mathbf{naep}}_{sygb}$, interpolating for even years y and grades $g \in \{3, 5, 6, 7\}$.



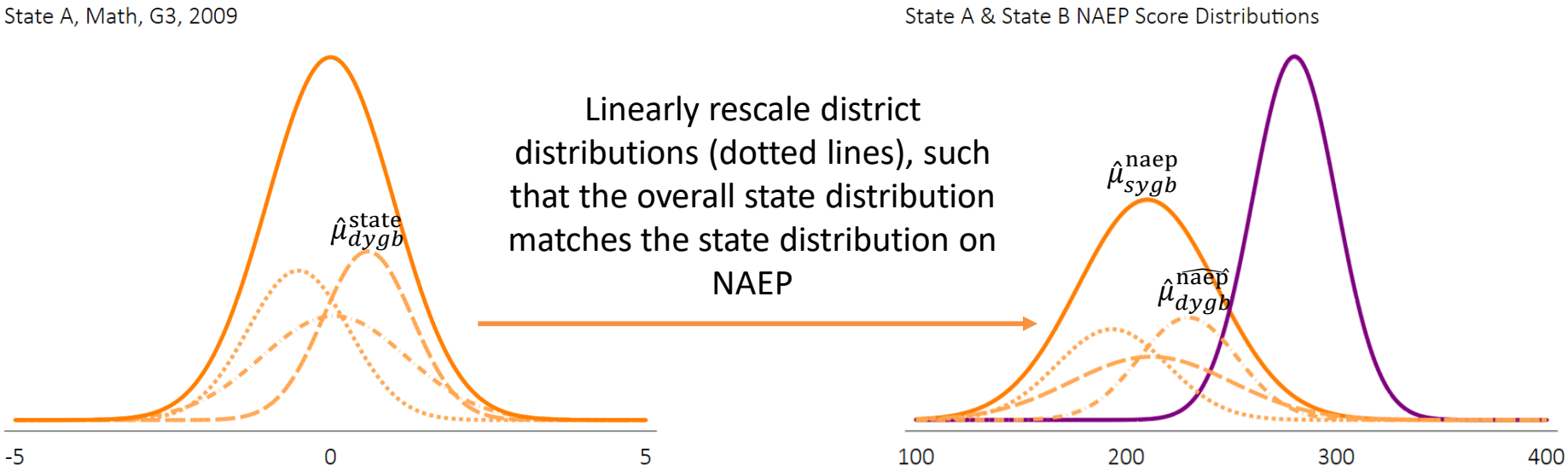
And each district's relative mean and SD in its own state test, on a standardized scale, $\hat{\mu}^{\mathbf{state}}_{dygb}$ and $\hat{\sigma}^{\mathbf{state}}_{dygb}$:



Estimate each district's absolute position on the NAEP scale, $\hat{\mu}^{\widehat{\mathbf{naep}}}_{dygb}$ and $\hat{\sigma}^{\widehat{\mathbf{naep}}}_{dygb}$.



