Where Were Scores From?

Plausible Values & Direct Estimation Approaches

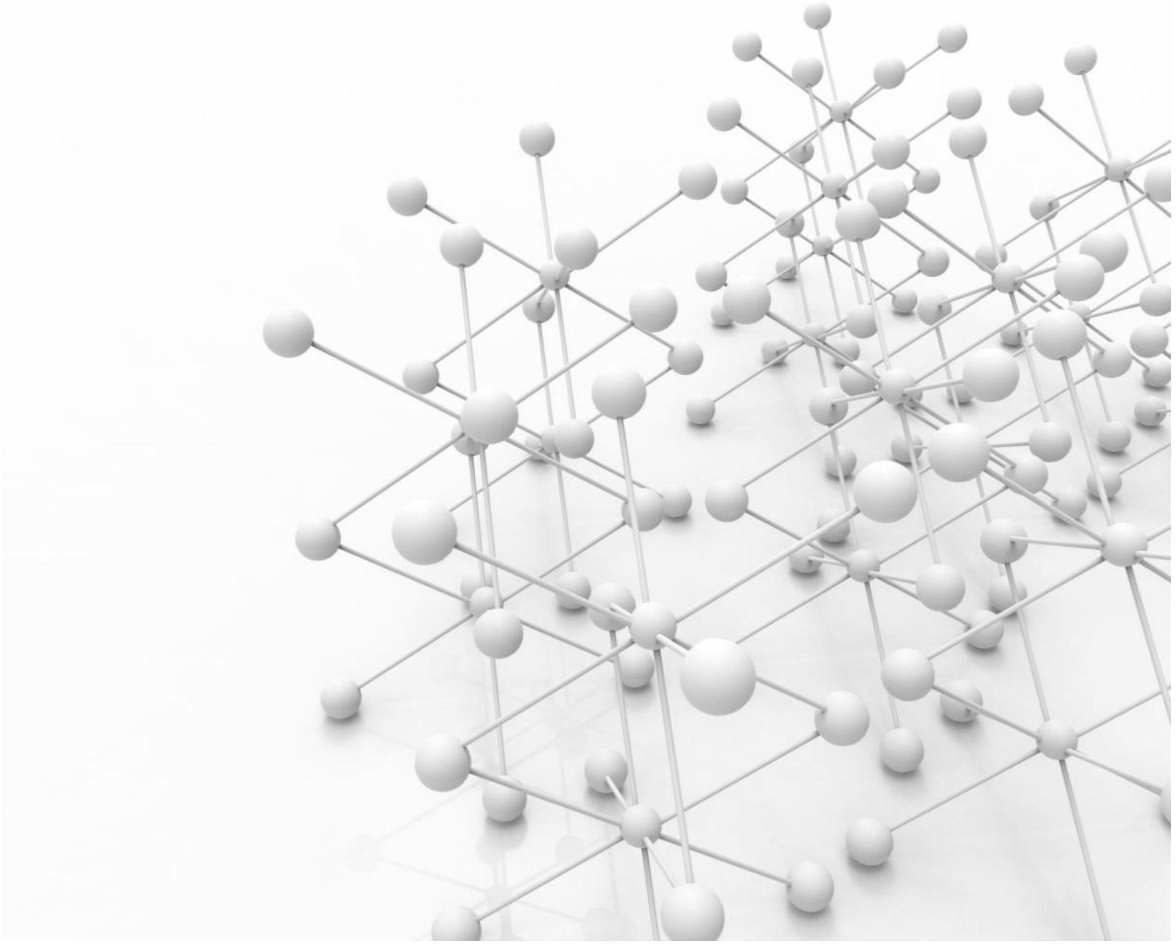
Ting Zhang, Ph.D.

Senior Researcher

IMPS Workshop| July 2022

Agenda

1. Psychometric models for TIMSS data
2. The Direct Estimation Approach
3. The Plausible Values Approach



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?

Assessment Blocks				
Student Achievement Booklet	Part 1		Part 2	
Booklet 1	M01	M02	S01	S02
Booklet 2	S02	S03	M02	M03
Booklet 3	M03	M04	S03	S04
Booklet 4	S04	S05	M04	M05
Booklet 5	M05	M06	S05	S06
Booklet 6	S06	S07	M06	M07
Booklet 7	M07	M08	S07	S08
Booklet 8	S08	S09	M08	M09
Booklet 9	M09	M10	S09	S10
Booklet 10	S10	S11	M10	M11
Booklet 11	M11	M12	S11	S12
Booklet 12	S12	S13	M12	M13
Booklet 13	M13	M14	S13	S14
Booklet 14	S14	S01	M14	M01

Assessment designs!

- Test design features
  - Large scale assessments such as 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.

