

A simple model and its practical consequences: a tinge of pessimism

Norman Verhelst

National Institute for Educational Measurement (Cito)

Arnhem, The Netherlands

Overview

- The story
- A bit of Reality
- The story revisited

Reference

Verhelst, N.D. & Kamphuis, F.H. (2009). *A Poisson-Gamma Model for Speed Tests*. Measurement and Research Department Reports, 2009-2. Arnhem: Cito

The student monitoring system

- Measurement of individual development
 - Common scale
- Estimation of distribution (norms)
 - Twice per grade (M3, E3, ..., M8)
- Several subjects
 - Arithmetic
 - Reading comprehension
 - **Technical reading**

Three Minute Test

- Three cards containing a list of words.
- Three levels: easy, medium, hard.
- Task: read aloud as many words as you can during **one minute** per card.
- Score per task: number of words read correctly
- Three parallel forms per level:
 - Permutations of each other
 - Strong evidence that the parallel forms were indeed parallel (not considered further)

Example: Three Minute Test (TMT)

- Easy version

- as
- fee
- oom
- uur
- zee
- oor
- ...
- poot (=150)

- Hard version

- banden
- geluid
- tante
- beker
- kuiken
- koffer
- ...
- brandweerwagen
(=150)

Test Design

- Complete design
 - Except for M3: only two cards
- Administration order: easy, medium, hard for everybody
- Sample sizes: about 900 per half-year group M3, E3, ..., M8

Models

- Measurement model

- What is the relation between the (latent) ability and the test performance?

$$f(\text{Obs} \mid \theta)$$

- Structural model

- The distribution of the latent ability in one or more populations? (M3, E3, M4, ..., M8)

$$g(\theta \mid \text{M3}), \quad g(\theta \mid \text{E3}), \dots$$

Measurement model: Poisson (1)

x_{vi} : observation (number read/number correct)

v : student index

i : task index

$$P(x_{vi}; \alpha) = \frac{\alpha^{x_{vi}}}{x_{vi}!} e^{-\alpha}, \quad (x_{vi} = 0, 1, 2, 3, \dots)$$

Measurement model: Poisson (2)

$$P(x_{vi}; \alpha) = \frac{\alpha^{x_{vi}}}{x_{vi}!} e^{-\alpha}, \quad (x_{vi} = 0, 1, 2, 3, \dots)$$

$$\alpha = \alpha_{vi} = \tau_i \times \theta_v \times \sigma_i$$

τ_i : time limit (in minutes)

σ_i : easiness of task i (dimensionless)

θ_v : ability (#subtasks/minute)

Parameter Estimation

- Task parameters (σ_i): JML or CML
- Person Parameters: ML
 - Conditionally unbiased: $E(\hat{\theta} | \theta) = \theta$

Person parameters

$$\hat{\delta}_v = \sum_i d_{vi} \tau_i \hat{\sigma}_i$$

δ is the corrected reading time (weights: σ_i)

$$\hat{\theta}_v = \frac{s_v}{\hat{\delta}_v} \quad E(\hat{\theta}_v | \theta) = \theta$$

$$SE(\hat{\theta}_v) = \sqrt{\frac{\theta_v}{\delta_v}} \approx \sqrt{\frac{\hat{\theta}_v}{\hat{\delta}_v}} = \frac{\sqrt{s_v}}{\hat{\delta}_v}$$

Two step procedure

- Estimate the task parameters σ_i
 - JML = CML
- Estimate latent distribution while fixing the task parameters at their CML -estimate

Advantage

If X_1 and X_2 indep. Poisson with parameters α_1 en α_2 ,
then $X_1 + X_2$ is Poisson distributed with parameter $\alpha_1 + \alpha_2$

$$s_v = \sum_i s_{vi} \square P[\theta_v \times \sum_i \tau_i \sigma_i] = P(\theta_v \delta)$$

Structural model:
distribution of reading speed (θ)

$$g(\theta; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta)$$

$$E(\theta) = \frac{\alpha}{\beta} \quad \text{Var}(\theta) = \frac{\alpha}{\beta^2}$$

Marginal distribution of the sum score s

$$f(s) = \int_0^{\infty} P(s | \theta) \times g(\theta) d\theta$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^{\infty} \frac{(\delta\theta)^s}{s!} e^{-\delta\theta} \times \theta^{\alpha-1} e^{-\beta\theta} d\theta$$

