

Large-Scale Assessments Methodology and Implications for Data Analyses

Ting Zhang

American Institutes for Research

AERA | April 2024

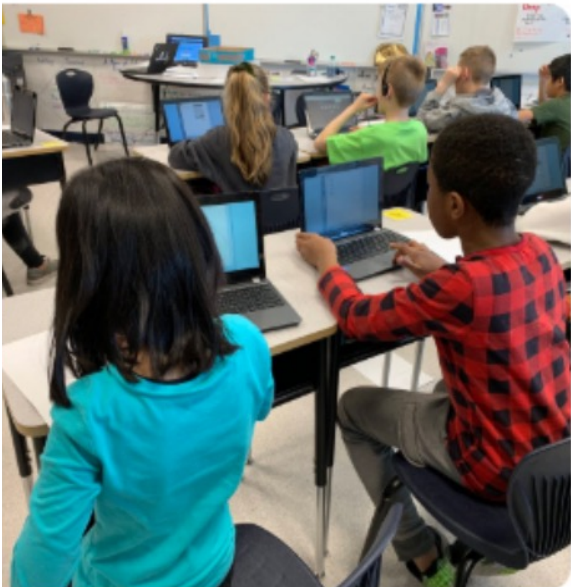
Agenda

- Overview of large-scale assessments (LSA)
- Sampling design
 - Complex sample design
 - Implication for variance estimation and sampling weights
- Testing design
 - Metrix sampling design
 - Introduction to plausible values
 - Implication for proficiency and variance estimation



What are educational large-scale assessments (LSA)?

Tests that focus on measuring and monitoring what populations know and can do in academic subject areas



- Populations are usually certain ages/grades in cities, states, countries
- Subpopulation measurement (gender, SES, race/ethnicity) is also prioritized
- Academic subject areas include mathematics, science, reading, social studies, computer literacy, etc.
- Measure contextual factors associated with achievements

Uses of LSAs

Provide information for achievement comparisons between educational systems or jurisdictions

- *How does the performance in one country compare with that of other countries?*

Examine trends in achievement

- *How does one country's achievement increase or decrease over time?*

Improve education by informing policy, research, and practices

- *What factors are associated with educational achievement? What can we learn from others about what works (and what doesn't)? What could be adopted by or adapted?*

Why Specialized Software Programs are Needed for LSA Data Analysis?

Study Designs

- Complex sample design (multi-stage, clustered sample design, sampling weights, sampling variance)
- Complex assessment design (matrix sampling design, use of Item Response Theory (IRT) and Plausible Values (PVs), and measurement variance)



Sample Design, Sampling Variance and Weights

Why Do We Use Samples?

Impossible to test everyone on everything

- Too many people
- Too many items
- Too expensive

Not necessary to test everyone on everything, e.g.,

- Blood sample
- Soup sample



Source: Sabine Meinck, Ph.D., IEA, Design of International Large-Scale Assessments and Implications for Multilevel Modeling

You're probably familiar with simple random sampling (SRS)