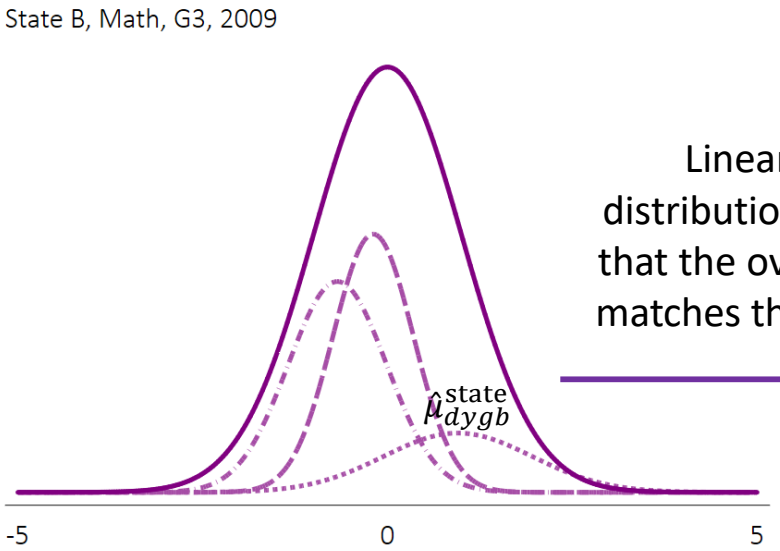
Linear Linking



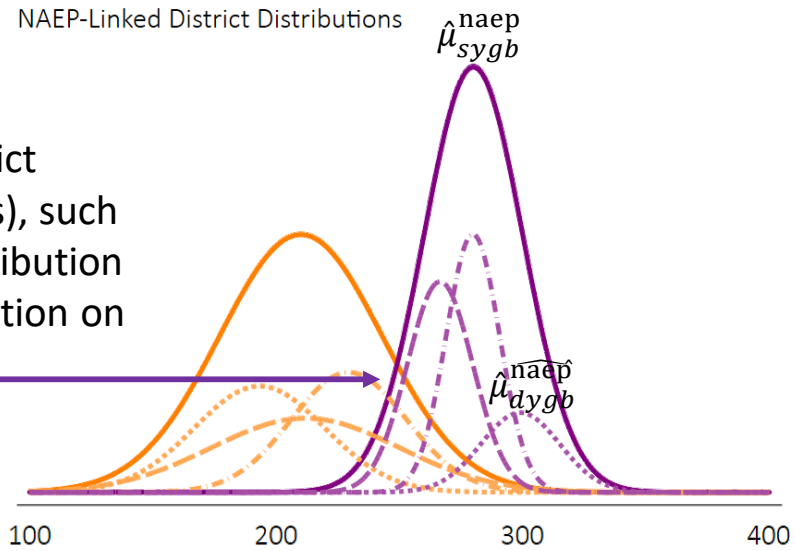
State A standardized distribution (Solid Line):
Mean = 0, SD = 1
District distributions (Dotted Lines)

State A NAEP Distribution (Orange Line):
Mean = 210, SD = 33
District NAEP-Linked Distributions (Dotted Lines)

Linear Linking



Linearly rescale district distributions (dotted lines), such that the overall state distribution matches the state distribution on NAEP

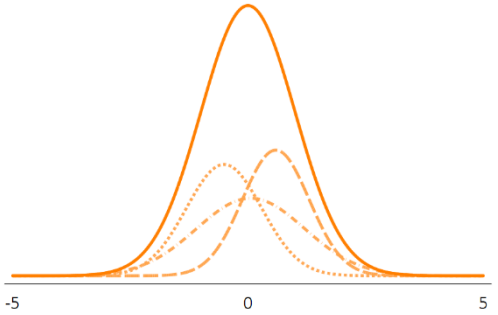


State B standardized distribution (Solid Line): Mean = 0, SD = 1
District Distributions (Dotted Lines)

State B NAEP Distribution (Purple Line): Mean = 280, SD = 20
District NAEP-Linked Distributions (Dotted Lines)

How can we correct standardized group means for imperfect reliability, $\hat{\rho}_{sygb}^{\text{state}}$?

Reliability $\rho = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_e^2}{\sigma_X^2}$.



Observed SDs of 1 overestimate "true" SDs of $\sqrt{\rho}$, so a district's mean should be: $\hat{\mu}_{dygb}^{\text{state}} / \sqrt{\hat{\rho}_{sygb}^{\text{state}}}$

Mean linkage: $\widehat{\hat{\mu}_{dygb}^{\text{naep}}} = \hat{\mu}_{sygb}^{\text{naep}} + \frac{\hat{\mu}_{dygb}^{\text{state}}}{\sqrt{\hat{\rho}_{sygb}^{\text{state}}}} * \hat{\sigma}_{sygb}^{\text{naep}}$

How can we correct standardized group SDs for imperfect reliability, $\hat{\rho}_{sygb}^{\text{state}}$?

