Where Were NAEP/TIMSS Scores From? Psychometric and Statistical Models

Ting Zhang, Ph.D.

Senior Researcher

AERA| June 2022

Overview

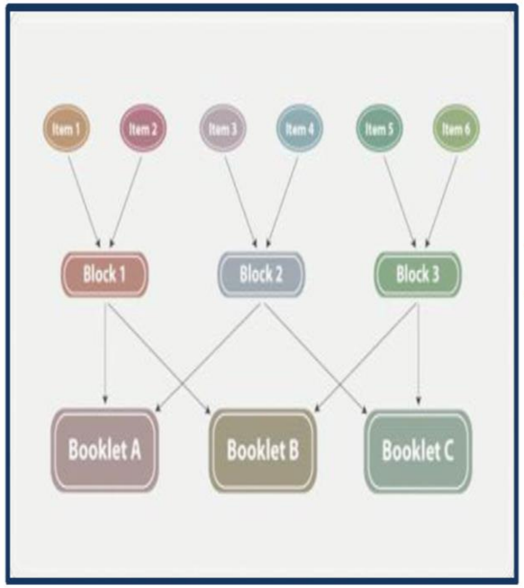
- Why plausible values?
- How plausible values are formed?
- How to use plausible values in a large-scale data analysis?

What are Plausible Values?

91	MRPS11	612	5	2	C	Plausible NAEP math value #1 (num & oper)
92	MRPS12	617	5	2	C	Plausible NAEP math value #2 (num & oper)
93	MRPS13	622	5	2	C	Plausible NAEP math value #3 (num & oper)
94	MRPS14	627	5	2	C	Plausible NAEP math value #4 (num & oper)
95	MRPS15	632	5	2	C	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?



Booklet	1st block	2nd block
A	1	2
B	2	3
C	3	4
D	4	5
E	5	6
F	6	7
G	7	1

Assessment designs!

- Test design features
  - Large scale assessments such as NAEP and TIMSS 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.
  - But at the assessment level, all items are measured across the assessed population.

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

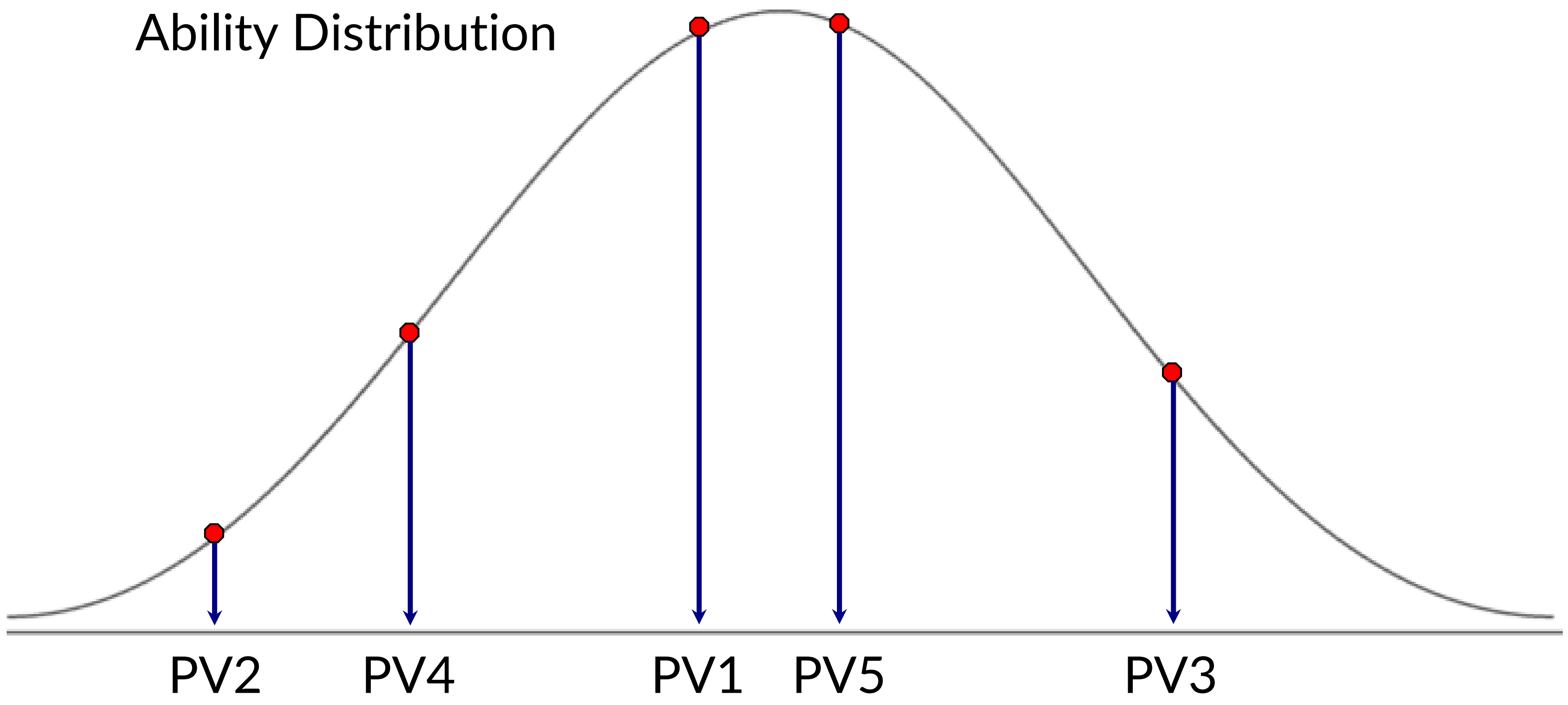
- Each student receives too few test questions to permit a reliable estimate of scale score at the individual level.
- Results have large measurement error and leads to inaccurate inference.

