

Where Were NAEP/TIMSS Scores From?

Plausible Values & Direct Estimation Approaches

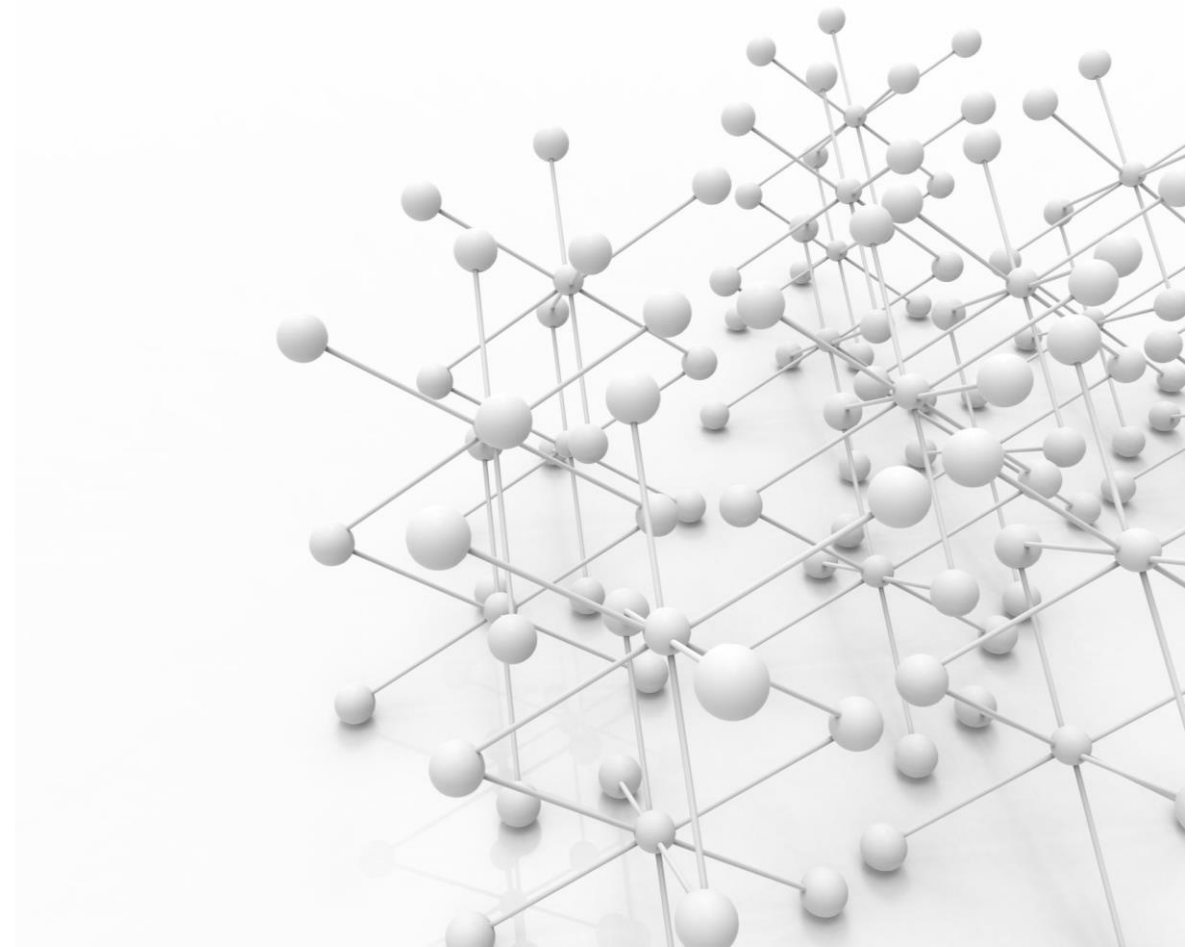
Ting Zhang, Ph.D.

Senior Researcher

NCME | March 2023

Agenda

1. Assessment design
2. Psychometric models for NAEP/TIMSS data
3. The Direct Estimation Approach
4. The Plausible Values Approach



What are Plausible Values?