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

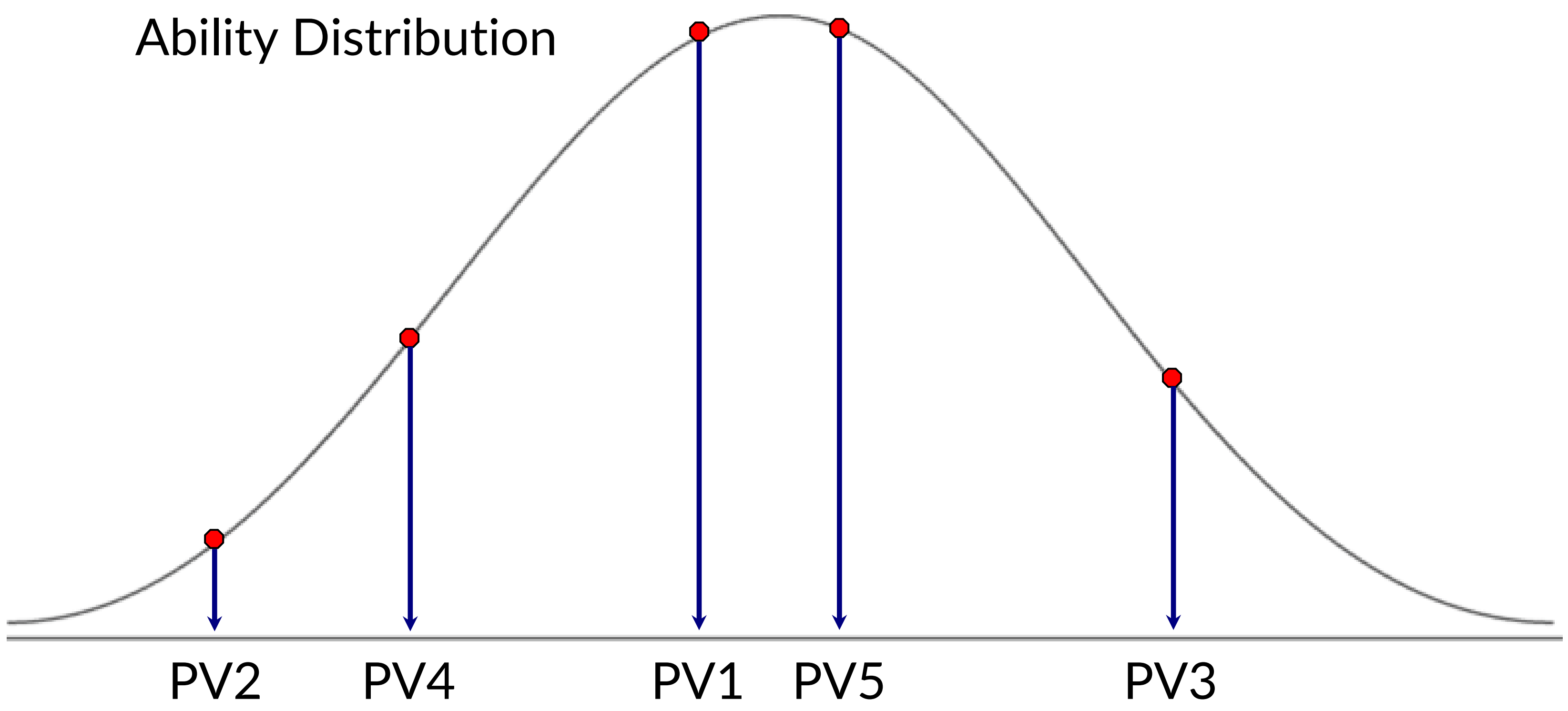
- Each student receives too few test questions to permit estimating an accurate scale score for that student.
- Results in large measurement error and leads to inaccurate inference.

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 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).

Ability Distribution and Plausible Values



How 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

Three parameter (3PL) logistic model

- The probability of a correct response to a multiple choice item depends on the ability of the person ***i*** and 3 properties (parameters) of the item ***j***

$$P\left(y_{ij}=1\,|\,\vartheta_i\right)=c_j+\frac{1-c_j}{1+e^{-1.7a_j\left(\vartheta_i-b_j\right)}}\equiv P_j\left(\vartheta_i\right)$$

- Where:
  - y_{ij} is the response of person *i* to question *j*
  - ϑ_i is the latent trait for person *i*
  - a_j is the discrimination (or slope) parameter for item *j*
  - b_j is the difficulty (or location) parameter for item *j*
  - c_j is the guessing parameter for item *j*