Negative Binomial (Gamma-Poisson)

$$f(s) = \frac{\Gamma(\alpha + s) \delta^s \beta^\alpha}{s! \Gamma(\alpha) (\delta + \beta)^{s+\alpha}}$$

$$p = \frac{\delta}{\delta + \beta} \qquad 1 - p = \frac{\beta}{\delta + \beta}$$

$$f(s) = \frac{\Gamma(\alpha + s)}{s! \Gamma(\alpha)} p^s (1 - p)^\alpha$$

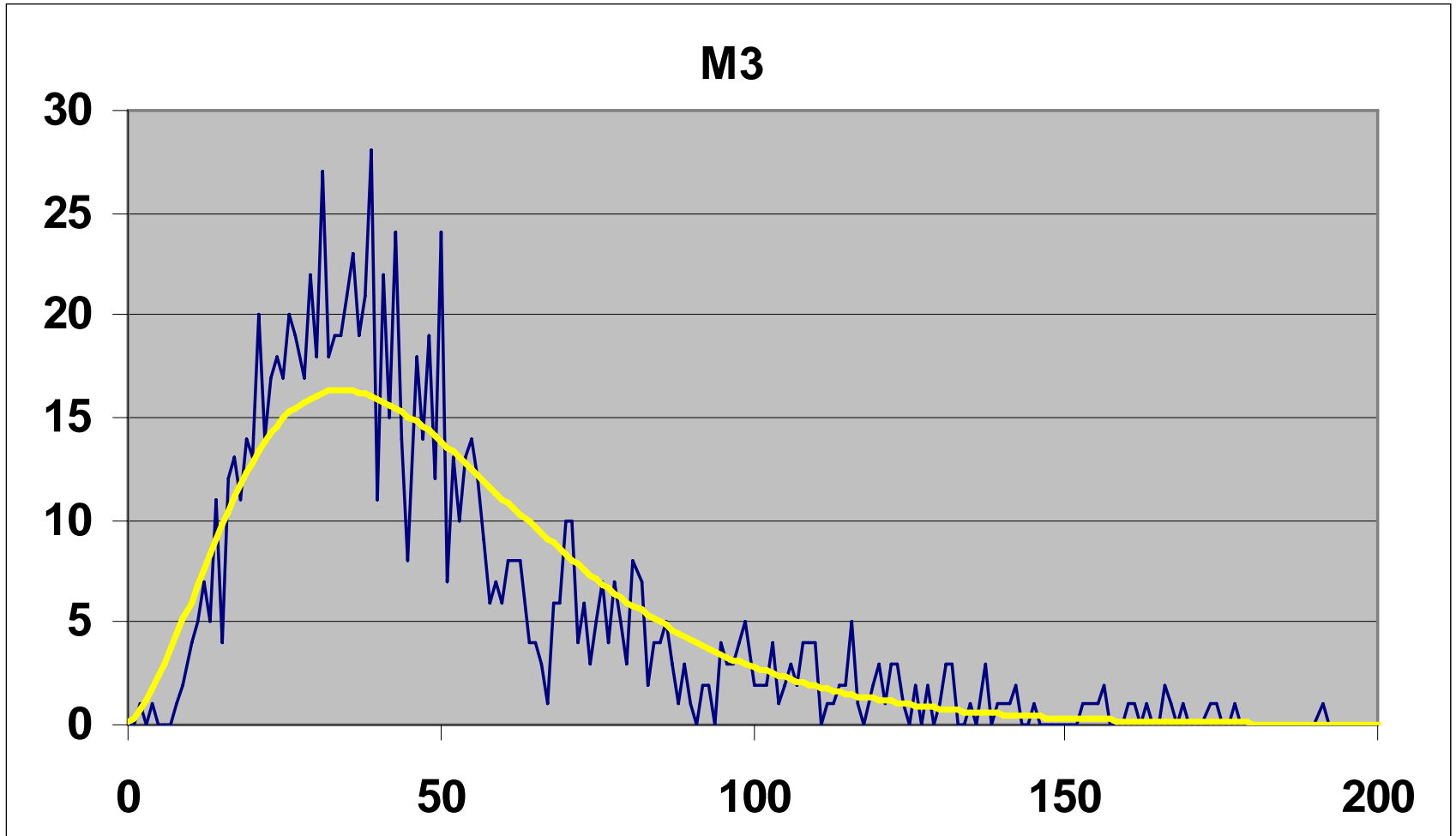
Negative binomial

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$$

$$\frac{\Gamma(\alpha + s)}{\Gamma(\alpha)} = \frac{\cancel{\Gamma(\alpha)}}{\cancel{\Gamma(\alpha)}} \times \prod_{j=0}^{s-1} (\alpha + j)$$

