Assessment Design and Implications for Analysis

PIAAC International Database Training
Prague, Czech Republic
May 13-15, 2014

Some of the Challenges...

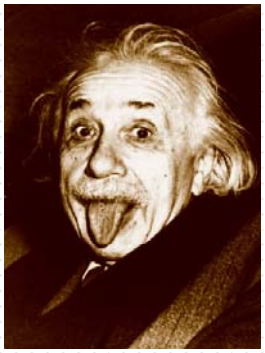
- There are several domains
- Domains are very broad
- Limited testing time (physical & psychological)
- Challenges:
  - ✓ Need to administer the items in a “sensible” design
  - ✓ Need to summarize performance on the items
  - ✓ Need to account for unreliability of estimates

How Does Sampling Help?

- Impossible to test everyone on everything
  - ✓ Too many people
  - ✓ Too many items
  - ✓ Too expensive
- Not necessary to test everyone on everything
  - ✓ Blood sample
  - ✓ Soup sample
- Some people are tested on some things
- Results should be seen in the context of the person and item sample design

How Does Sampling Help?

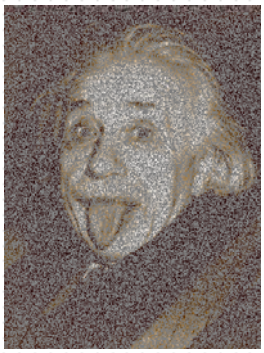
- We select a few people from the entire population
- To those selected, we give a selected set of items
- This is known as multiple matrix sampling



How Does Sampling Help?

- Some examples…





General Sample Design

- Population of interest
  - ✓ Adults 16-65 years of age
  - ✓ Other “special samples” per country choice

General Sample Design

- Multistage stratified cluster sampling design
  - ✓ Multistage
    - There are different stages/levels of selection
    - For example: Municipalities–Block–Household-Person
  - ✓ Stratified (implicit or explicit)
    - Selection takes place across different segments of the population
    - Achieved by systematic selection across sorted list, or targeted selection within different groups
  - ✓ Cluster
    - Multiple individuals are selected from within segments of the population
    - Segments could be municipalities, blocks, etc.

Why Do We Do It This Way?

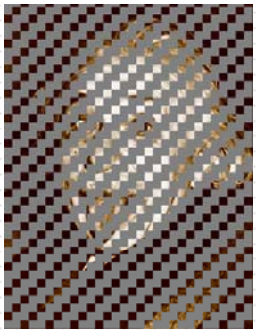
- Among many reasons…
  - ✓ Availability of information
  - ✓ Cost reduction
  - ✓ Ensure representation of target population groups
  - ✓ Achieve desired precision levels for target groups
  - ✓ Redundancy

What are the Consequences?

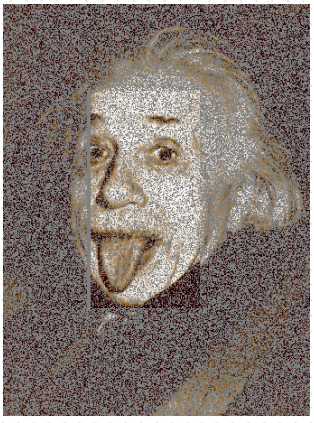
- We DO NOT have a “simple random selection/sample” (SRS) from the population
  - ✓ Think of selecting clusters, and then persons within clusters
    - Persons within a cluster are likely to be more similar to each other than to persons in other clusters
- This matters when we compute sampling errors

Selecting Individuals vs. Clusters

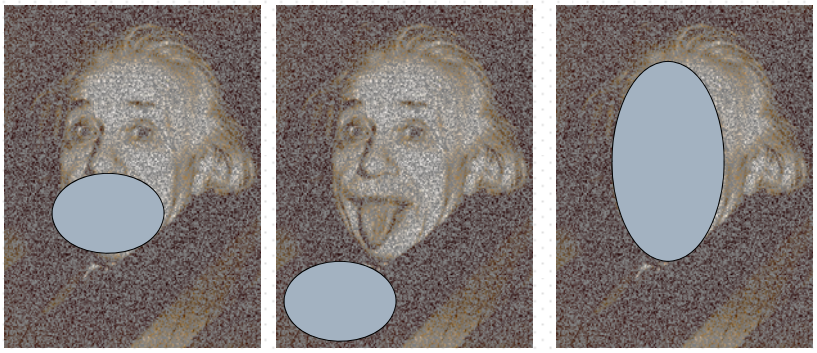
Random vs. Systematic Selection



Oversampling



Exclusions from the Population



Sampling Weights

- Sampling weights are inverse of the probability of selection for a person
- They take into account characteristics of the sample and selection procedure
  - ✓ Stratification or disproportional sampling of subgroups
  - ✓ Adjustments for nonresponse
  - ✓ Selection probability of each person is known
  - ✓ Poststratification to external control totals
- Sampling weights must be used ALWAYS to get correct population estimates