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 represent a student’s proficiency.
- Plausible values are based on student responses to the subset of items they receive and available background information (Mislevy, 1991).

Plausible Values



How Plausible Values 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;

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: Population model (conditional model)

The values of $\boldsymbol{\theta}$ are derived from a latent regression equation, referred to as the conditioning model

$$\theta_i = \boldsymbol{\Gamma}'\boldsymbol{X}_i + \varepsilon_i$$

- Where θ_i are the latent distribution that represents a student's proficiency
- Where $\boldsymbol{X}_i$ are is the observed responses to survey items
  - In operation, we don't use the raw variables for $\boldsymbol{X}$, rather we reduce the dimensions of x to principal components which account for 90% of the variance in $\boldsymbol{X}$
- $\boldsymbol{\Gamma}$ are the latent regression parameters
- ε_i's follow multivariate normal distribution with mean zero and variance-covariance matrix $\boldsymbol{\Sigma}$

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

3rd stage: Final model

- Plausible values are drawn from the poster[…] latent trait […]ed responses to[…] items, x_i, and survey questionnaire items, y_i:

Latent regression

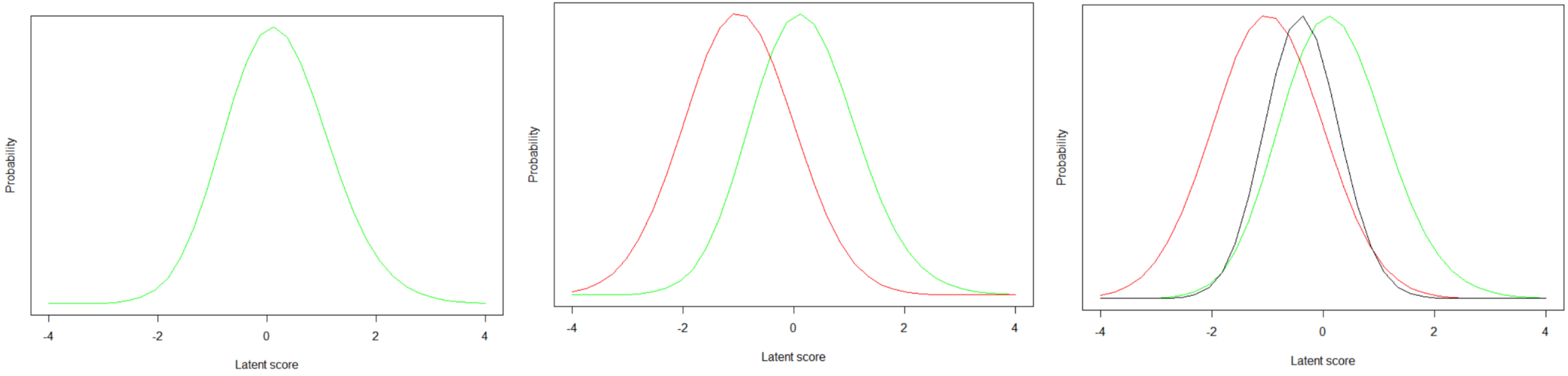
Likelihood function based on the item responses