$$f(s) = \frac{\prod_{j=0}^{s-1} (\alpha + j)}{s!} p^s (1-p)^\alpha$$

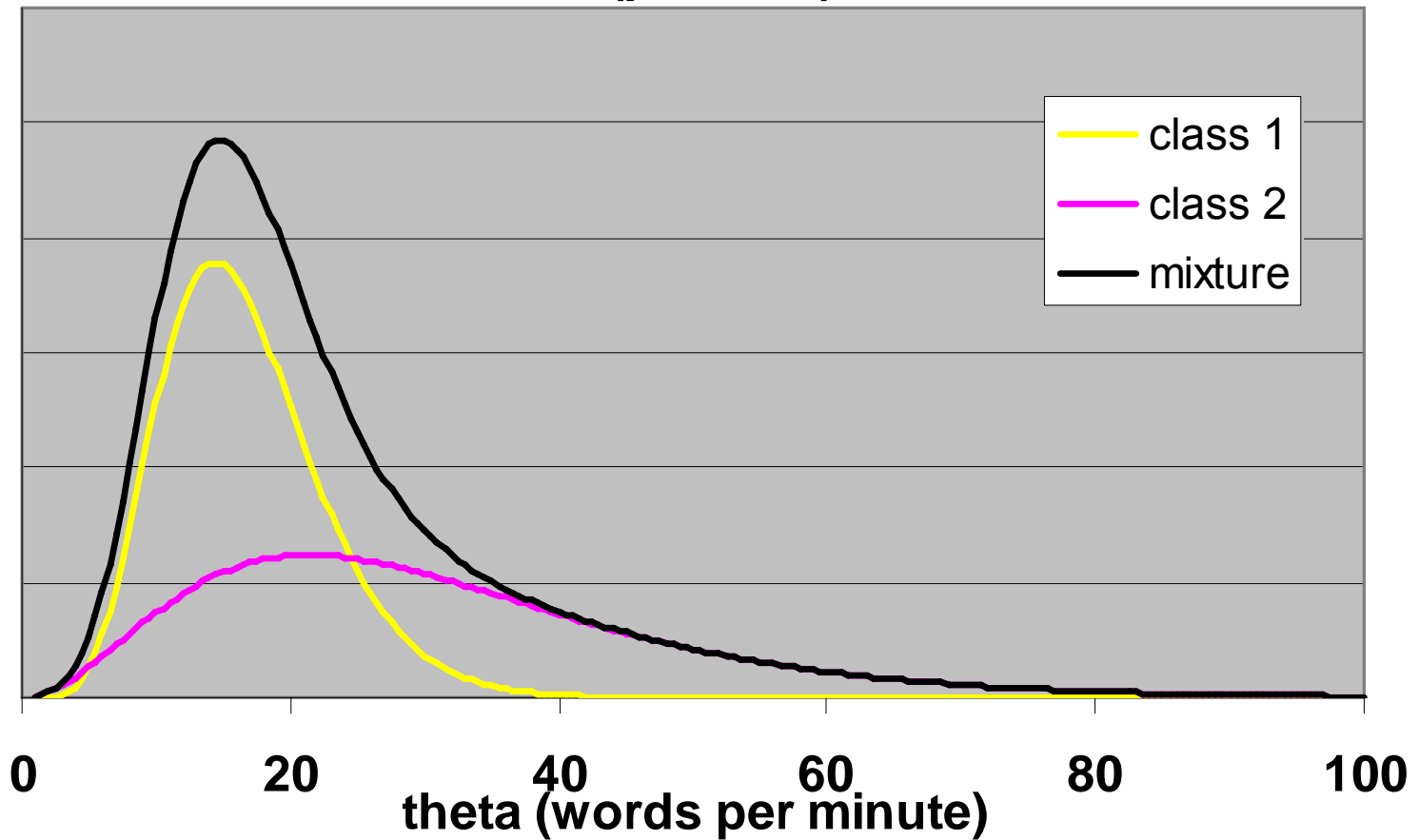
Validation (TMT)