Computing Means

$$Unweighted\_Mean(x) = \frac{\sum x}{N}$$

$$Weighted\_Mean(x) = \frac{\sum wgt * x}{\sum wgt}$$

An Example…

- See file: “Example of Using Weights.xlsx”

Estimating Sampling Variance

- SRS of 5,000 persons from across the country covers the population diversity better than a sample of 200 clusters with 25 persons each
  - ✓ A stratified multistage cluster design has more uncertainty associated with its estimates than a SRS of the same size
- The increase in uncertainty is directly related to the differences between and within the clusters

Estimating Sampling Variance

- Consider the following extreme cases:
  - ✓ All clusters are <u>different</u> from each other, but within them all persons are <u>identical</u>
  - ✓ All clusters are <u>identical</u> to each other, but within them all persons are <u>different</u>
- How many clusters would you choose?
- How many persons would you choose?

Estimating Sampling Variance

- Sampling automatically results in some uncertainty (called “error” or “variance”)
- Which factors can influence the magnitude of variance in a sample?
  - ✓ How we sample
  - ✓ Sample size
  - ✓ Variability within the population (between and within clusters)
- Think of a simple random sample (SRS) and how these factors influence the variance of an estimate

Estimating Sampling Variance

- **Main point:** These samples were clearly not selected using SRS
- Most analysis procedures (**standard** SAS, **standard** SPSS, etc.), assume SRS, and should NOT be used when analyzing these data
  - ✓ Special procedures have been developed to calculate sampling variance for these cases

Estimating Sampling Variance

- Because we cannot sample from the population over and over again, we use replicate samples
- Brief explanation
  - ✓ Systematically delete subsamples from the full sample to form replicate samples
  - ✓ Adjust weights of the remaining units to account for the deleted units – new weights are called replicate weights
  - ✓ Produce an estimate using the full sample weight and an estimate from each set of replicate weights
  - ✓ Calculate the variance of the replicate estimates from the full sample estimate

Estimating Sampling Variance

- The variation of the replicates from the full sample estimate provides a measure of the variance of the full sample estimate
- Advantages of replication:
  - ✓ Convenient to use
  - ✓ Effects of nonresponse and other adjustments can be reflected in replicate weights
  - ✓ Estimates can be computed for subpopulations
  - ✓ Readily applicable to most statistics
- Disadvantage
  - ✓ Very computer intensive…

Estimating Sampling Variance

- Several Methods
  - ✓ Jackknife Repeated Replication (JRR)
    - Variant 1 (JK1)
    - Variant 2 (JK2)
  - ✓ Balanced Repeated Replication (BRR)
    - Fay's variant

How the JK1 works ...

Cluster	R0	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
1	1.00	0.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11
2	1.00	1.11	0.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11
3	1.00	1.11	1.11	0.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11
4	1.00	1.11	1.11	1.11	0.00	1.11	1.11	1.11	1.11	1.11	1.11
5	1.00	1.11	1.11	1.11	1.11	0.00	1.11	1.11	1.11	1.11	1.11
6	1.00	1.11	1.11	1.11	1.11	1.11	0.00	1.11	1.11	1.11	1.11
7	1.00	1.11	1.11	1.11	1.11	1.11	1.11	0.00	1.11	1.11	1.11
8	1.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0.00	1.11	1.11
9	1.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0.00	1.11
10	1.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0.00
11	1.00	0.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11
12	1.00	1.11	0.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11
13	1.00	1.11	1.11	0.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11
14	1.00	1.11	1.11	1.11	0.00	1.11	1.11	1.11	1.11	1.11	1.11
15	1.00	1.11	1.11	1.11	1.11	0.00	1.11	1.11	1.11	1.11	1.11
16	1.00	1.11	1.11	1.11	1.11	1.11	0.00	1.11	1.11	1.11	1.11
17	1.00	1.11	1.11	1.11	1.11	1.11	1.11	0.00	1.11	1.11	1.11
18	1.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0.00	1.11	1.11
19	1.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0.00	1.11
20	1.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0.00

How the JK2 works ...