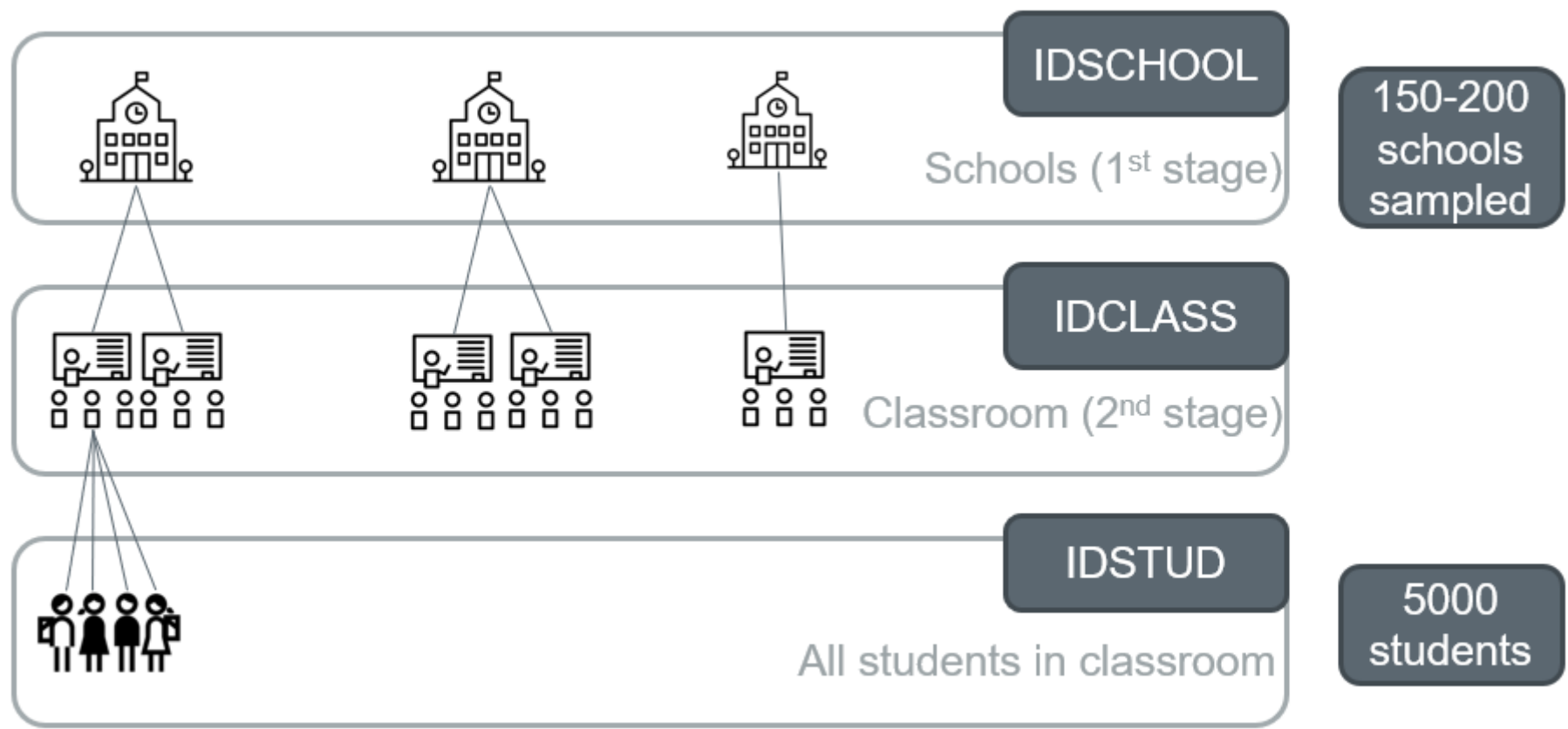
- Each person has an equal probability of selection
- But, to conduct a study with nationally representative samples, SRS not feasible:
 - Time!
 - Cost!

Complex sample designs

Why “complex”?

- Multiple stages of stratified sampling
- Homogenous clusters are sampled
- Selection probabilities differ for different sampling units

Multiple sampling stages (TIMSS)



What are the implications of multistage cluster sampling?



Cluster effect: Individuals within classrooms/schools tend to be more similar than individuals between schools

What are the implications of multistage cluster sampling?

- In general, the sampling variance of a clustered sample tends to be larger than the sampling variance of a simple random sample of the same size.
- In studies using a complex sampling design, standard errors tend to get larger, partly due to the sampling variance.
- Variances are essential in statistical tests of significance
 - Biased variances could make differences between scores appear significant when in fact they are not

How to handle clustering:

Three recommended methods

- Replication methods. TIMSS and NAEP use jackknife repeated replication
- Taylor series approximations
- Hierarchical linear models

Sampling schools and students:

Unequal probabilities of selection

- Schools are sampled with **probability proportional to their size** from the list of all schools in the population
 - Larger schools are more likely to be selected than small ones
 - Students in large schools have a lower probability of selection than classes in smaller schools
 - Result: the overall probability of selection of students is more similar across different size schools

When we sample with unequal probabilities of selection, aren't we introduce bias into our analytical results?