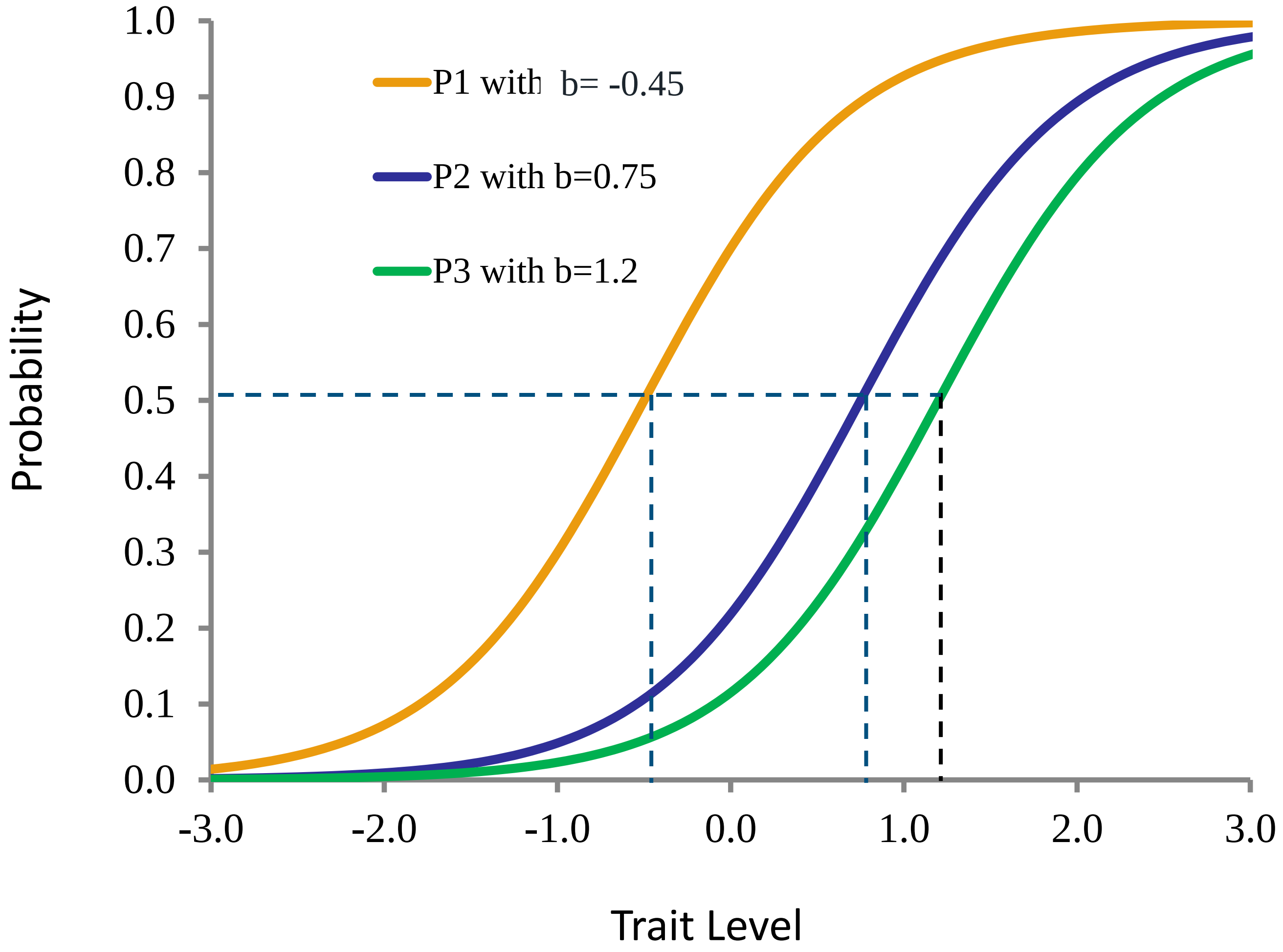
Two parameter (2PL) logistic model

- The probability of a correct response to a dichotomous item depends on the ability of the person ***i*** and 2 properties (parameters) of the item ***j***

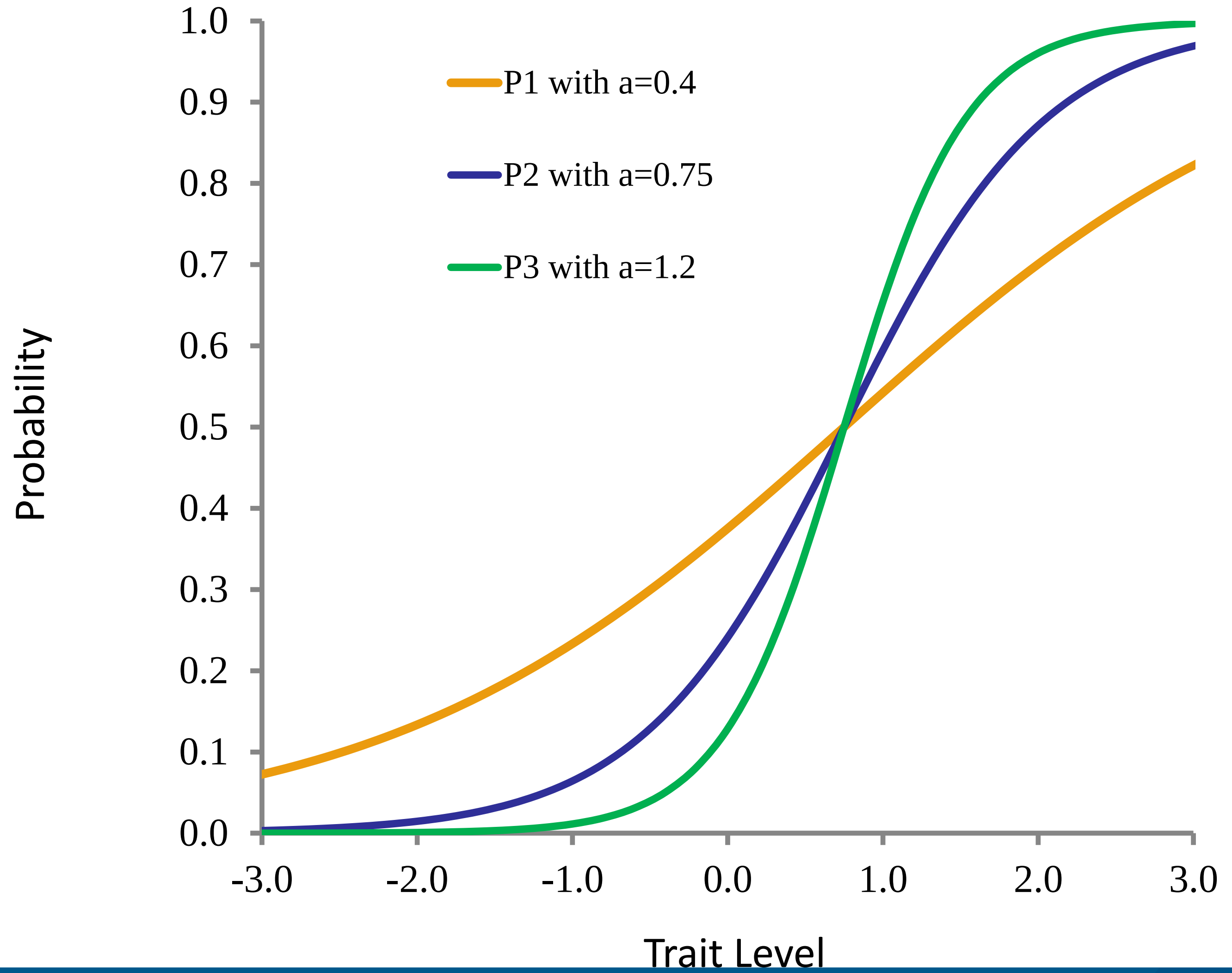
$$P\left(y_{ij}=1 \mid \vartheta_i\right)=\frac{1}{1+e^{-1.7 a_j\left(\vartheta_i-b_j\right)}} \equiv P_j\left(\vartheta_i\right)$$

- Where:
  - y_{ij} is the response of person *i* to question *j*
  - ϑ_i is the latent trait for person *i*
  - a_j is the discrimination (or slope) parameter for item *j*
  - b_j is the difficulty (or location) parameter for item *j*