Strata	Cluster	R0	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
1	1	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	2	1.0	2.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	3	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	4	1.0	1.0	2.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
3	5	1.0	1.0	1.0	2.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	6	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
4	7	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0
	8	1.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0	1.0	1.0	1.0
5	9	1.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0	1.0	1.0
	10	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0
6	11	1.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0	1.0
	12	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0
7	13	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0
	14	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0
8	15	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0
	16	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0
9	17	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0
	18	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0
10	19	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0
	20	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0

How the BRR Fay‘s Variant works...

Strata	Cluster	R0	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12
1	1	1.0	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5
	2	1.0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
2	3	1.0	1.5	0.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5
	4	1.0	0.5	1.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5
3	5	1.0	1.5	0.5	0.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5
	6	1.0	0.5	1.5	1.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5
4	7	1.0	1.5	1.5	0.5	0.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5
	8	1.0	0.5	0.5	1.5	1.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5
5	9	1.0	1.5	0.5	1.5	0.5	0.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5
	10	1.0	0.5	1.5	0.5	1.5	1.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5
6	11	1.0	1.5	0.5	0.5	1.5	0.5	0.5	1.5	0.5	1.5	1.5	1.5	0.5
	12	1.0	0.5	1.5	1.5	0.5	1.5	1.5	0.5	1.5	0.5	0.5	0.5	1.5
7	13	1.0	1.5	0.5	0.5	0.5	1.5	0.5	0.5	1.5	0.5	1.5	1.5	1.5
	14	1.0	0.5	1.5	1.5	1.5	0.5	1.5	1.5	0.5	1.5	0.5	0.5	0.5
8	15	1.0	1.5	1.5	0.5	0.5	0.5	1.5	0.5	0.5	1.5	0.5	1.5	1.5
	16	1.0	0.5	0.5	1.5	1.5	1.5	0.5	1.5	1.5	0.5	1.5	0.5	0.5
9	17	1.0	1.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5	0.5	1.5	0.5	1.5
	18	1.0	0.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5	1.5	0.5	1.5	0.5
10	19	1.0	1.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5	0.5	1.5	0.5
	20	1.0	0.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5	1.5	0.5	1.5
11	21	1.0	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5	0.5	1.5
	22	1.0	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5	1.5	0.5
12	23	1.0	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5	0.5
	24	1.0	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5	1.5

Estimating Sampling Variance

- Think of the following:
  - ✓ If I take clusters out, recalculate, and results DO change…
    - What can I say about the rest of the clusters not sampled?
    - What can I say about other samples I could have drawn?
  - ✓ If I take clusters out, recalculate, and results DO NOT change…
    - What can I say about the rest of the clusters not sampled?
    - What can I say about other samples I could have drawn?

Calculating Sampling Variance

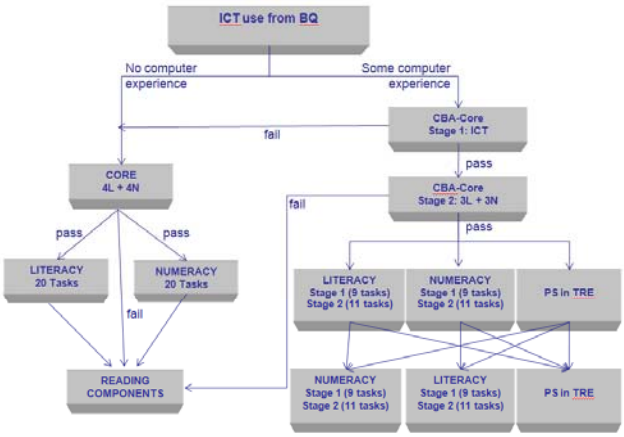
$$Var_{\varepsilon} = f * \sum_{r=1}^{R} \left(\varepsilon_r - \varepsilon_0\right)^2$$

- When…
  - ✓ Using JK1: $f = \frac{VENREPS - 1}{VENREPS}$
  - ✓ Using JK2: $f = 1.0$
  - ✓ Using BRR (w/Fay): $f = \frac{1}{VENREPS * (1 - VEFAYFAC)^2}$
- Method and number of replicates are country-specific

Rationale for IRT Scaling of Data

- IRT: Item Response Theory
  - ✓ Response to an item depends on the interaction between the "ability" of the respondent and characteristics of the item
  - ✓ Allows us to summarize data across multiple items, even if/when different people take different items
  - ✓ Facilitates linking when dealing with multiple test forms
    - Different people take different, but overlapping, sets of items

PIAAC Assessment Design



PIAAC Assessment Design

- Not everyone takes everything
- Path taken is dependent on person skills
  - ✓ No task is given to a representative sample of the entire population
- We need to …
  - ✓ Put everyone on the same scale
  - ✓ Compute uncertainty due to reduced and set of items taken