If observed SDs are 1, error variance is:

$$\sigma_e^2 = 1 - \rho.$$

Decompose a district's observed variance:

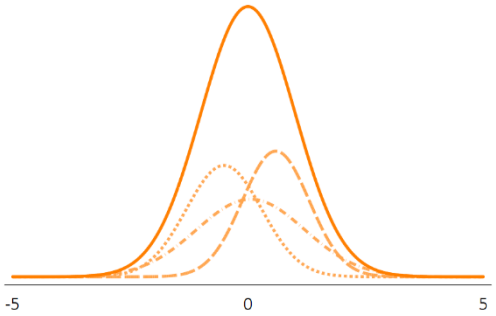
$$\sigma_d^2 = \sigma_{T_d}^2 + \sigma_e^2$$

So the true district SD on a standardized scale is:

$$\sigma_{T_d} = \sqrt{\sigma_d^2 - (1 - \rho)}$$

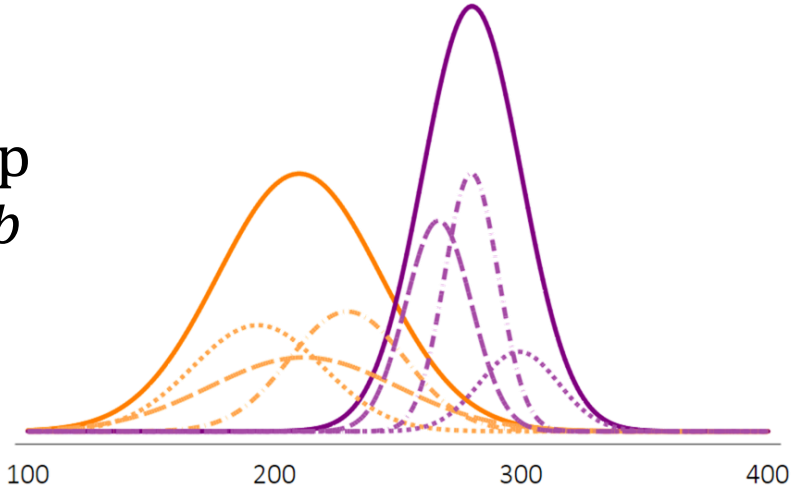
And the true district SD in terms of true state SD units, $\sqrt{\rho}$, is:

$$\sigma_{T_d} = \sqrt{\frac{\sigma_d^2 - (1 - \rho)}{\rho}}$$



Reliability-adjusted linear linking equations:

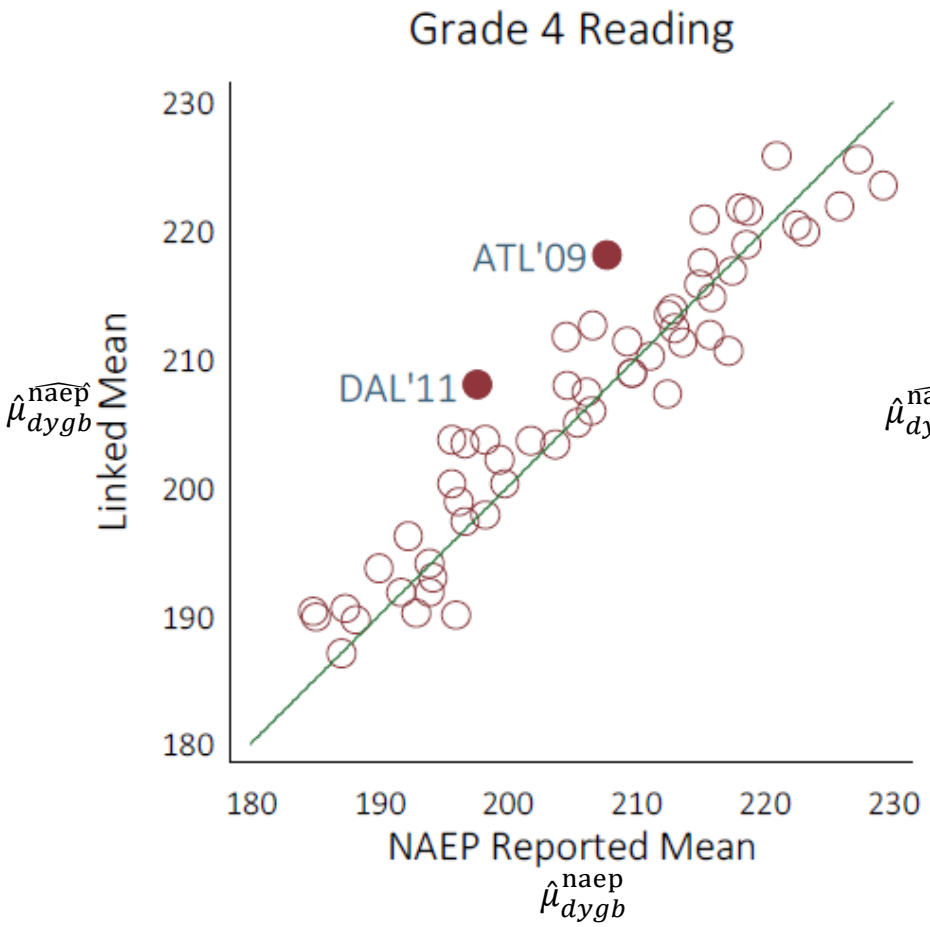
$$\widehat{\hat{\mu}^{\text{naep}}_{dygb}} = \hat{\mu}^{\text{naep}}_{sygb} + \frac{\hat{\mu}^{\text{state}}_{dygb}}{\sqrt{\hat{\rho}^{\text{state}}_{sygb}}} * \hat{\sigma}^{\text{naep}}_{sygb}$$

