Detecting and Treating Errors in Cognitive Assessments and Questionnaires

May 11th-12th, OECD, Paris

Matthias von Davier
National Board of Medical Examiners

NBME®

Response Errors in Assessments

- What are Errors in Assessments?
- Assumptions about Response Processes
- Careless Errors in Questionnaires
- Systematic Errors in Test Responses
- Combining the Evidence
- Outlook

NBME®

Response Errors in Assessments

- What are Errors in Assessments?
- Assumptions about Response Processes
- Careless Errors in Questionnaires
- Systematic Errors in Test Responses
- Combining the Evidence
- Outlook

NBME®

What are Errors in Responses?

- Responses that do not relate to what we want to measure:
  - Cheating (getting answers without work)
  - Training programs(?) Guessing(?)
  - Not paying attention (lack of motivation)
  - Interruptions, nuisance variables,
  - Interviewer helps, or rushes respondents

What are Errors in Responses?

NBME®

What are Errors in Responses?

- May 2016: Some 3,000 students in Thailand must retake university entrance exams after a cheating scam involving cameras and smartwatches was uncovered.
- The sophisticated scam happened at Rangsit University in Bangkok. The university says three people filmed their test papers using tiny cameras embedded in their glasses.
- They then transmitted the images to an outside team, who sent the correct answers to the smartwatches of three other students taking the exams.
- One admitted he was being charged $24,000 (£17,000) to receive the right answers to get into medical school.

(http://www.bbc.com/news/world-asia-36253769)

What are Errors in Responses?

THE CHRONICLE OF HIGHER EDUCATION NEWS OPINION DATA ADVICE JOBS

SECTIONS FEATURED: Lesser-Known Truths About Academe Get the Daily Briefing The Future of Work Report How to Be a Dean

STUDENTS

How Students Cheat in a High-Tech World

OCTOBER 26, 2016

Cheating has always involved elaborate schemes, but now they are increasingly complex and multinational. *Chronicle* reporters look at how students in the United States use internet searches to find surrogates overseas to do their work for them, and how those surrogates can raise their standard of living by writing one paper after another. Cheating technology has also infiltrated

NBME®

http://www.chronicle.com/resource/how-students-cheat-in-a-high-t/6122/

Response Errors in Assessments

- What are Errors in Assessments?
- Assumptions about Response Processes
- Careless Errors in Questionnaires
- Systematic Errors in Test Responses
- Combining the Evidence
- Outlook

NBME®

Response Process Assumptions

Measurement models describe how observed behavior relates to underlying variables of interest:

- How does solving math problems relate to quantitative literacy / numeracy?
- How does answering questions about a text relate to reading literacy?
- How does solving chess problems relate to being a chess master?
- …

Test Theory = Response Model

- Test theories are mathematical models that describe how responses on tests relate to underlying variables (skills / attitudes, …)
- Most theories rest on some fundamental regularities, their model assumptions
- Theories differ in how well these are spelled out formally, and in how strong the models assumptions are

Test Theory = Response Model

1. Monotonicity: The likelihood of a correct response increases with increasing skill,
2. The tasks [questions] have the same order of difficulty for all persons [given skill level],
3. Absence of other Influences: Given a person's skill level, responses vary randomly around a [skill-adjusted, conditional] expectation.

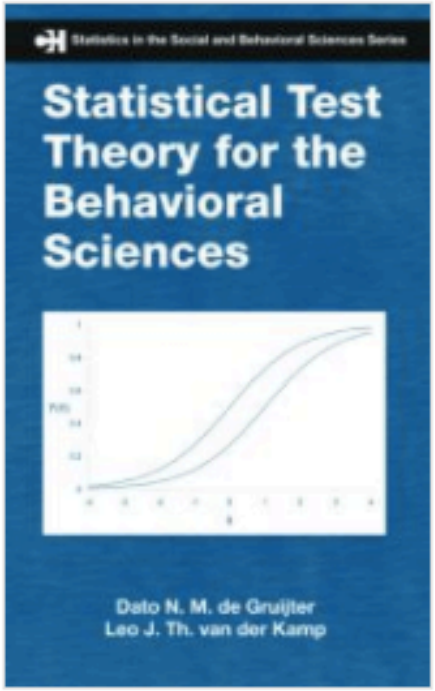
Test Theory = Response Model

Example: Item Response Theory (IRT). This model is the standard approach for complex test designs:

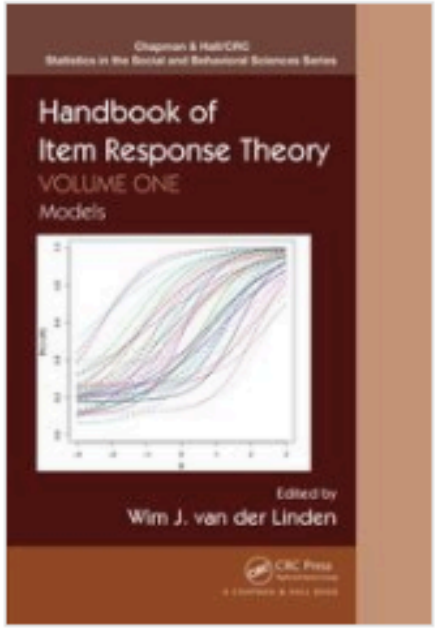
- $P(X=1|u,i) = f(\theta_u,\beta_i)$,
- *e.g. the inverse logit,* $f(\theta_u,\beta_i) = 1/(1+\exp[\beta_i-\theta_u])$

Probability of a correct response is a monotonic function of two variables: Person skill level θ_u and item parameter β_i. Responses are independent given θ_u.

Modern Test Theory is a Part of Applied Statistics

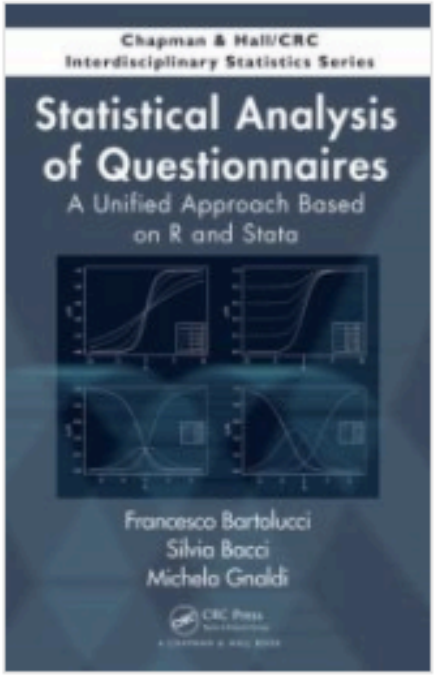


Statistical Test Theory for the Behavioral Sciences



Handbook of Item Response Theory, Volume One

Analysis of Multivariate Social Science Data, Second Edition



Statistical Analysis of Questionnaires

NBME®

Response Errors in Assessments

- What are Errors in Assessments?
- Assumptions about Response Processes
- Careless Errors in Questionnaires
- Systematic Errors in Test Responses
- Combining the Evidence
- Outlook

NBME®

Errors in Questionnaires

NBME®

Errors in Questionnaires

How much do you agree with the following statement? “Compared to others in similar jobs I have a lot of freedom managing my time.”

- Fully agree
- Agree somewhat
- Neither / nor
- Disagree somewhat
- Fully disagree

Errors in Questionnaires

NBME®

Errors in Questionnaires

How much do you agree with the following statement? “The handling of the new 2017 Ferrari California T is outstanding!”

- Fully agree
- Agree somewhat
- Neither / nor
- Disagree somewhat
- Fully disagree

Nuisance Variance in Responses

- Response styles, e.g.:
  - Acquiescence (or opposite, Nay-saying)
  - Extreme-response tendency
  - Mid-point tendency, …
- Other nuisance factors
  - Social desirability, faking good, context effects, order effects, anchoring effects
  - ‘Users of dating sites are 10% taller and weigh 10% less than the general population’

Remedies for Nuisance in Responses

- Response styles, e.g.:
  - Mixture IRT Models (model based clustering)
  - IRT-Tree Models [multinomial choice trees]
  - Predicting / adjusting individual thresholds
- Faking good
  - Over-Claiming Questionnaire
  - Social Desirability Scale
  - Randomized Responses,