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

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

Likelihood distribution from the final model

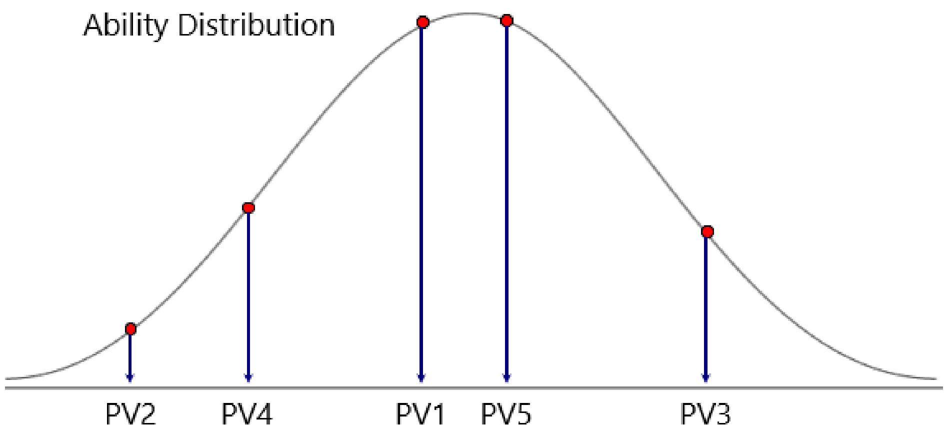
$$f(\theta_i|\boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) \propto \boxed{\phi(\theta_i; \boldsymbol{\Gamma}'\boldsymbol{X}_i, \boldsymbol{\Sigma})}\boxed{\prod f_i(Y_{ij}|\theta_i, \beta_j)}$$



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

What are Plausible Values?

- Instead of individual student scores, we draw multiple potential values from the posterior distribution of the latent trait given the observed responses to both the assessment items and survey questionnaire.
  - TIMSS: 5 PVs
  - NAEP: 20 PVs starting 2013; 5 PVs prior to 2013
- 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, …, 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^*:

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

measure variance + sampling variance

Examples using NAEP Primer data

- All plausible values were used

```
> es1 <- edsurveyTable(composite ~ dsex, data = sdf)
> es1$data
    dsex    N    WTD_N      PCT   SE(PCT)     MEAN  SE(MEAN)
1   Male 8486 8511.974 50.27015 0.5016796 276.7235 0.8207151
2 Female 8429 8420.489 49.72985 0.5016796 275.0458 0.9402535
```

- Only one plausible value was used

```
> es2 <- edsurveyTable(mrpcm1 ~ dsex, data = sdf)
> es2$data
    dsex    N    WTD_N      PCT   SE(PCT)     MEAN  SE(MEAN)
1   Male 8486 8511.974 50.27015 0.5016796 276.8186 0.8180693
2 Female 8429 8420.489 49.72985 0.5016796 275.2309 0.9230248
```

Poll

4. Why are the SEs from the 2nd example SMALLER than the SEs from the 1st example?

A. The sampling weights are not applied

B. The outcome variables are entirely different

C. The measurement variance is missing

D. The sampling variance is missing

Takeaways

- When conducting a NAEP or TIMSS 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

Tools analyzing data with Plausible Values

- *EdSurvey* package in R is designed to analyze NCES data with plausible values and complex sampling design.
- *Dire* package in R analyze NAEP and TIMSS data and conduct direct estimation for students' scale scores.
- Standard statistical software packages can also be used, such as SAS, Stata, or SPSS
- For simple analyses (e.g. comparing group means, simple correlations, summary tables), check out the NAEP Data Explorer and International Data Explorer.

Packages should only be used when they include methods to take into account the measurement and sampling errors.

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.

Advancing Evidence.
Improving Lives.

Ting Zhang

Senior Psychometrician / Statistician
202.403.6646
tzhang@air.org