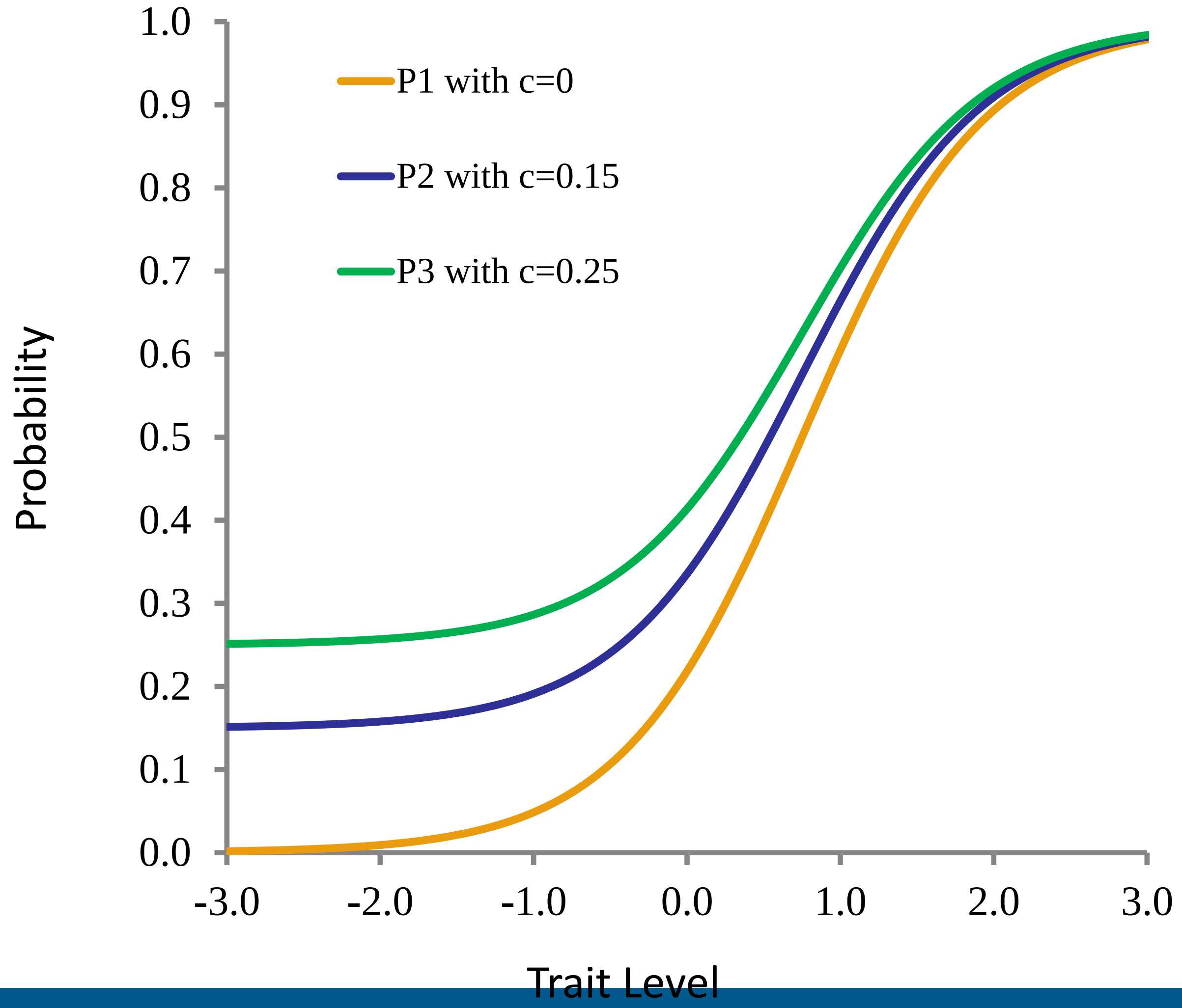
Item Characteristic Curves
Difficulty parameter



Discrimination parameter



Guessing parameter



Generalized partial credit (GPC) model

- The probability of getting a particular score on a constructed response item, e.g., a score of 2 in a writing task with a score scale 0 to 4, depends on the ability of the person ***i*** and properties of the item ***j***

$$P\left(y_{ij}=k\mid \vartheta_i\right)=\frac{e^{\left(\sum_{k=0}^{g}1.7a_j\left(\vartheta_i-b_j+d_{jk}\right)\right)}}{\sum_{g=0}^{m_j-1}e^{\left(\sum_{k=0}^{g}1.7a_j\left(\vartheta_i-b_j+d_{jk}\right)\right)}}$$