Response Errors in Assessments

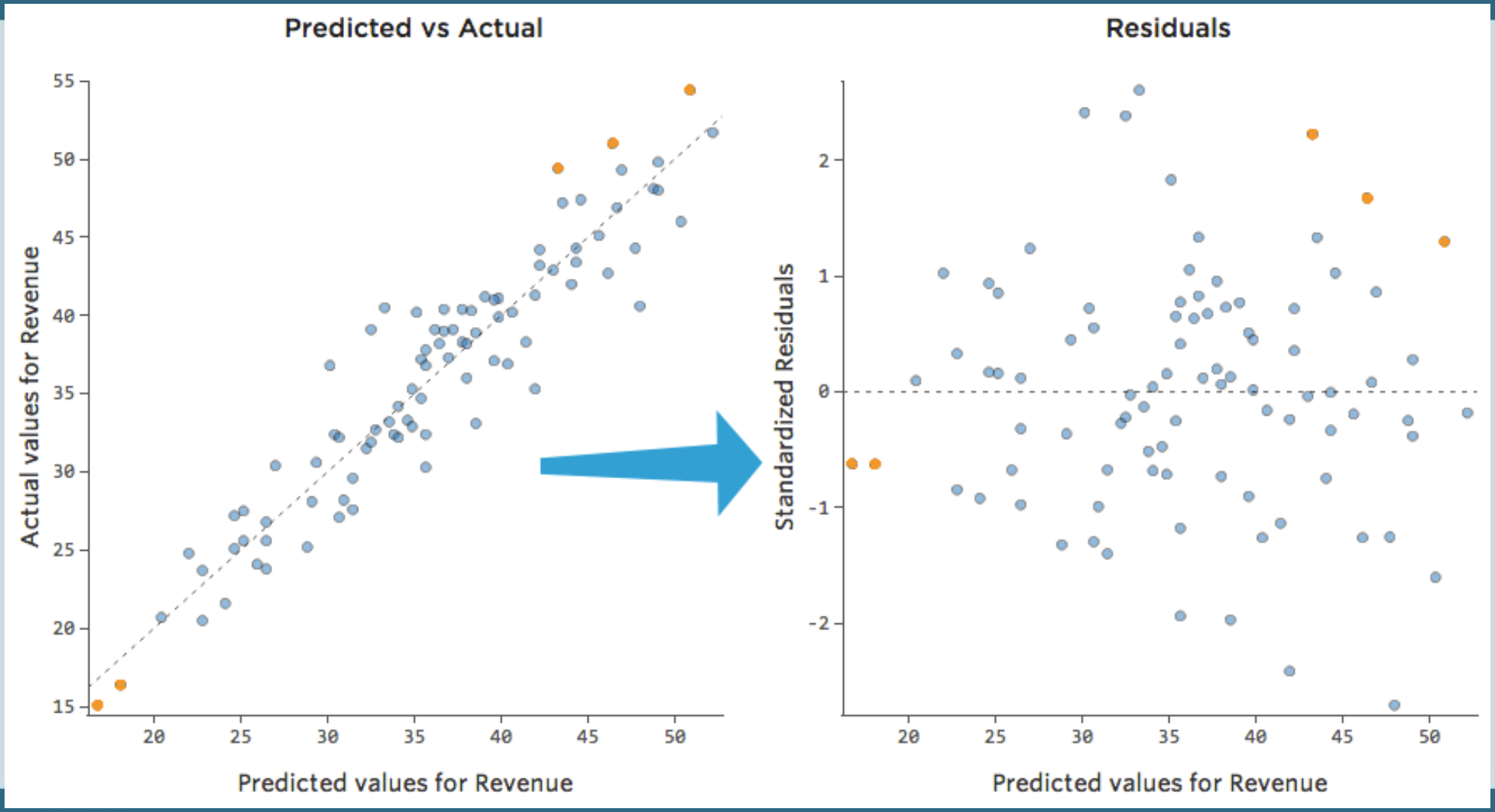
- What are Errors in Assessments?
- Assumptions about Response Processes
- Careless Errors in Questionnaires
- Systematic Errors in Test Responses
- Combining the Evidence
- Outlook