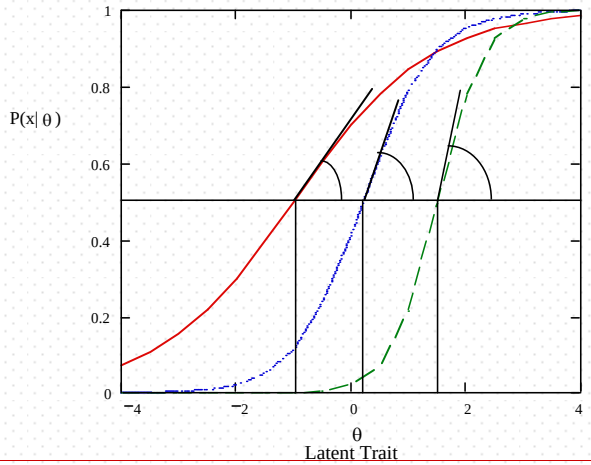
Why Item Response Theory?

- Addresses limitations of classical test theory
  - ✓ Difficulty of a task is defined in reference to probability of a correct answer, given the ability of the individual
  - ✓ Ability of an individual is defined in reference to the likelihood of a correct answer given the difficulty of the task
  - ✓ Once the metric is chosen, the relationship between difficulty and ability can be calculated
  - ✓ Metric is not in reference to particular population or item selection, but in reference to the relationship between the individual and the task

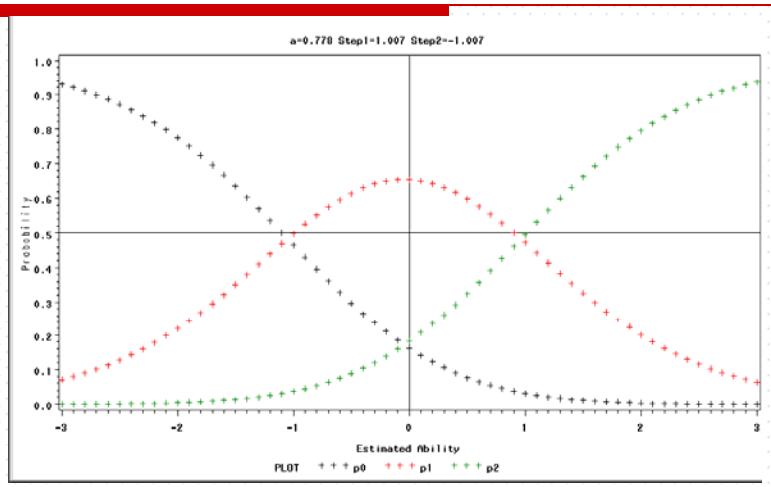
Why Item Response Theory?

- Facilitates dealing with multiple matrix sampling designs
  - ✓ Each person receives a subset of items (booklet)
  - ✓ Each item is administered to a subset of the people
- Total score is based on characteristics of the items taken
  - ✓ High jump example:
    - 66% correct: Jumps 2.00m & 2.05m; fails 2.10m
    - 66% correct: Jumps 1.50m & 1.55m; fails 1.60m

Item Characteristic Curve



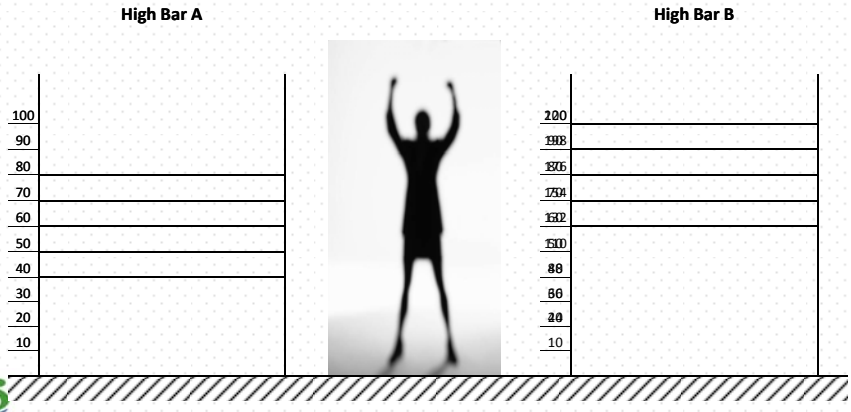
Item Characteristic Curve



Advantages of IRT

- It allows us to:
  - ✓ Evaluate the effectiveness of a test at different levels of ability
  - ✓ Evaluate the ability of respondents who took different set of items
  - ✓ Evaluate a set of items on a same scale established by a different set
  - ✓ Design tests to measure best at specific ability level
  - ✓ Develop new tests and investigate them without administering them
  - ✓ Develop item statistics that do not change when the group of examinees change

How high can you jump…?