Where

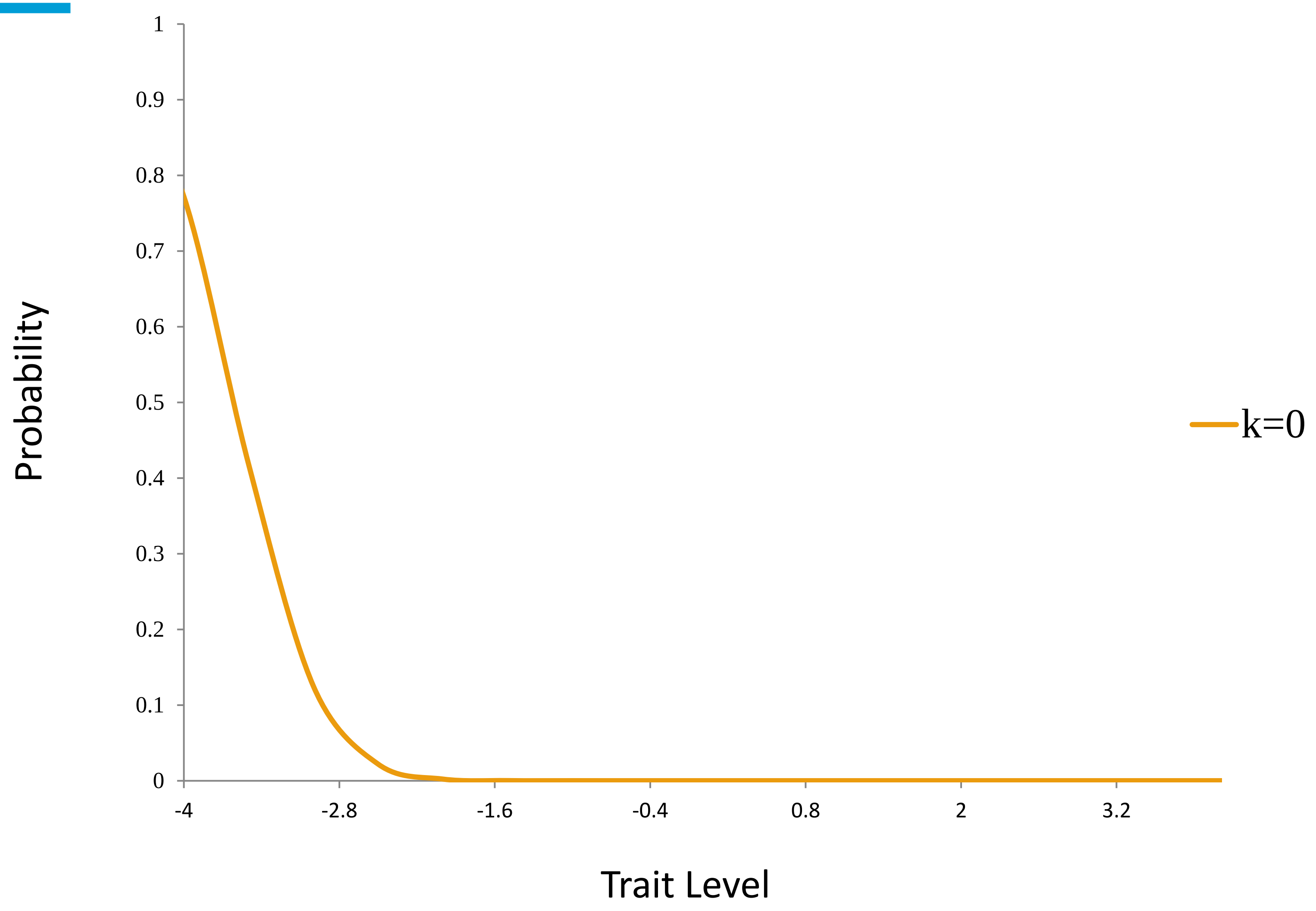
ϑ_i is the ability of person i

a_j is the discrimination parameter

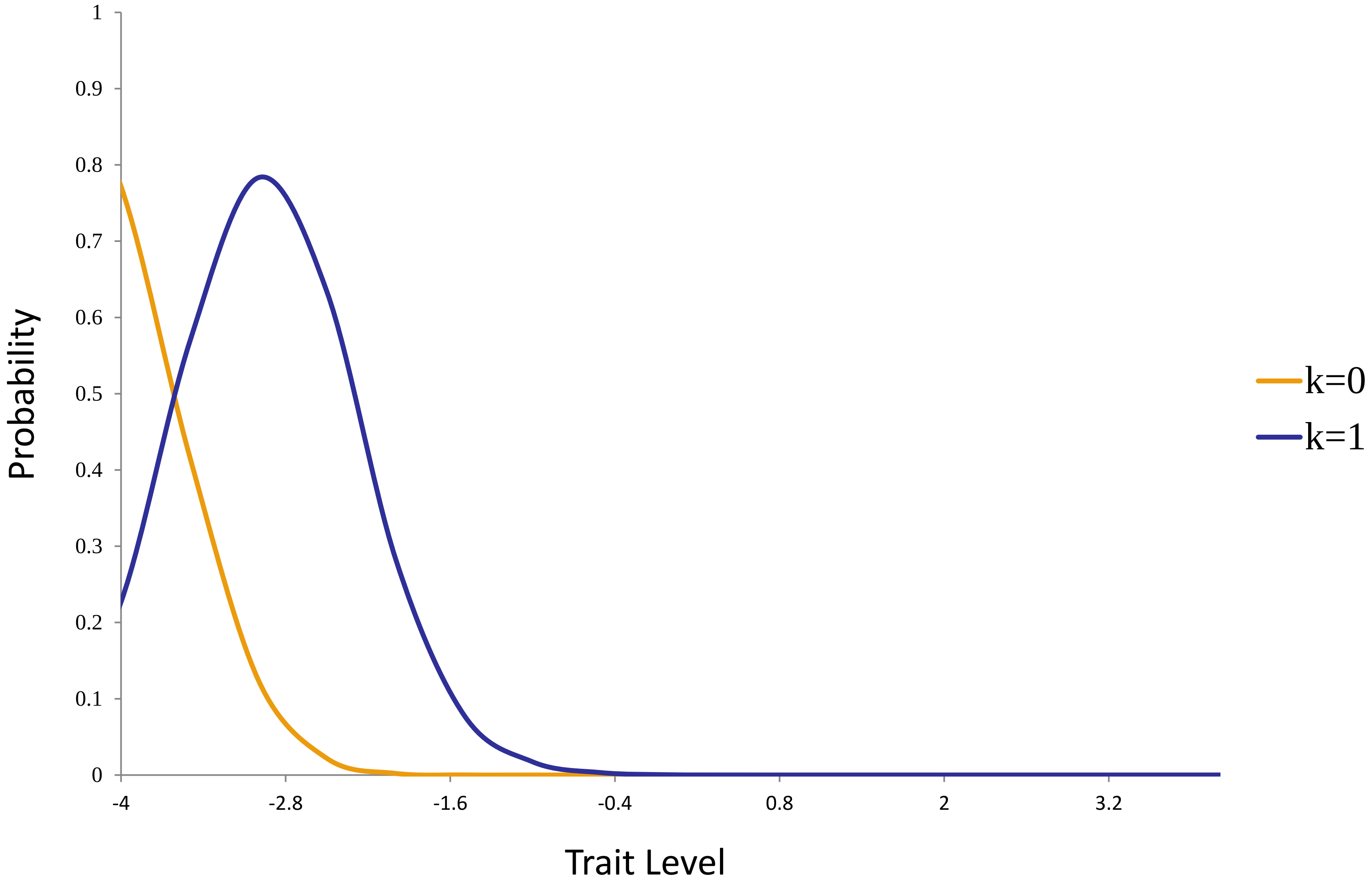
b_j is the base difficulty parameter

d_{jk} is the set of category parameters that add to or subtract from the base difficulty parameter