NBME®

Systematic Errors in Tests

- Order of task difficulties is not the same for a certain subgroup compared to all, e.g., some solve hard problems, but not easy problems
- Responses to some tasks do not correlate with other variables (rapid responses uncorrelated with skills and background data)
- Inferences are made based on response likelihoods, or based on response residuals

Residuals in Regression and IRT



NBME®

From: http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/

Response Residuals in IRT

$$\frac{x_{ui} - P(X = 1|u, i)}{\sqrt{P(X = 1|u, i)[1 - P(X = 1|u, i)]}}$$

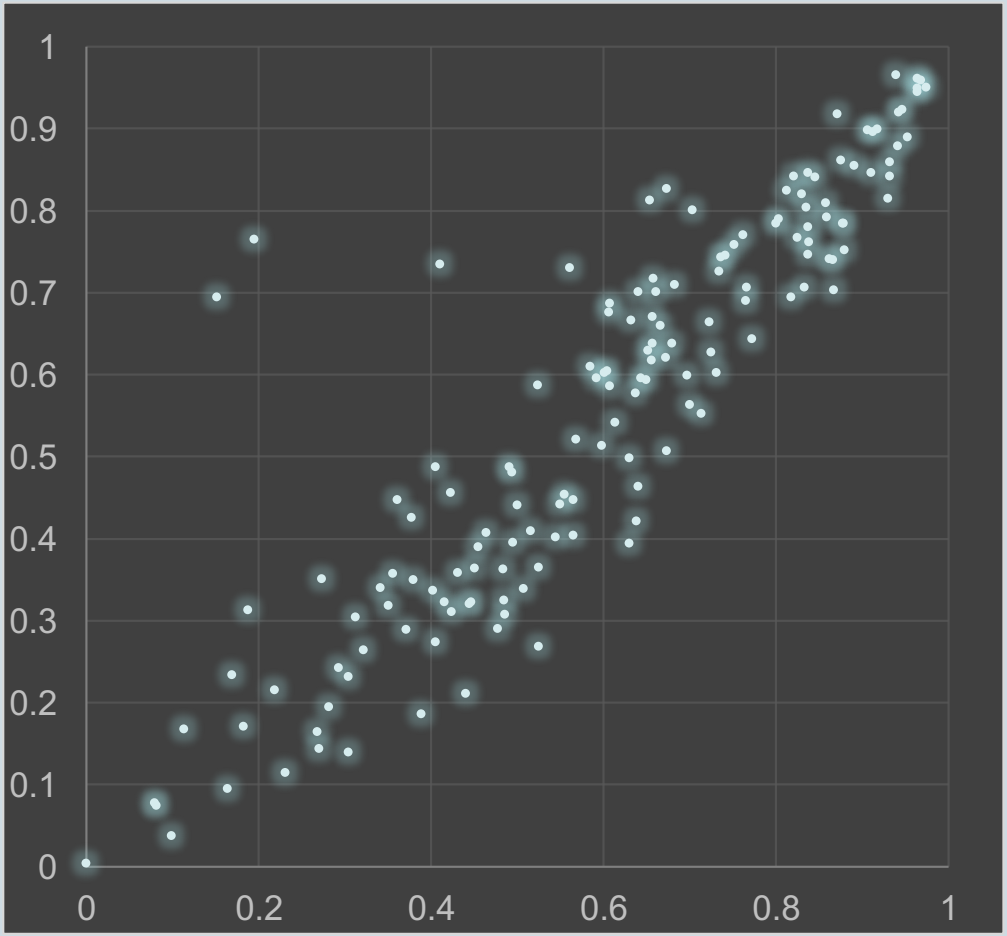
For each response x_{ui}. These can be squared and aggregated across items for person fit, or across respondents per item for item fit

Errors in Test Responses

- High Stakes (Consequential for Students)
  - Test prep ‘heuristics’, repeating the test,
  - Cheating, pay someone else to take a test,…
- Low Stakes (Surveys, PIAAC, PISA, etc.)
  - High stakes for systems?
    - Curb-siding survey questionnaire responses
    - Providing responses, backfilling missing responses
  - Motivation of test takers