Sample Weights

- To account for unequal probabilities of selection, sampling weights should be used in all statistical calculations where inferences are made to populations
 - Otherwise, population estimates (e.g., means, percentages) will be biased
 - Most standard software have ways to specify weights

TIMSS Student Weights

Final student weights = school-level sampling probability * student-level sampling probability * non-participation adjustments

- Students are assigned sampling weights to adjust for over- or under-representation of particular groups in the final sample
- **Student weight is the inverse of the probability of selection**
- Students with higher weight values are representing more people in the target population
- Use of sampling weights is necessary for computation of sound, representative estimates
- Weights adjust for nonparticipation
- **Sum of the overall student weights equals the number of students in the target population**

Probabilities and Weights

- **Student weight is the inverse of the probability of selection**
- Suppose a school has a probability of selection of 0.1 and each student within a school has a probability of selection of 0.2. What is the students' probability of selection?

School prob

Student within school prob

Joint Prob

1/10

2/10

2/100 or 1/50

- Student weight = inverse probability

$$50/1 = 50$$

Q: If a school has a 0.2 probability of selection and once school is selected the student has a 0.05 probability of selection, what is the students weight?

A. 100

B. 200

C. 20

D. 500

Implications for Large-Scale Data Analysts

- Many statistical software packages assume the data being analyzed come from a simple random sample with independent observations and equal selection probability
- LSAs clustering of observations in schools and classes prevents them from being independent
 - Need to use the JK, Taylor series or HLM method for sampling variance
- Complex sampling leads to unequal selection probability
 - Weights need to be applied for unbiased estimates
- Statistical software exists for “complex samples”
 - NCES’s EdSurvey and Dire!

Questions?

Testing Design and Plausible Values

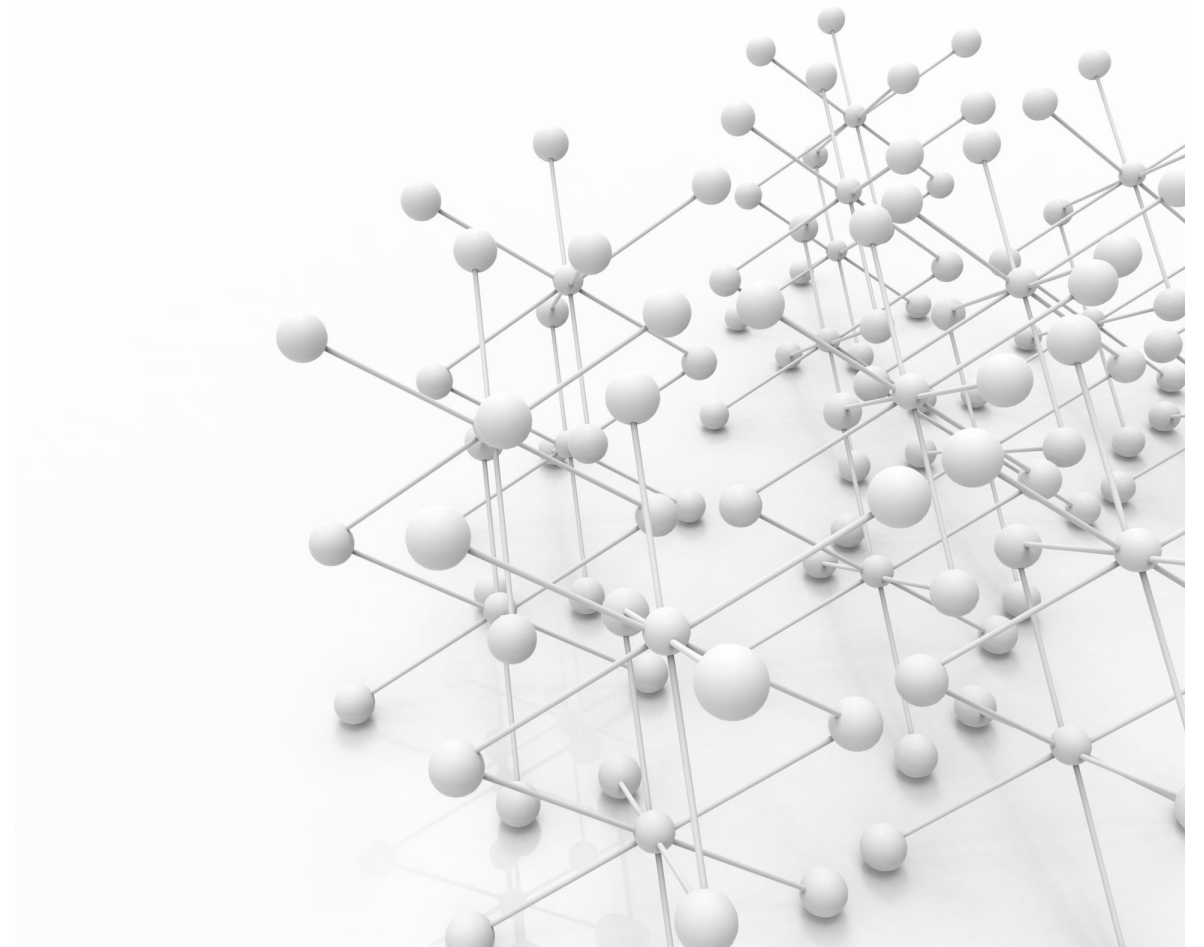
Ting Zhang, Ph.D.