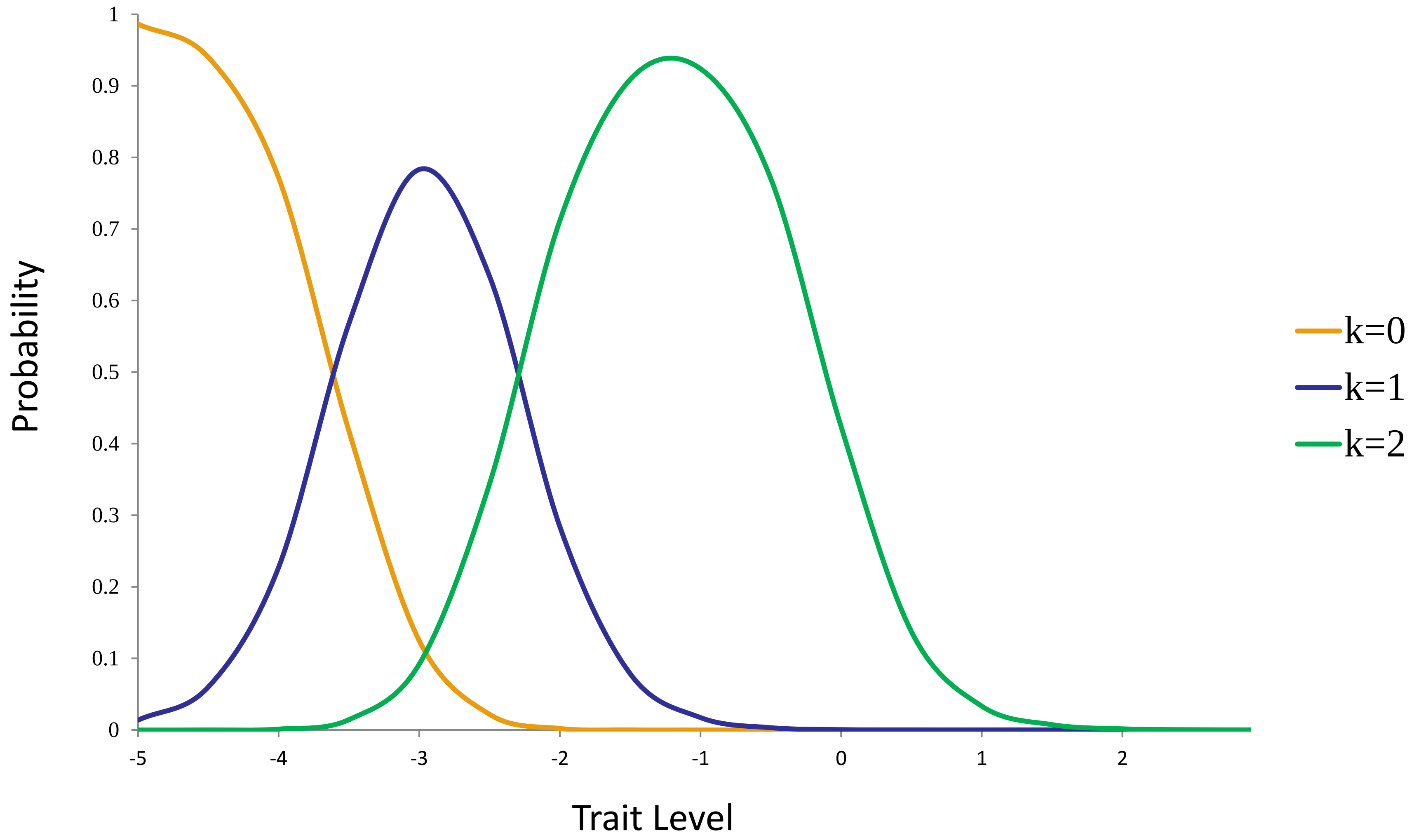
Category Response Curves
Probability of score 0 for item j



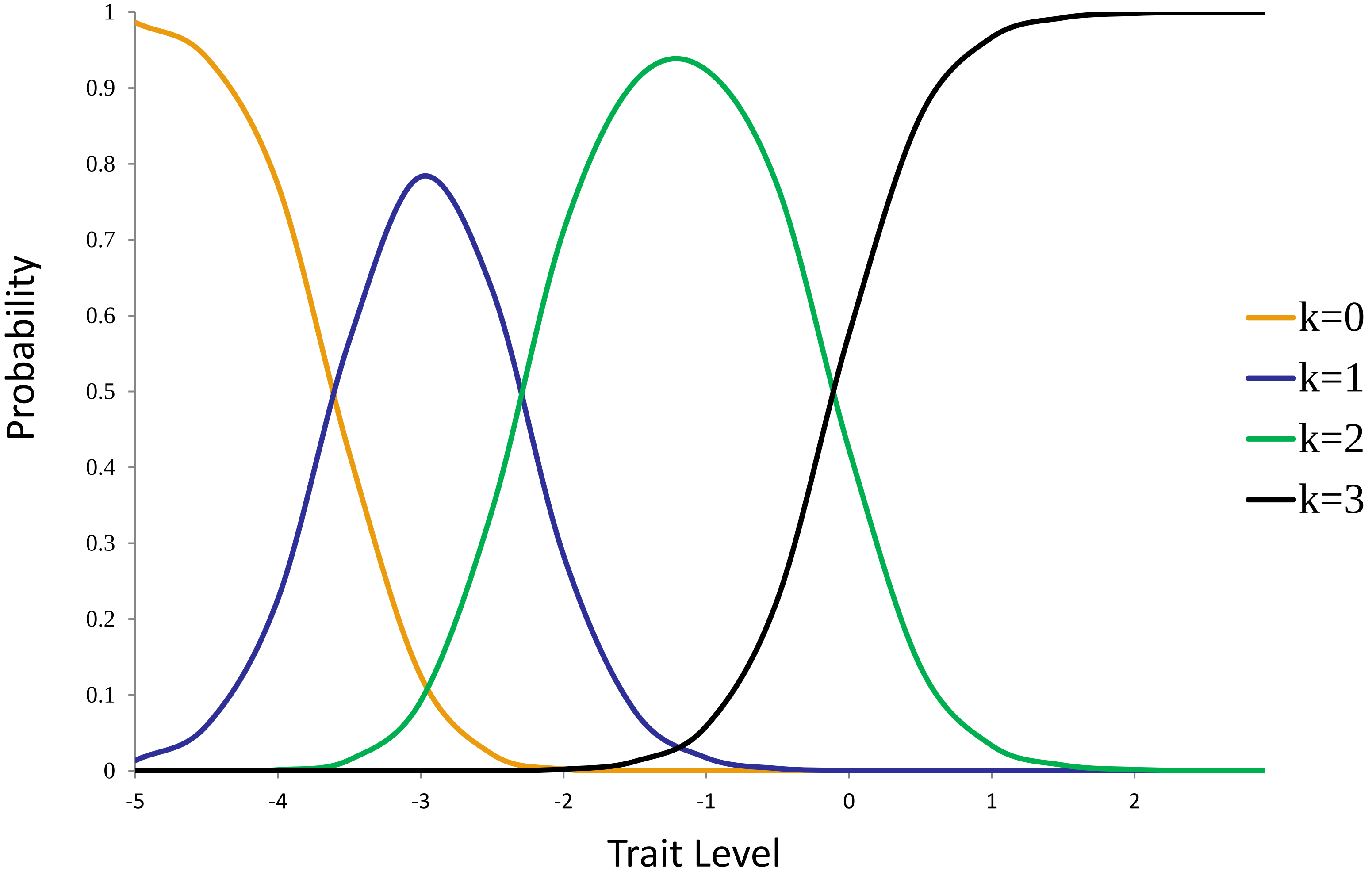
Probability of score 0 and 1, each for item j