Item Probabilities in Groups

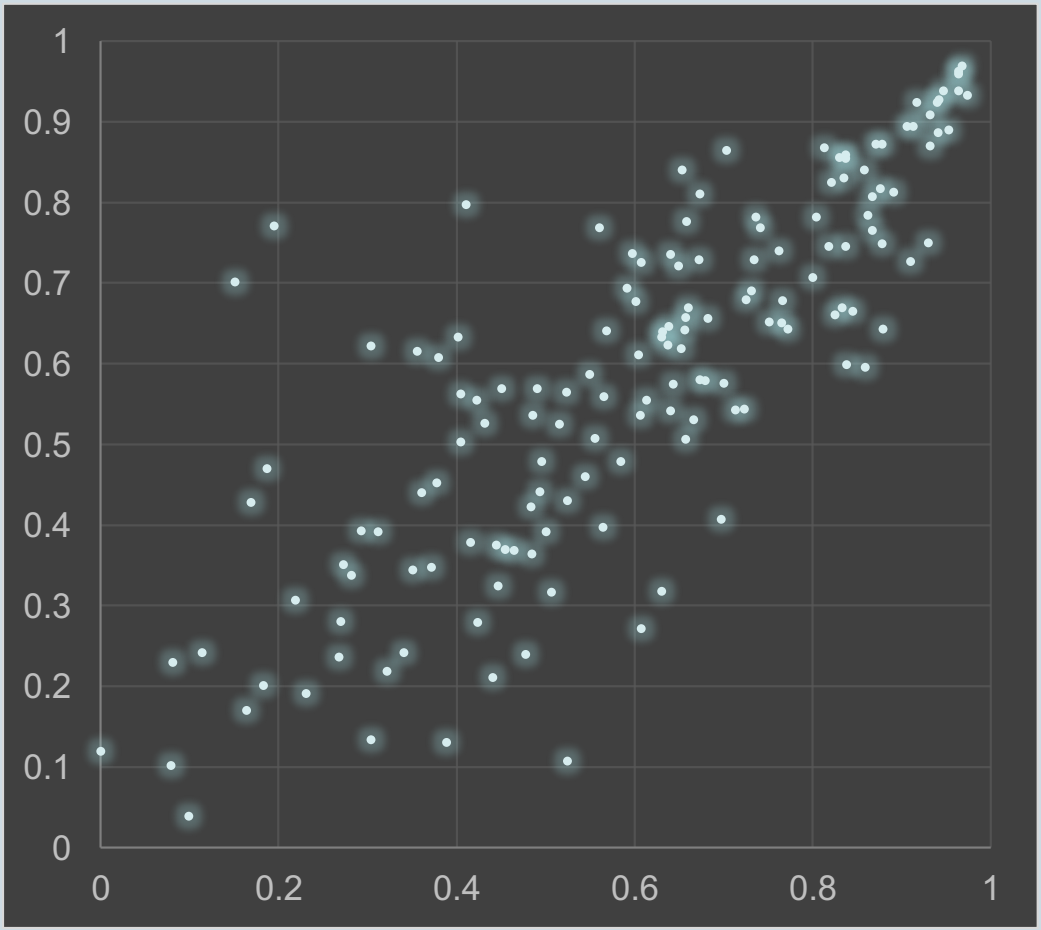
- Great agreement of groups A and B:
  - Relative frequency of solving items is ordered
  - High correlation of group probabilities across the items