Senior Researcher

AERA | April 2024

Agenda

1. Intro
2. Testing Design
3. Plausible Values
4. Implication for data analysis



What are Plausible Values?

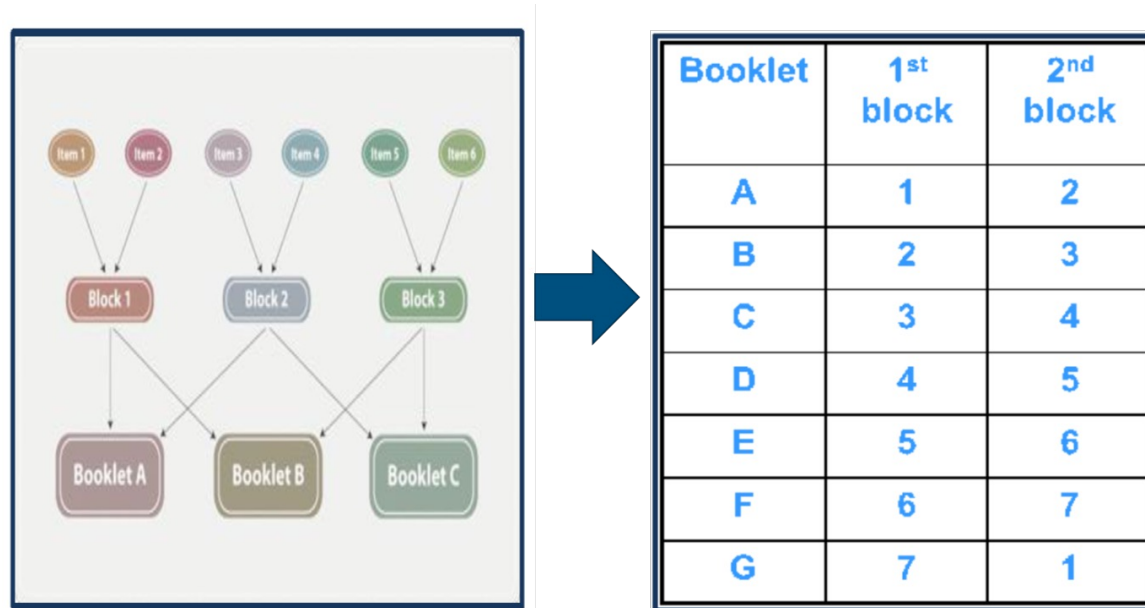
91	MRPS11	612	5	2	C
92	MRPS12	617	5	2	C
93	MRPS13	622	5	2	C
94	MRPS14	627	5	2	C
95	MRPS15	632	5	2	C

Plausible	NAEP	math	value	#1	(num & oper)
Plausible	NAEP	math	value	#2	(num & oper)
Plausible	NAEP	math	value	#3	(num & oper)
Plausible	NAEP	math	value	#4	(num & oper)
Plausible	NAEP	math	value	#5	(num & oper)

- Proficiency estimates for an individual student, drawn at random from a conditional distribution of potential scale scores.
- All available plausible values should be used when calculating summary statistics for groups of students



Why use Plausible Values?