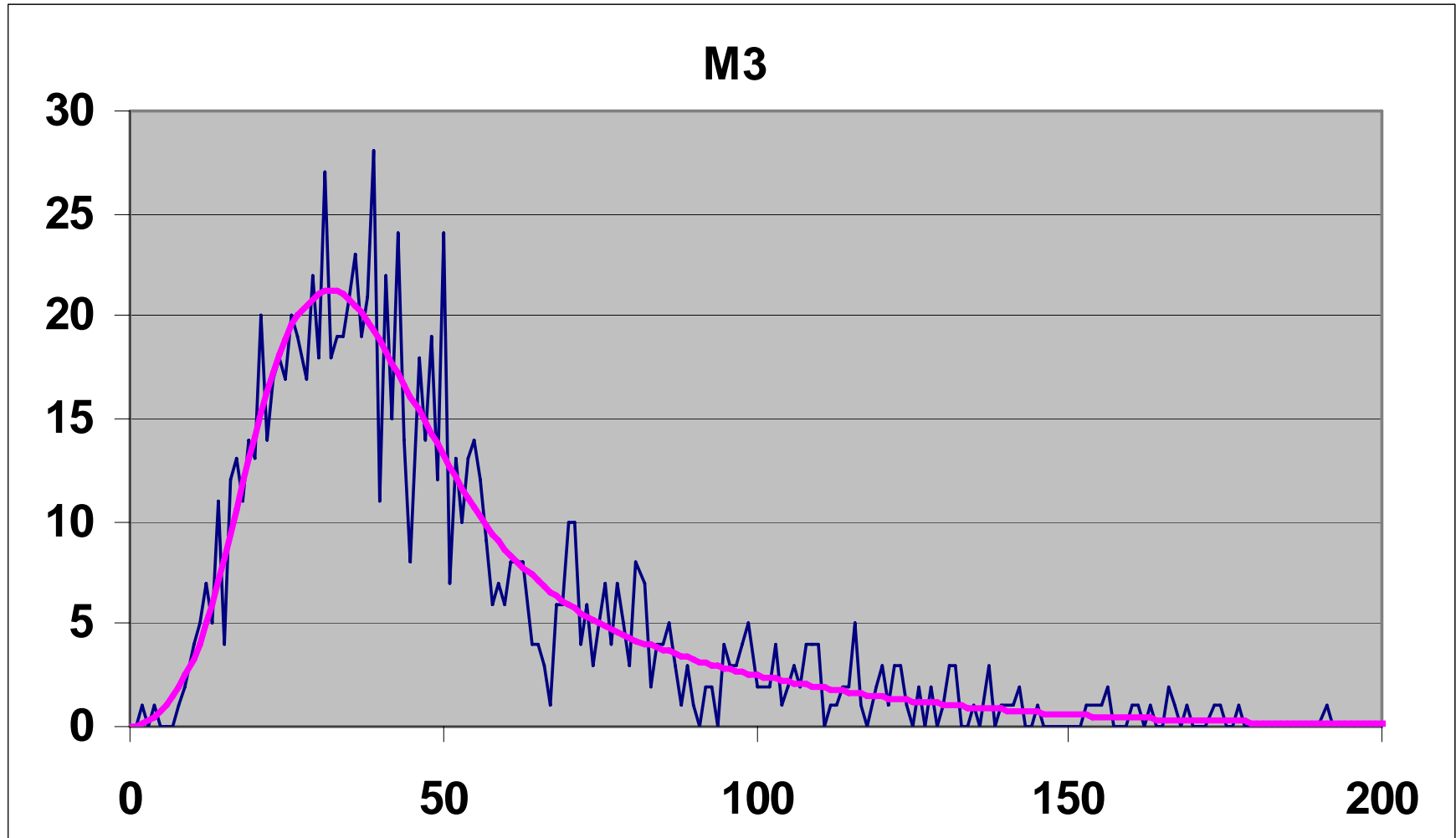
Latent class model

- Population consists of two latent classes of size π and $1 - \pi$ respectively
- The latent variable is gamma distributed in each class
- Parameters
 - π
 - α_1 en β_1
 - α_2 en β_2
- EM-algorithm