NBME®

Item Probabilities in Groups

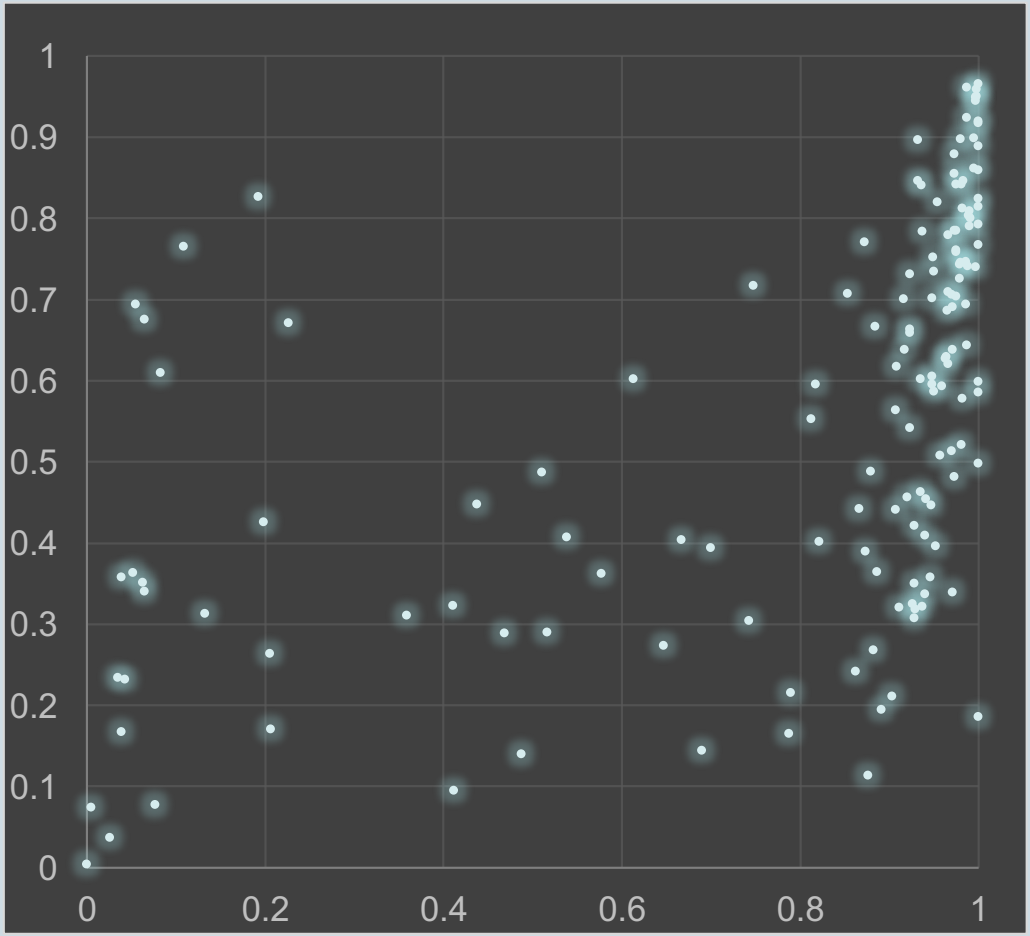
- Good agreement of groups A and C:
  - Relative frequency of solving items is ordered
  - High correlation of group probabilities across the items



NBME®

Item Probabilities in Groups

- Bad agreement of groups A and D:
  - Relative frequency of solving items is in part unordered
  - Low correlation of group probabilities across the items



Comparing Groups vs. Reference

- When the individual groups are compared against multi-group (PIAAC combined) item difficulties (P+), the correlations increase further (for most groups).
- Correlations between group level P+ and overall P+ of 0.9 and higher are common
- The deviant group D reaches only r=0.58