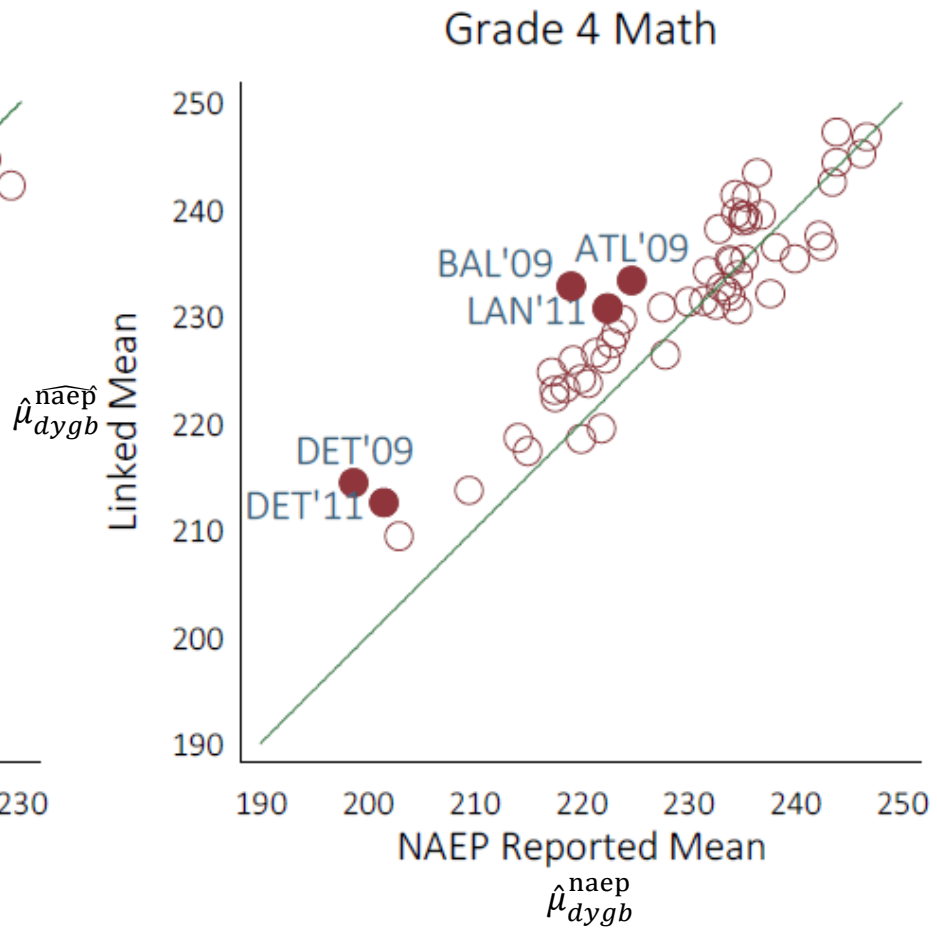


$$\widehat{\hat{\sigma}^{\text{naep}}_{dygb}} = \left[\frac{\left(\hat{\sigma}^{\text{state}}_{dygb}\right)^2 + \hat{\rho}^{\text{state}}_{sygb} - 1}{\hat{\rho}^{\text{state}}_{sygb}}\right]^{1/2} \cdot \hat{\sigma}^{\text{naep}}_{sygb}$$