Item Response Theory

- Persons of high ability should answer easy items correctly
- Persons of low ability should not answer difficult items correctly
- Probability of a correct answer depends on item parameters and ability of examinee
- Does not make assumptions of normal distribution but assumes unidimensionality of measurement

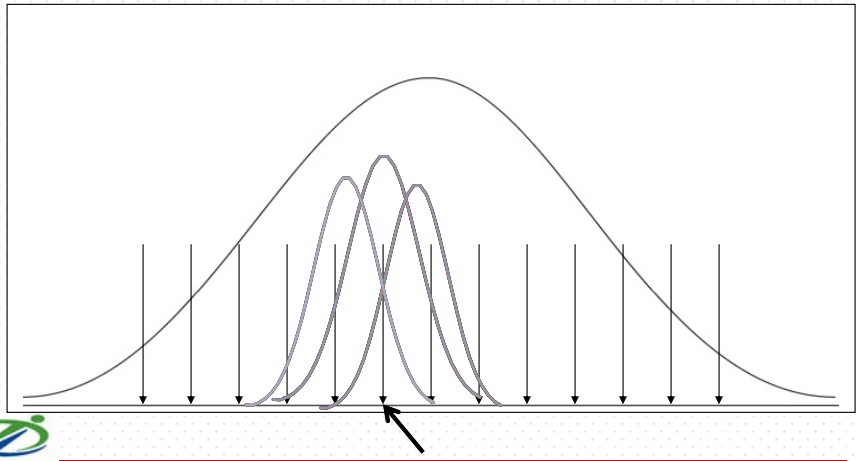
Scaling Procedures

- Achievement is initially estimated using scale scores computed based on IRT
- IRT allows for performance in a domain to be summarized on a common scale even when different persons are administered different items
- In addition to IRT, PIAAC makes use of multiple imputations, or "plausible values" methodology

Scaling Procedures

- Plausible values
  - ✓ Random draws from the estimated ability distribution of students with similar item response patterns and background characteristics
- Think of a regression where the predictors are item responses and background data

What Happens with Few Items?



Using Plausible Values

- Plausible values are optimal for obtaining population estimates
- Plausible values should not be used for individual reporting
- When using plausible values
  - ✓ Compute statistic with each plausible value and average the results
  - ✓ Compute variance due to imputation

$$Var_{\bar{\varepsilon}} = \left(1 + \frac{1}{P}\right) * \frac{\sum_{p=1}^{P} \left(\varepsilon_{0,p} - \bar{\varepsilon}_{0,p}\right)^2}{P-1}$$

Scaling Procedures

- Items calibrated using national subsamples
  - ✓ Countries contribute equally to setting the scales
- Item parameters are used to obtain initial ability estimates
- Person abilities are then estimated using auxiliary/background variables
- Resulting scores are sets of plausible values
  - ✓ PIAAC used 10 plausible values

Calculating Standard Errors

- Not involving plausible values:

$$SE_{\varepsilon} = \sqrt{f * \sum_{r=1}^{R} \left(\varepsilon_r - \varepsilon_0\right)^2}$$

- Involving plausible values:

$$SE_{\bar{\varepsilon}} = \sqrt{\left[\sum_{p=1}^{P}\left(f * \sum_{r=1}^{R}\left(\varepsilon_{r,p} - \varepsilon_{0,p}\right)^2\right) * \frac{1}{P}\right] + \left[\left(1+\frac{1}{P}\right) * \frac{\sum_{p=1}^{P}\left(\varepsilon_{0,p} - \bar{\varepsilon}_{0,P}\right)^2}{P-1}\right]}$$

Example:

- Example taken from New Zealand (TIMSS 1999)
- Difference appears to be statistically significant using standard statistical software!

	Mean	Standard Error	
		SRS	SMP + IMP
Overall Mathematics Score	491	1.5	5.2
Mathematics Score for Girls	495	2.0	5.5
Mathematics Score for Boys	487	2.1	7.6
Difference Between Girls & Boys	8	2.9	8.3

In Summary...

- If you do not take into account the sample and assessment design in your analysis, chances are you will end up with the wrong answer
- If we did not have to do this, we wouldn't!!
- Programs like IDB Analyzer, Data Explorer and Wesvar take sample and assessment design into account

Thank You!

IERI
IEA ETS

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