Response Errors in Assessments

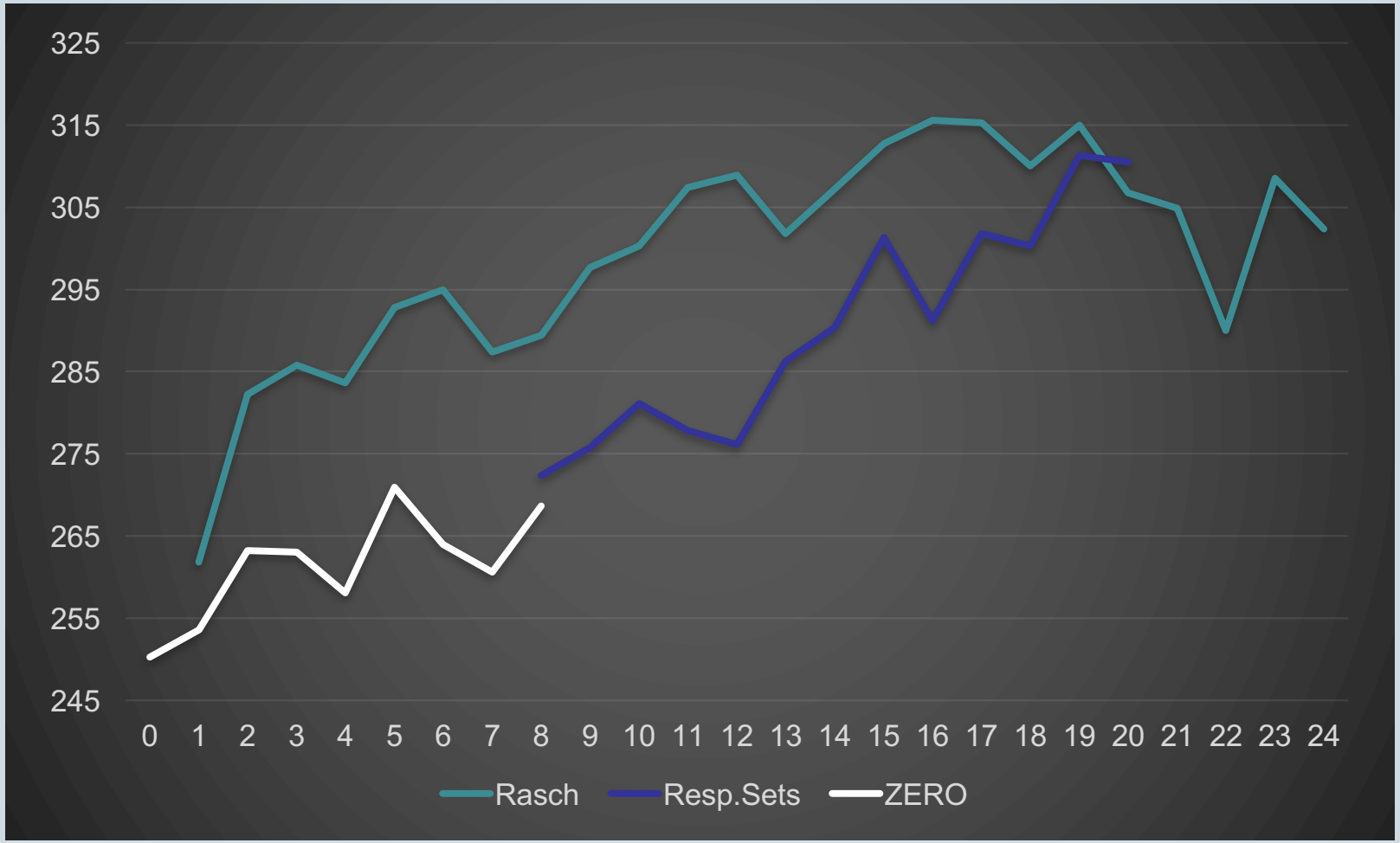
- What are Errors in Assessments?
- Assumptions about Response Processes
- Careless Errors in Questionnaires
- Systematic Errors in Test Responses
- Combining the Evidence
- Outlook

NBME®

Combining the Evidence

- Data from Questionnaires and from Test responses can be combined to examine the effect of nuisance variables on results
- A mixture IRT model to identify 3 classes was used on PIAAC skill use data, and identified:
  - Regular responders
  - Extreme Responders
  - Zero Inflated Responders (no opportunity)

Combining the Evidence



NBME®

Combined Results

- Results on the PIAAC Numeracy Scale were 25-30 points (>0.5SD) lower for the extreme responders compared to regular responders with the same observed skill-use score.
- Do extreme responders not care (motivation), do they exaggerate responses (faking), or do they not have the proficiency to understand skill use items (low reading skills)?

Response Errors in Assessments

- What are Errors in Assessments?
- Assumptions about Response Processes
- Careless Errors in Questionnaires
- Systematic Errors in Test Responses
- Combining the Evidence
- Outlook

NBME®

Outlook

- Computer based testing provides a host of new sources of information and tools
  - Timing data, Process data
  - Response changes, automated response scoring
  - More complex test designs
    - making response copying almost impossible
    - allows tests to be targeted at the respondents level
  - New technologies will facilitate even better QC