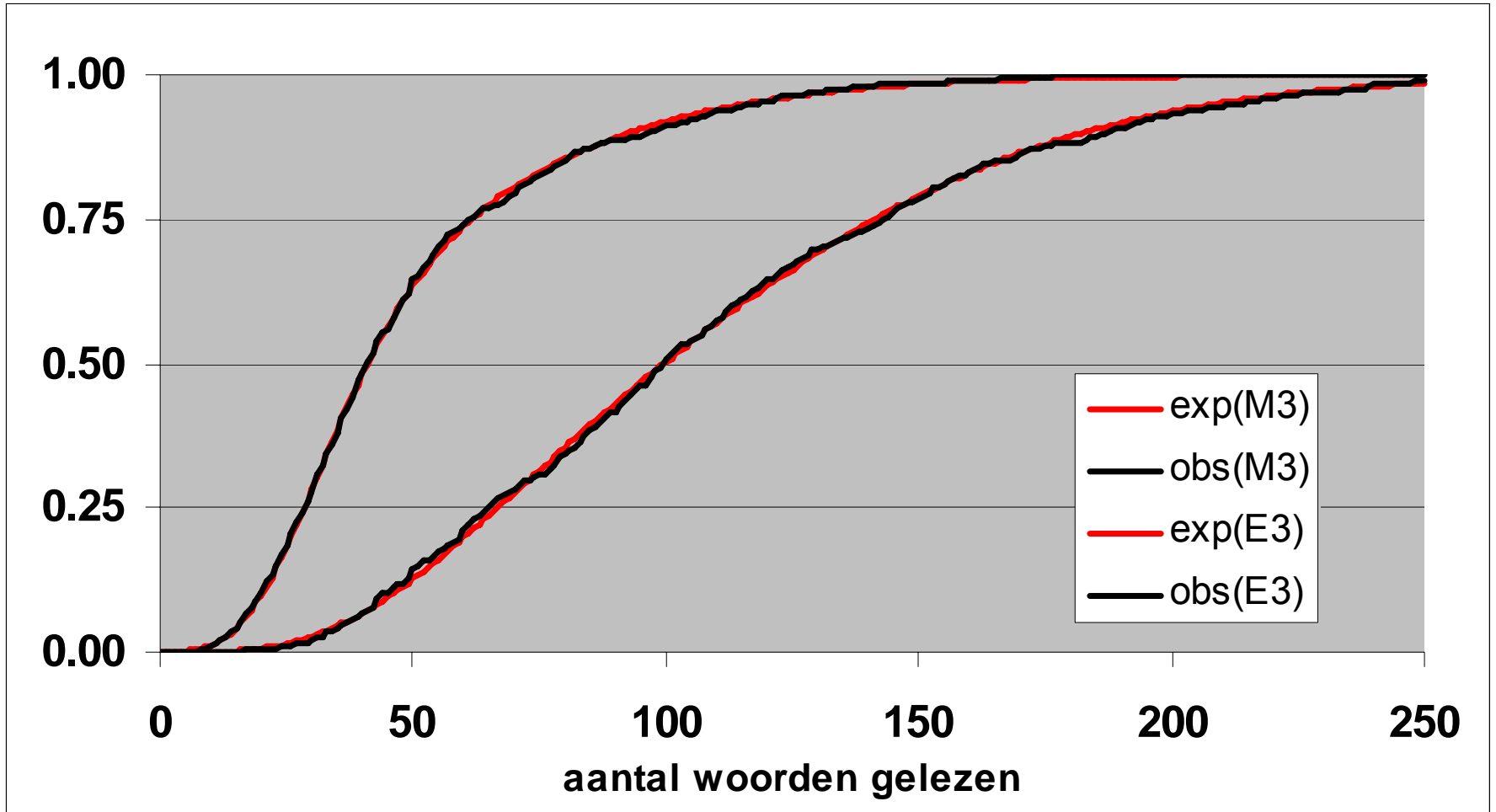
M3 ($\pi = 0.54$)



Validation (TMT)



Validation (TMT)



A Bit of Reality

- Reporting in the student monitoring system:
 - Administer the test, **either completely or incompletely**
 - Estimate θ
 - Report on a five-point scale, using the norms
→ → →