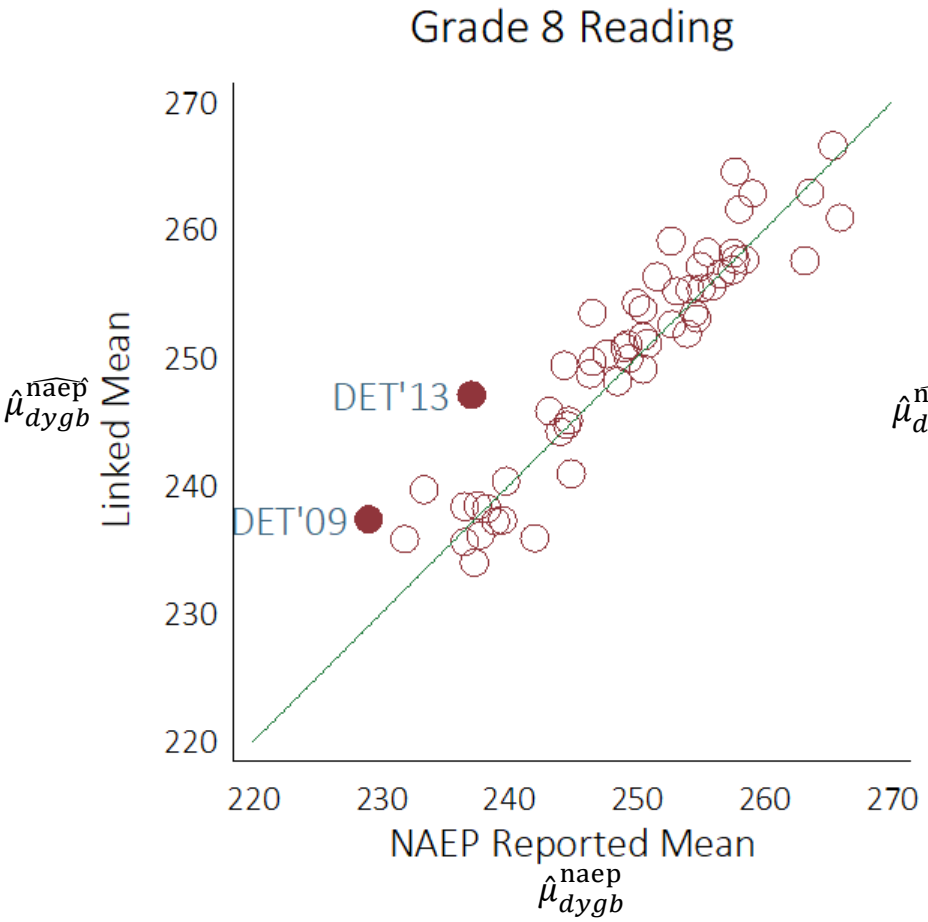
How can we validate the linking?