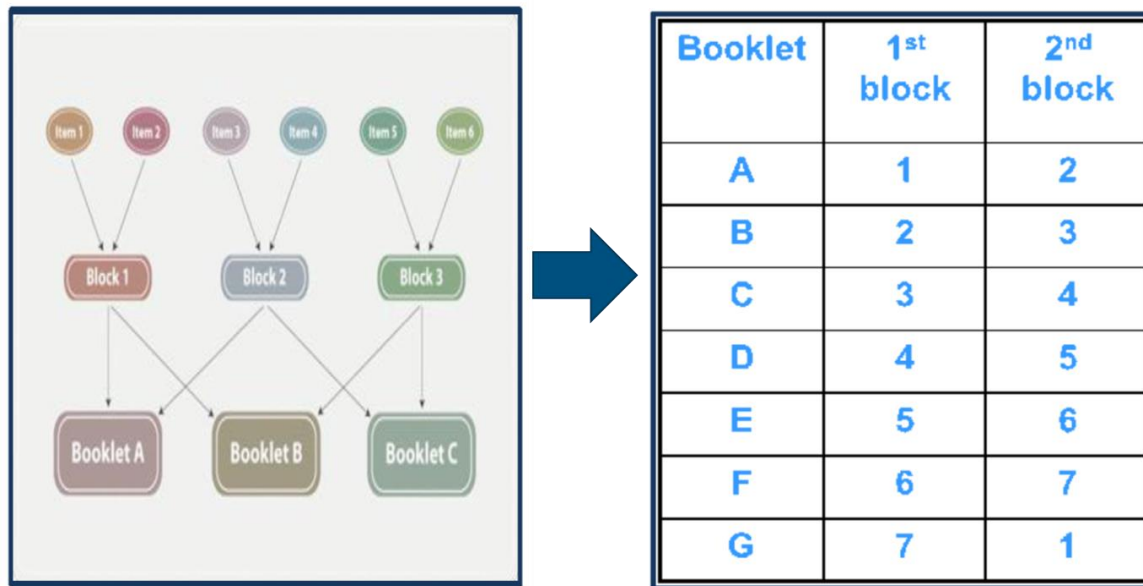
91	MRPS11	612	5	2	C
92	MRPS12	617	5	2	C
93	MRPS13	622	5	2	C
94	MRPS14	627	5	2	C
95	MRPS15	632	5	2	C

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 assessment 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.

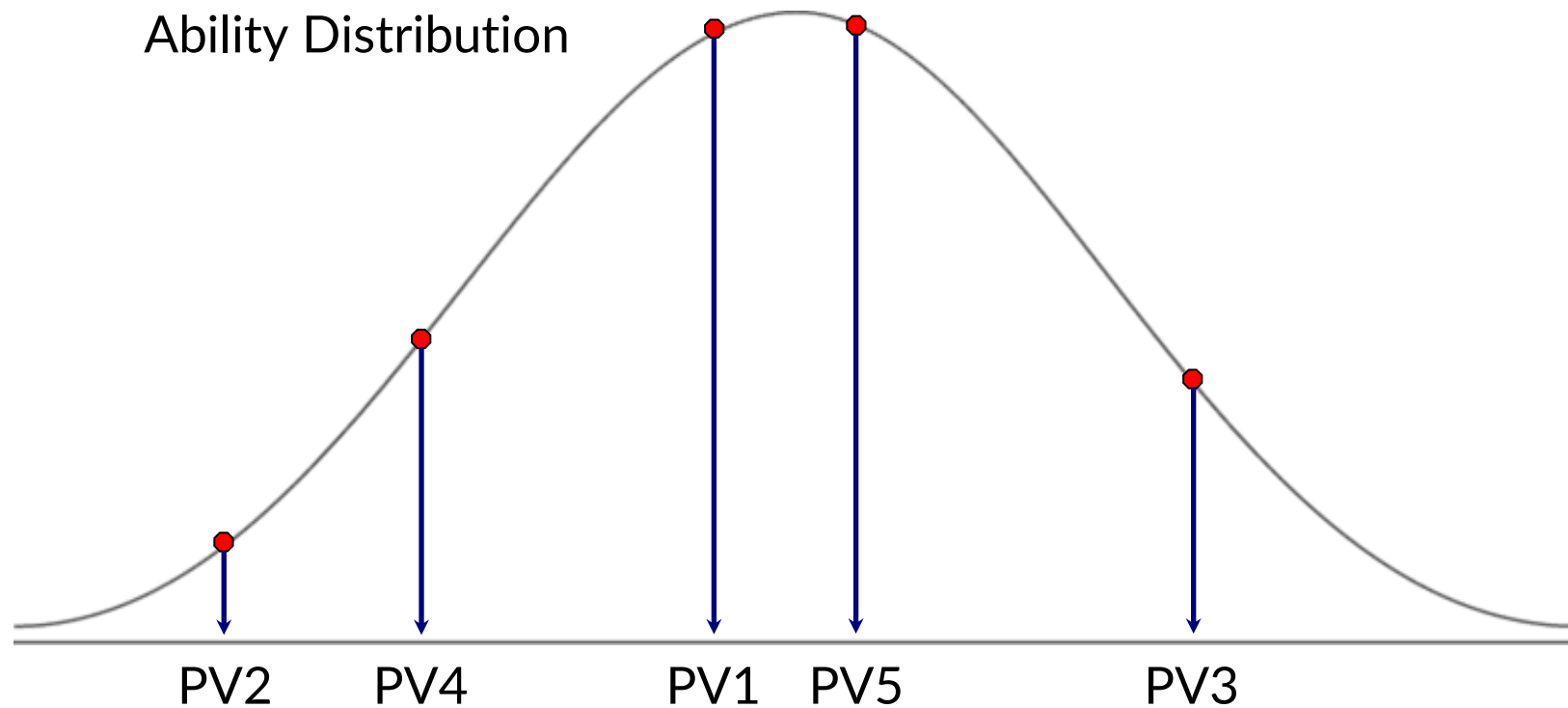
How can an assessment program work without accurate scores for individual students?