Assessment designs!

- Large-scale assessments such as NAEP use a large item pool of test questions to provide comprehensive coverage of each subject domain.
- To keep the burden of test-taking low and encourage school participation, each student is administered a small number of items.
- At the assessment level, all items are measured.

Advantages and trade-offs of the testing design


Advantages

- Cost efficient and avoids overburdening students and schools
- Achieves broad coverage of the targeted content domain
- Allows sufficiently precise estimates of proficiency distributions of the target population and sub-populations,
 - uses IRT and Multiple Imputations to create student scale scores – plausible values.

Trade-offs

- Each student receives too few test questions to permit estimating an accurate scale score for that student.
- Results in large measurement error and leads to inaccurate inference.

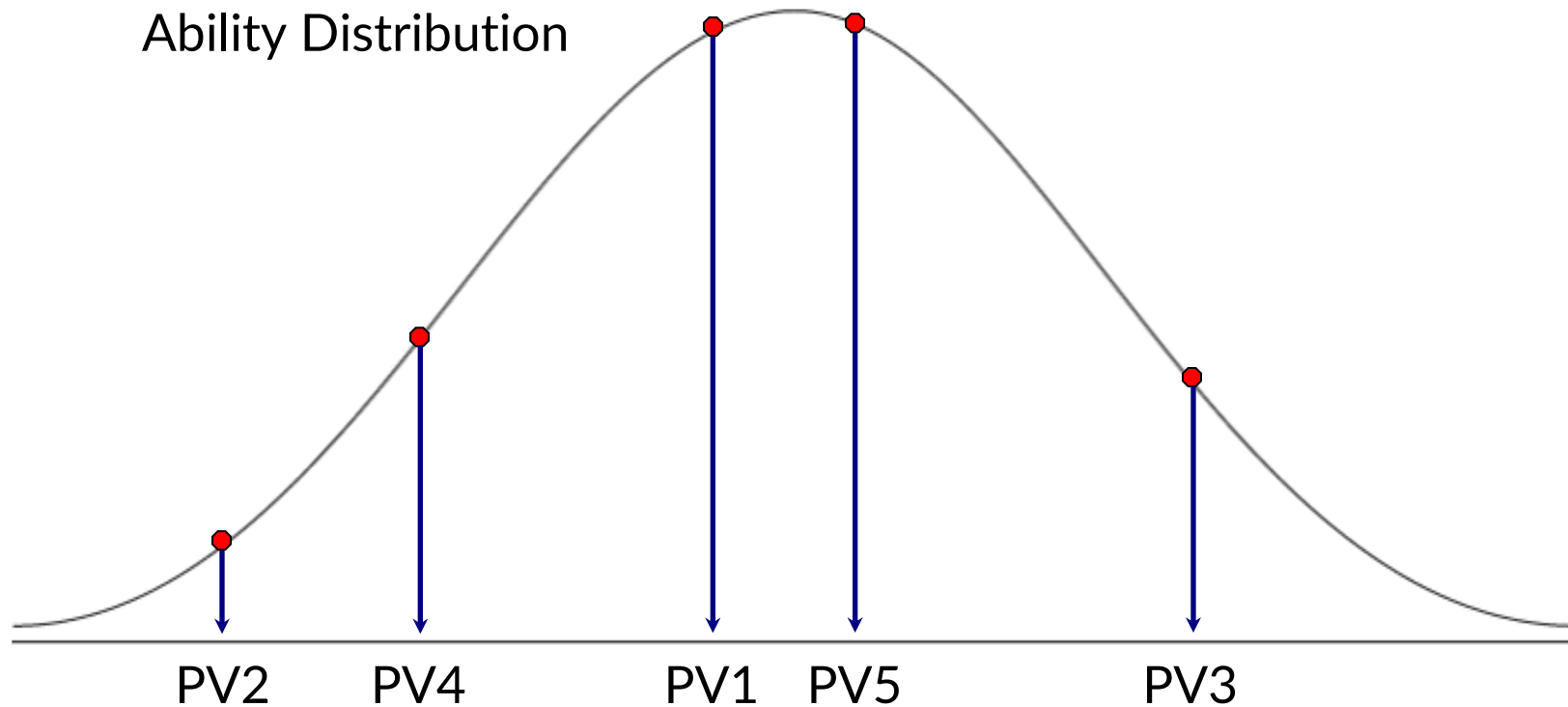
How can an assessment program work without accurate scores for individual students?