Probability of score 0, 1, and 2, each for item j



Probability of score 0, 1, 2, and 3, each for item j



AIR

Likelihood of the pattern of answers from an examinee

- Likelihood of examinee i providing a given pattern of responses to a set of q items is:

$$L_i(Z_i \mid \vartheta_i) = \prod_{j=1}^{q} \prod_{k=0}^{m_j-1} P_{jk}(\vartheta_i)^{z_{ijk}} \left(1 - P_{jk}(\vartheta_i)\right)^{1-z_{ijk}}$$

Where

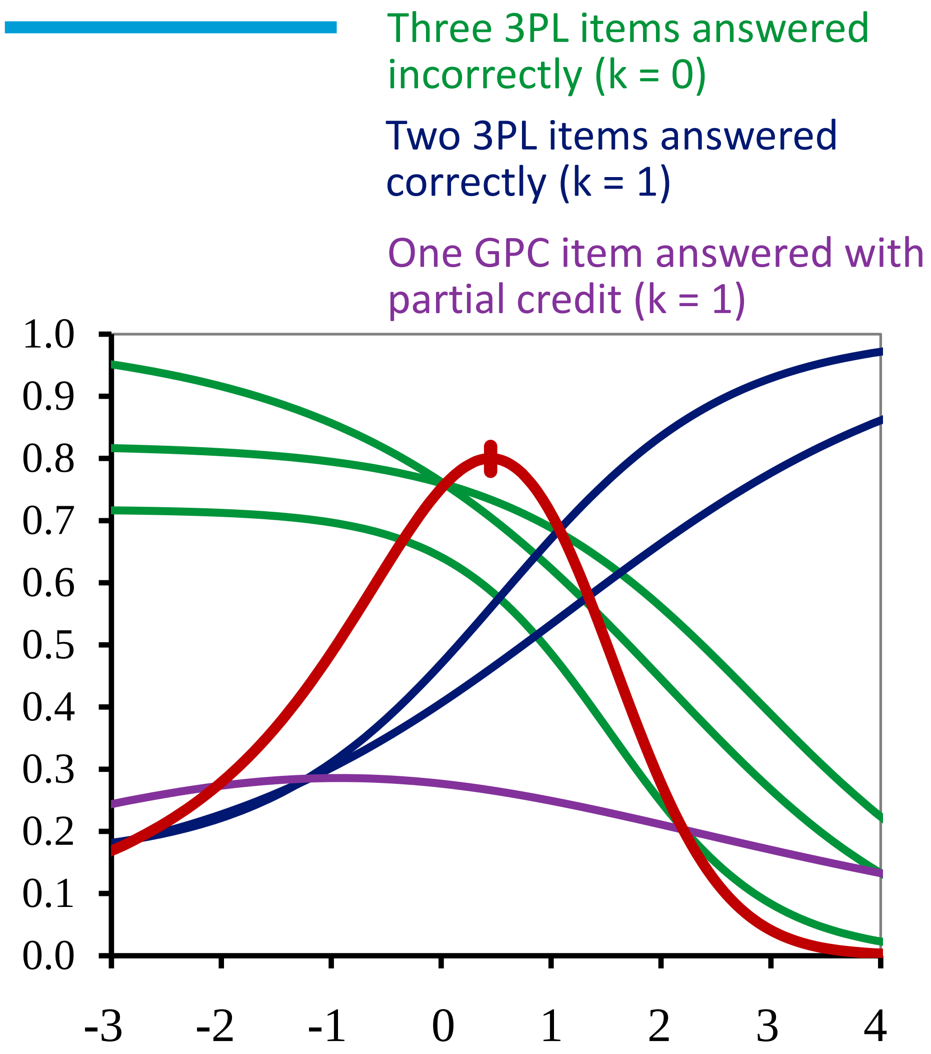
z_{ijk} is a response in category k to item j by person i

Z_i is the vector of response to q items by examinee i to all j items

- This likelihood function for the proficiency θ_i is called the posterior distribution of the θ's for each examine.

AIR

Example: Likelihood function based on six TIMSS math items



Notes:

1. At each point along the horizontal axis, the value of the item characteristic curves are multiplied together (and rescaled vertically for this graph)

2. For the purpose of this explanation, the likelihood function, shown in red, is approximated to be a normal distribution

3. The maximum value could be considered the student's test score, if the curve were not so wide

Why don't we stop here?