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 items, x_i , and survey questionnaire items, y_i :

Latent regression

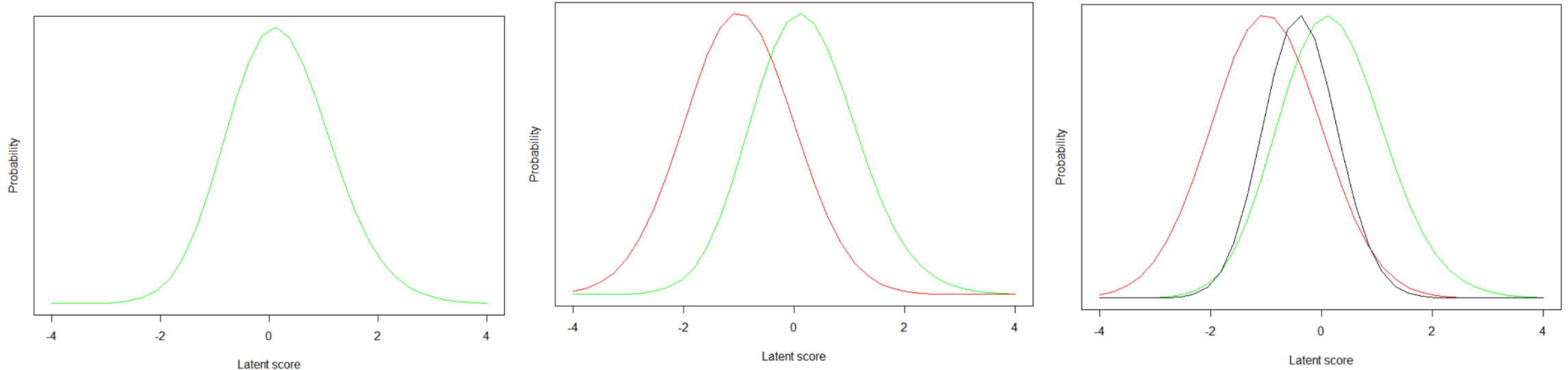
Likelihood function
based on the item
responses

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

- $\boldsymbol{\beta}$ 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 Direct Estimation approach

It is the 3rd stage final model

Through the final model, obtain regression estimates for conditioned subgroups/contextual variables

Weights and sampling design variables (e.g., PSU and STRATA) applied

Taylor series method is used for variance estimation

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

Pro and con of the PV approach

- Already available in the datasets. Existing statical packages to handle them.
- You can only use covariates that were included in the conditioning model for PVs
 - Adding new variable to analytical model may bias the results.
- Computation intensive. Have to rely on testing companies and special internal software.

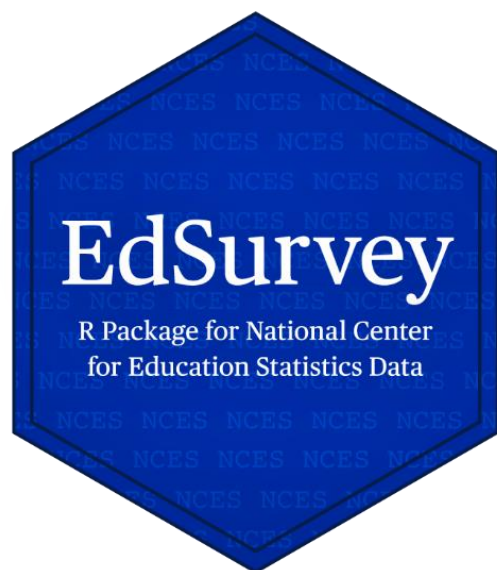


Advantages of the Direct Estimation approach



1. Allows to add covariates (e.g., PC or external variables) without biasing the analytical results
2. Allow to analyze NAEP/TIMSS data properly with selected important covariates and your variables of interest

Why use EdSurvey and Dire?



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 Psychometrician / Statistician

202.403.6646

tzhang@air.org

AMERICAN INSTITUTES FOR RESEARCH[®] | [AIR.ORG](https://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](https://air.org).