Solution: treat the
scale score as
missing data!

- One way of taking the uncertainty associated with the estimates into account, and of obtaining unbiased group-level estimates, is to use multiple imputation to impute what we know about the students and obtain the distribution that represents a student's proficiency.
- Plausible values are based on student responses to the subset of items they receive and available background information (Mislevy, 1991).

Ability Distribution and Plausible Values



How NAEP/TIMSS scores are generated?

(von Davier, Gonzalez & Mislevy 2009)

The first stage

- requires estimating IRT parameters for each cognitive question.

The second stage

- results in latent regressions that imputing scale performance with all information in the student, teacher, and school questionnaires.

The third stage

- combines the previous two stages.

The fourth stage

- draws multiple plausible values from a posterior distribution.

1st stage: Item response theory (IRT)

Estimating IRT parameters for each cognitive question, and a likelihood function for proficiency.

Common IRT models used in large scale assessments

- Dichotomous items: the two- or three-parameter logistics item response model
- Polytomous items: the generalized partial credit model

2nd stage: Conditioning model (population model)

The values of θ are derived from a latent regression equation, referred to as the conditioning model

$$\theta_i = \Gamma' X_i + \varepsilon_i$$

- Where θ_i are the latent distribution that represent a student's proficiency
- Where X_i are the observed responses to survey items
 - In operation, we don't use the raw variables for X , rather we reduce the dimensions of x to principal components which account for 90% of the variance in X
- Γ are the latent regression parameters
- ε_i 's follow multivariate normal distribution with mean zero and variance-covariance matrix Σ



3rd stage: Final model

- Plausible values are drawn from the posterior latent trait given observed responses to test items, x_i , and survey questionnaire items, y_i :

Latent regression

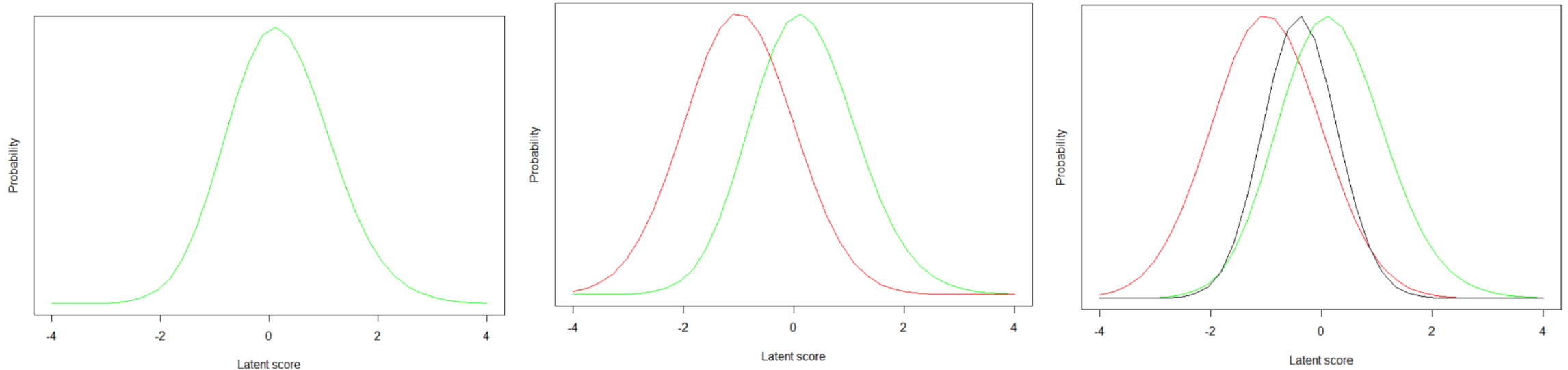
Likelihood function
based on the item
responses

