Adaptive Assessment and Psychometrics

Teaching for Learning Core Lecture 1

Matthieu Brinkhuis

September 11, 2018

Utrecht University, Information and Computing Sciences
m.j.s.brinkhuis@uu.nl
Today’s lecture:

• What is learning analytics?
• Adaptive Assessment and Psychometrics (Veldkamp, Matteucci, and Eggen, 2011)
  1. Psychometrics (Borsboom, 2006)
  2. Computerized adaptive testing (Glas and Geerlings, 2009)
  3. Adaptive estimation (Klinkenberg, Straatemeier, and Maas, 2011)
Intended learning outcomes

By the end of this lecture, you’ll be able to:

• Understand different types of (adaptive) assessment
• Apply basic understanding of (adaptive) assessment in your research project
• Analyze applications of an adaptive assessment system
Table of Contents

1. Introduction
2. Psychometrics
3. Computerized adaptive testing
4. Adaptive estimation
5. Application
Introduction
Learning analytics

Definition
“the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs” (Siemens, 2010)

- Learning analytics: measurement and evaluation
- Special Interest Group uu.nl/ads
“[...] in the near future it will be possible to continuously assess and store the unfolding life history (trajectory in behavior space) of each individual” (Molenaar, 2004).

It should be possible to create completely individualized educational programs, by teachers that are fully informed on the progress and learning difficulties of each student.

First some building blocks.
Psychometrics
Psychometrics

Psychometrics is psychological measurement, related to areas in psychology and education that is study measurement, testing and assessment.

Started early in the 20th century. Relations with econometrics, biometrics, and generally with data science where it involves measurement.
Latent Variable Theory

Latent variables
We treat a variable as observed when the inference from data structure to variable structure can be made with certainty and as latent when this inference is prone to error. (Borsboom, 2008)

- length?
- age?
- intelligence?
Classical test theory

\[ X = T + E \]  \hspace{1cm} (1)

where \( X \) is the total score, \( T \) the true score and \( E \) the measurement error (Verhelst, 2004a).

However, several difficulties:

- What if test is replaced, do they still measure the same construct?
- How to compare different tests and different populations?
We assume ability $\theta$ to be a latent continuum. If your $\theta$ increases, so does the probability of making an item with difficulty $\delta$ correct. (Rasch, 1960; Verhelst, 2004b)

**Item response function**

We define the following *item response function* for which we can draw an *item response curve*.

$$Pr(X = 1) = \frac{\exp(\theta - \delta)}{1 + \exp(\theta - \delta)}$$  \hspace{1cm} (2)
Advantages of IRT: possibility to carry out meaningful measurement in incomplete designs, for example compare test takers with respect to some proficiency even if they did not all take the same test.
Figure 1: Item response curve of the Rasch model
Now how do we estimate the parameters of such a model? Example of JML likelihood of the Rasch model (Rasch, 1960):

\[ \Lambda = \frac{\prod_n \prod_i \exp(x_{ni}(\theta_n - \delta_i))}{\prod_n \prod_i (1 + \exp(\theta_n - \delta_i))} \]  

(3)

\[ \log \Lambda = \sum_{n}^{N} \theta_n r_n - \sum_{i}^{I} \delta_i s_i - \sum_{n}^{N} \sum_{i}^{I} \log(1 + \exp(\theta_n - \delta_i)) \]  

(4)

where \( r_n \) and \( s_i \) are the sufficient statistics of the person and the item.
Figure 2: Minorization of the log likelihood (van Ruitenburg, 2005)
Computerized adaptive testing
Now if we have all sorts of items with known item difficulties $\delta$, we can administer a test tailored to your ability $\theta$ by selecting the best items (Eggen, 1999).

Note there are different CATs:

- Originally developed for obtaining an efficient estimate of ability (Eggen, 2004)
- Classification (Groen, Eggen, and Veldkamp, 2014)
- Stage-wise testing (Zwitser and Maris, 2015) and more ...
Figure 3: One of the many applications of CATs
What is Computerized adaptive testing? iii

Figure 4: CAT workings visualized (Computer Adaptive Tests 2018)
What is Computerized adaptive testing?

Figure 5: Sequential probability ratio test in (adaptive) classification testing
What is Computerized adaptive testing?

"...of all tests with the same power the SPRT requires on the average fewest observations." (Wald and Wolfowitz, 1948).

What could be issues here with the underlying assumptions?

- model fit & calibration
- parameter change over time
- on average fewest
Adaptive estimation
When do we need adaptive estimation?

- Everyone gets different practice items
- Everyone develops differently
- Feedback (and decision support) in several ways
- System influences itself (complexity)
What is adaptive estimation?

Here we define adaptive estimation as techniques needed to deal with the challenges named before. Several techniques exist, e.g., Kalman filtering (GPS). We focus on tracking instead of modelling.
Elo rating formula for continuous measurement (Elo, 1978)

\[ R_n = R_o + K(W - W_e) \]

- What would be \( W_e \)?
- Elo ratings as self-correcting
- Sequential and simple: powerful

Enormous (and growing) amount of applications next to sports
• looks (Schwartz, 2007)
• animal dominance (Albers and de Vries, 2001)
• programmers (Forišek, 2009)
• plant breeding (Simko and Pechenick, 2010)
• extensions, like TrueSkill (Herbrich, Minka, and Graepel, 2006)
• extensions, using extra information (Klinkenberg, 2014).
Differences with adaptive testing

Two main differences:

1. No assumptions of parameter invariance
2. No standard errors
Application
Figure 6: Mathgarden landing page: the garden.
**Figure 7:** Mathgarden example item, here with multiple choice responses.
The Math Garden in numbers (early 2018):

- over 853,300,000 responses
- over 452,000 K-12 children
- 5,300 schools and many more household subscriptions
- 26 arithmetic domains
- totaling more than 37,000 different items
- rate about 900,000 per school day

Not suitable for classical psychometrics or statistics.
By the end of this lecture, you’ll be able to:

• Understand different types of (adaptive) assessment
• Apply basic understanding of (adaptive) assessment in your research project
• Analyze applications of an adaptive assessment system

Next lecture, we will extend this by:

• Evaluate an application of an adaptive assessment system (possibly your project)
• Create new measures for evaluating adaptive assessment systems
Thank you.


Klinkenberg, S et al. (2011). “Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation”. In: *Computers & Education* 57.2, pp. 1813–1824. DOI:

Molenaar, PCM (2004). “A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever”. In: *Measurement: Interdisciplinary Research & Perspective* 2.4, pp. 201–218. DOI:
10.1207/s15366359mea0204_1 (cit. on p. 7).