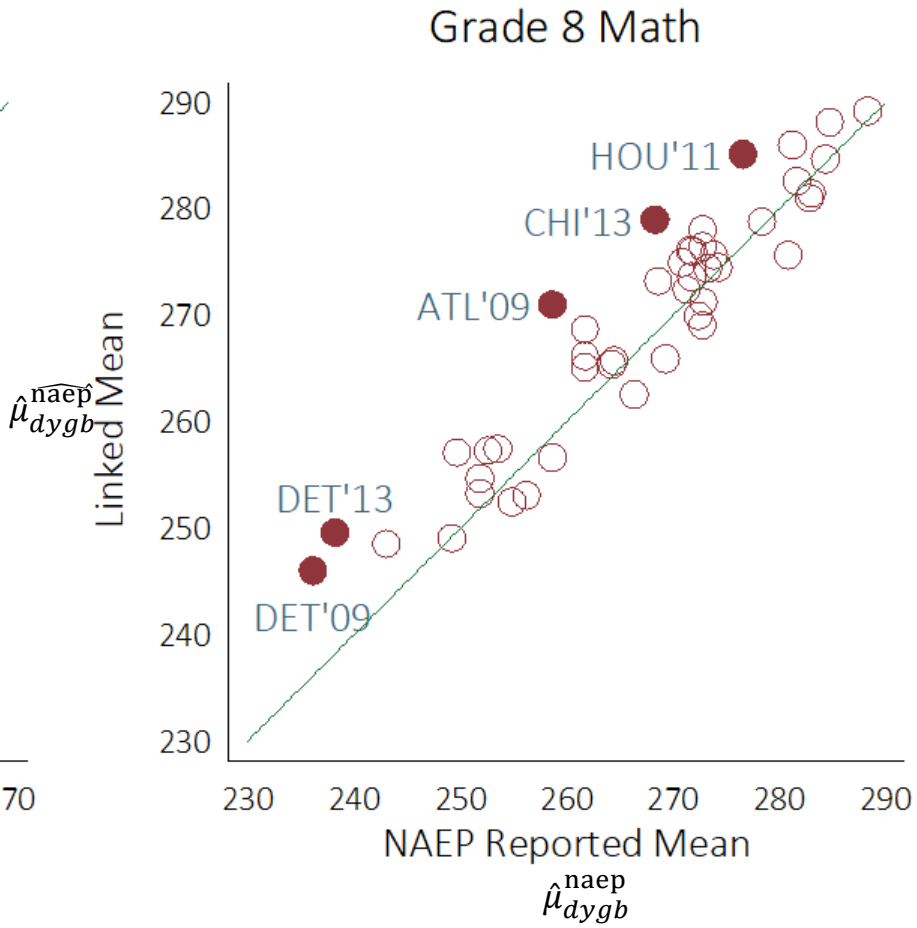
How can we validate the linking? G4.





How can we validate the linking? G8.





How can we precision-adjust estimates of linked-vs.-true bias, RMSE, and correlations?

$$\hat{\mu}_{idygb} = \alpha_{0dygb}\left(\hat{\mu}^{\widehat{\text{naep}}}_{idygb}\right) + \alpha_{1dygb}\left(\hat{\mu}^{\text{naep}}_{idygb}\right) + e_{idygb}$$

$$\alpha_{0dygb} = \beta_{00} + u_{0dygb}$$

$$\alpha_{1dygb} = \beta_{10} + u_{1dygb}$$

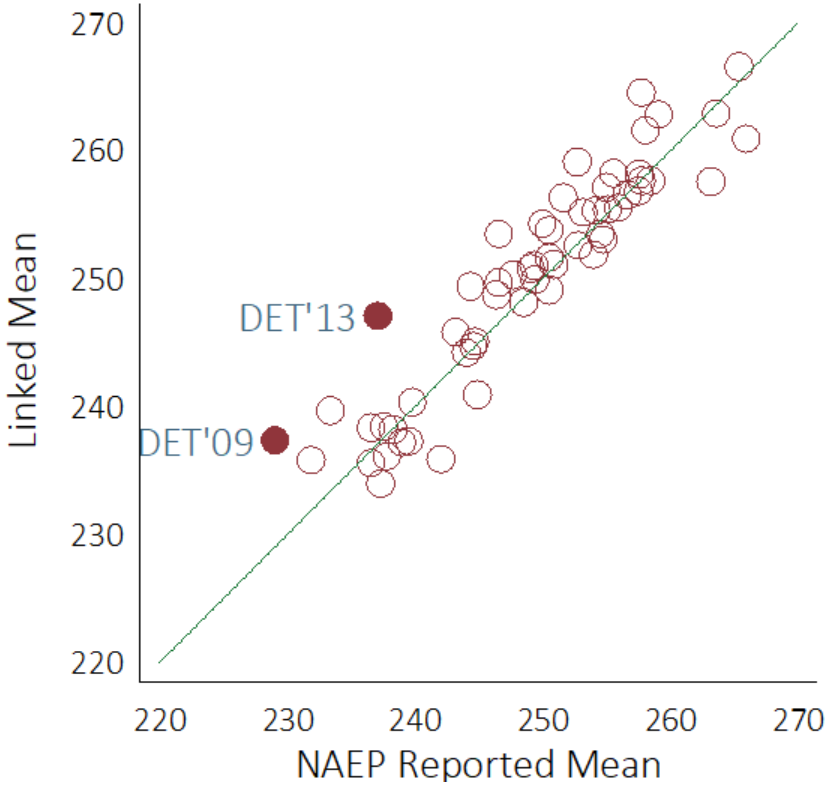
$$e_{idygb} \sim N\left(0, \omega^2_{idygb}\right);\ \boldsymbol{u}_{dygb} \sim MVN(0, \boldsymbol{\tau}^2)$$