$$f(\theta_i | \mathbf{X}_i, \mathbf{Y}_i, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) \propto \phi(\theta_i; \boldsymbol{\Gamma}'\mathbf{X}_i, \boldsymbol{\Sigma}) \prod f_i(Y_{ij} | \theta_i, \beta_j) \text{ Where}$$

- $\boldsymbol{\beta}$ are the item parameters
- $\boldsymbol{\Gamma}$ are the latent regression parameters
- $\boldsymbol{\Sigma}$ is a covariance matrix
- $\phi(\theta_i; \boldsymbol{\Gamma}'\mathbf{X}_i, \boldsymbol{\Sigma})$ is a normal distribution with mean $\boldsymbol{\Gamma}'\mathbf{X}_i$ and covariance $\boldsymbol{\Sigma}$

Likelihood distribution from the final model

$$f(\theta_i | \mathbf{X}_i, \mathbf{Y}_i, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) \propto \phi(\theta_i; \boldsymbol{\Gamma}'\mathbf{X}_i, \boldsymbol{\Sigma}) \prod f_i(Y_{ij} | \theta_i, \beta_j)$$



green line = student likelihood, red line = prior/conditioning model, **black line** = overall (convolution of both)

The Plausible Values approach

- Draw multiple potential values from the posterior distribution from the 3rd stage final model.
- Rubin's multiple imputation method need to be used to calculate the measurement error (imputation error) and sampling error

How do we analyze Plausible Values?

- Let $t = t(\theta)$ be the population parameter of interest and M be the number of plausible values
- Use each plausible value, $\widehat{\theta}_m$, from a set to evaluate t , yielding \hat{t}_m for $m = 1, \dots, M$
- Estimate $t^* = \sum_{m=1}^M \hat{t}_m / M$

Variance estimation from Plausible Values

- **Variance due to measurement error** (also known as between imputation variance)

$$B_M = \sum_{m=1}^M (\hat{t}_m - t^*)^2 / (M - 1)$$

- Compute the **sampling variance** of \hat{t}_m, U_m using jackknife variance approaches, and average sampling variance, U , across all plausible values

$$U^* = \sum_{m=1}^M U_m / M$$

- **Final estimate of variance of t^* :**

$$V = \left(1 + \frac{1}{M}\right) B_M + U^*$$

measurement variance
+ sampling variance

Takeaways

- When conducting a NAEP analysis that involves plausible values (PVs).

Always

- Use the full set of the PVs
- Apply the appropriate sampling weight(s)
- Calculate correct variance estimation, which usually has two components
 - » Measurement/imputation variance
 - » Sampling variance

Reference

- Mislevy, R. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56, 177–196.
- Rubin, D. (1987). *Multiple imputation for nonresponse in sample surveys*. New York: John Wiley.
- Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, 39(2), 142-151.
- von Davier, M., Gonzalez, E., & Mislevy, R. (2009). Plausible values: What are they and why do we need them? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, 2, 9–36.
- von Davier, M., & Sinharay, S. (2014). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 155–174). Boca Raton, FL: Taylor & Francis.

Ting Zhang

Senior Researcher
tzhang@air.org

AMERICAN INSTITUTES FOR RESEARCH[®] | AIR.ORG

Notice of Trademark: "American Institutes for Research" and "AIR" are registered trademarks. All other brand, product, or company names are trademarks or registered trademarks of their respective owners.

Copyright © 2021 American Institutes for Research[®]. All rights reserved. No part of this publication may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, website display, or other electronic or mechanical methods, without the prior written permission of the American Institutes for Research. For permission requests, please use the Contact Us form on AIR.ORG.