2nd stage: Conditioning model (population 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 represent a student's proficiency
- Where $\boldsymbol{X}_i$ are 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 $\boldsymbol{\Sigma}$

NAEP
NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

AIR

3rd stage: Final model

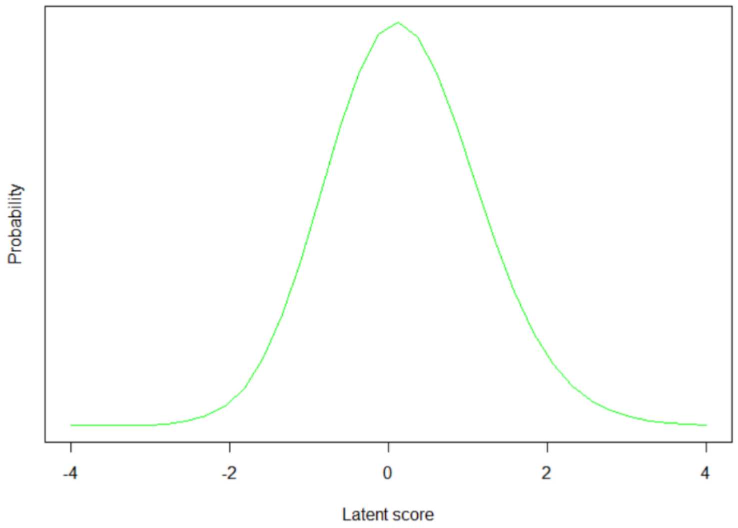
- Plausible values are drawn from the posterior distribution of the latent trait given the observed responses to both the assessment items, x_i, and survey questionnaire items, y_i:

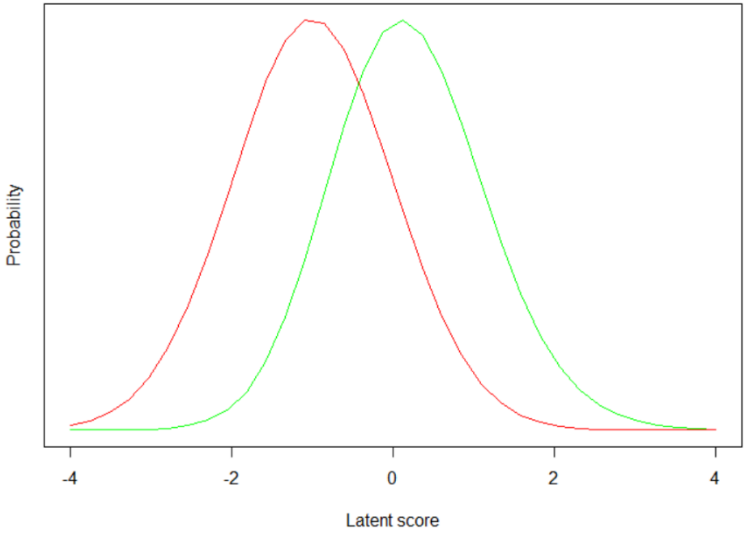
$$f(\theta_i|\boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) \propto \boxed{\phi(\theta_i; \boldsymbol{\Gamma}'\boldsymbol{X}_i, \boldsymbol{\Sigma})} \boxed{\prod f_i(Y_{ij}|\theta_i, \beta_j)}$$ Where

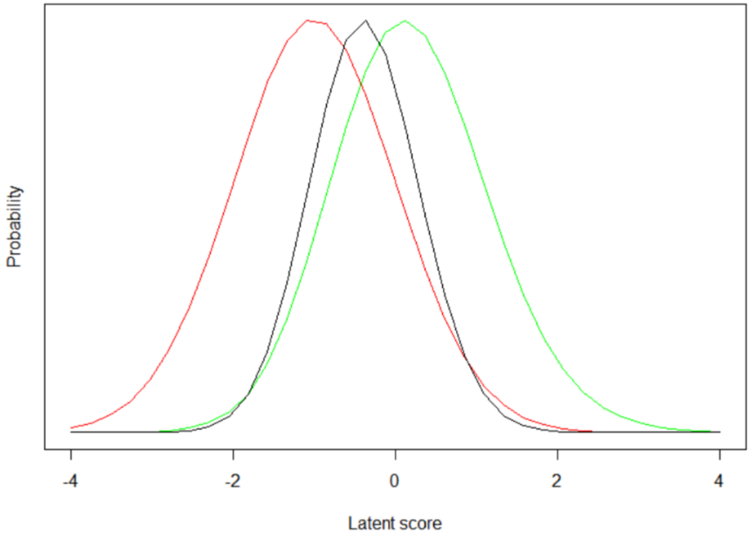
  - $\boldsymbol{\beta}$ are the item parameters
  - $\boldsymbol{\Gamma}$ are the latent regression parameters
  - $\boldsymbol{\Sigma}$ is a covariance matrix
  - $\phi(\theta_i; \boldsymbol{\Gamma}'\boldsymbol{X}_i, \boldsymbol{\Sigma})$ is a normal distribution with mean $\boldsymbol{\Gamma}'\boldsymbol{X}_i$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

The Direct Estimation approach

It is the 3rd stage final model

Proficiency levels, standard errors (measurement errors), and regression estimates are obtained through MML

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

Taylor series method is used for sampling 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

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

AIR

Advantages of the Direct Estimation approach

1. Allows to add covariates (e.g., PC or external variables) without biasing the analytical results
2. Support generating of new PVs with selected important covariates and your variables of interest

Why use EdSurvey and Dire?

Dire offers direct estimation and draw PVs functions. Good alternative for AM.

EdSurvey support both the PV approach and DE approach and acts as an interface to Dire, streamlines all procedures required for DE:

- Implements scoring, IRT parameters, scaling information automatedly
- Makes item adjustments made by NCES
- Allows users to use new PVs in the same way NCES PVs were used--all EdSurvey functions are available for the new PVs!
- Takes care of complex design features such as weighting and variance estimation behind the scene

Dire
R Package for National Center for Education Statistics Data

EdSurvey
R Package for National Center for Education Statistics Data

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

Advancing Evidence.
Improving Lives.

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^*:

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

measurement variance
+ sampling variance