$$\boldsymbol{\tau}^2 = \begin{bmatrix} \tau^2_{00} & \tau^2_{01} \\ \tau^2_{01} & \tau^2_{11} \end{bmatrix}$$

Bias: $\hat{B} = \hat{\beta}_{00} - \hat{\beta}_{10}$

$$\widehat{RMSE} = \left[\hat{B}^2 + \mathbf{b}\hat{\boldsymbol{\tau}}^2\mathbf{b}'\right]^{1/2}$$

Correlation: $\hat{r} = \dfrac{\hat{\tau}^2_{01}}{\hat{\tau}_{00}\hat{\tau}_{11}}$



Linking error in cross-state district comparisons?

				Recovery		
Subject	Grade	Year	n	RMSE	Bias	Correlation
Reading	4	2009	17	3.95	2.12	0.96
		2011	20	3.69	1.25	0.96
		2013	20	2.62	0.20	0.98
	8	2009	17	2.92	1.12	0.95
		2011	20	2.20	0.63	0.97
		2013	20	3.62	1.67	0.93
Math	4	2009	17	6.09	4.10	0.93
		2011	20	4.97	2.60	0.94
		2013	20	3.60	1.46	0.95
	8	2009	14	5.21	3.40	0.95
		2011	17	3.77	2.09	0.96
		2013	14	4.54	1.47	0.94
Average		2009	14-17	4.70	2.69	0.95
		2011	17-20	3.79	1.64	0.96
		2013	14-20	3.66	1.20	0.95
		Reading	17-20	3.23	1.17	0.96
		Math	14-20	4.77	2.52	0.95
		All	14-20	4.07	1.84	0.95

Linking Caveats

When comparing districts across states, what are five reasons why a high-scoring SEDA linked district may not actually have a high NAEP score, had one been reported?

1. Student sampling.
   - That district samples its higher achieving students, (a) *disproportionately over other districts in the state;* and, (b) *disproportionately on the state test over NAEP*
2. Student motivation.
3. Tested Content.
4. Inflation.
5. Cheating.

NOTE: All explanations involve a difference (between districts) in differences (between tests)

Stanford cepa | Center for Education Policy Analysis

CEPA Working Paper No. 16-09

Linking U.S. School District Test Score Distributions to a Common Scale

Author/s: Sean F. Reardon , Demetra Kalogrides , Andrew Ho

Year of Publication: 2017

There is no comprehensive database of U.S. district-level test scores that is comparable across states. We describe and evaluate a method for constructing such a database. First, we estimate linear, reliability-adjusted linking transformations from state test score scales to the scale of the National Assessment of Educational Progress (NAEP). We then develop and implement direct and indirect validation checks for linking assumptions. We conclude that the linking method is accurate enough to be used in analyses of national variation in district achievement, but that the small amount of linking error in the methods renders fine-grained distinctions among districts in different states invalid. Finally, we describe several different methods of scaling and pooling the linked scores to support a range of secondary analyses and interpretations.

https://cepa.stanford.edu/sites/default/files/wp16-09-v201706.pdf

Learning Objectives for Part II

1. **Generalizability Theory**:
   - *How can we describe and improve the precision of teacher evaluation scores?*

2. **Differential Item Functioning**:
   - *How can we detect items that are not functioning well across countries?*

3. **Linking:**
   - *How can we compare scores across different tests?*