Percentile	Report
$\leq P10$	E
P11 – P25	D
P26 – P50	C
P51 – P75	B
$> P75$	A

A Bit of Reality (2)

- TMT was used in a survey like research, shortly after its publication
- An unacceptably high proportion of D and E reports was observed
- A formal complaint was sent to Cito
- Details showed that
 - High D and E rates were observed in the lower grades
 - Only cards 2 and 3 had been used

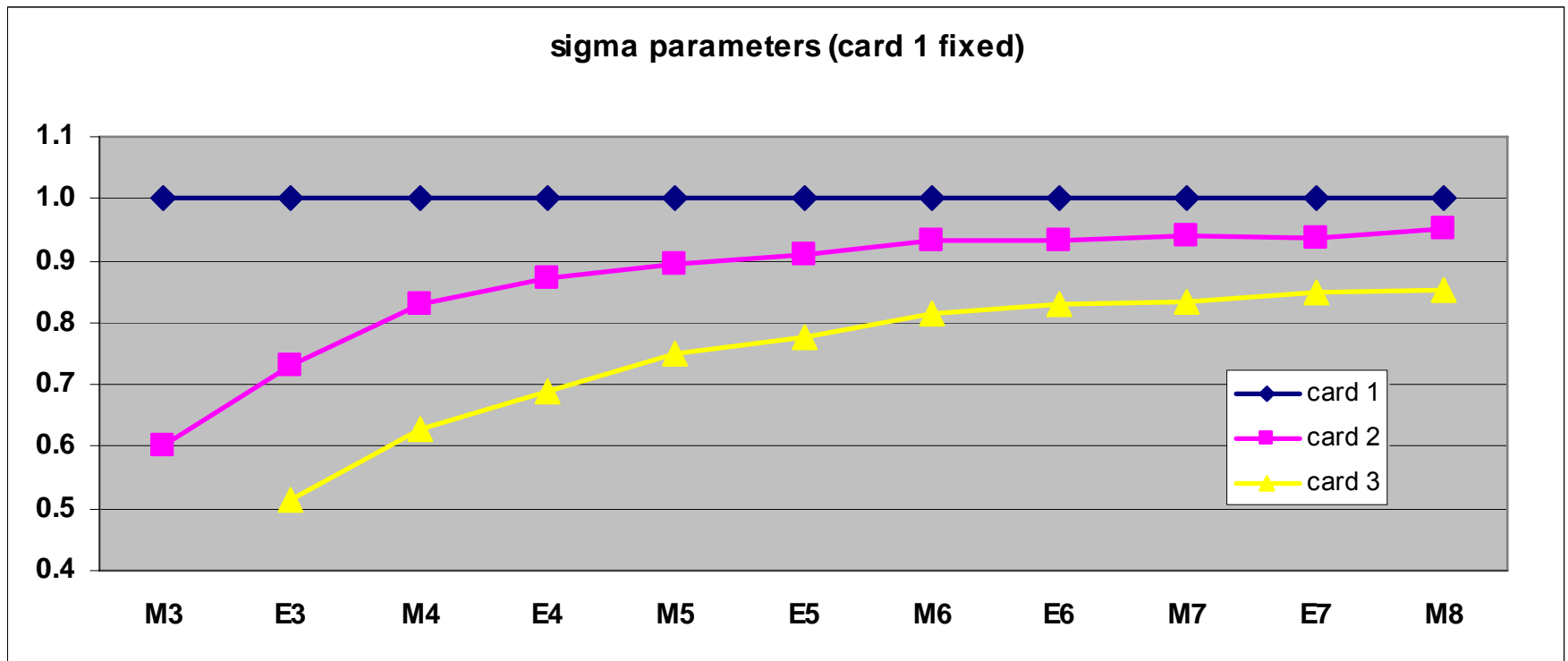
The story revisited

- What was overlooked?
 - Local independence?
 - Something else, essential but having little impact on the validation procedure used?
- Possible answers (all equivalent):
 - No specific objectivity
 - DIF
 - The ratio σ_i/σ_k not constant across grades

Research on DIF

- Items can have different (psychometric) characteristics across grades
- Some linking is needed, otherwise the concept of a single scale across grades would be empty
- So, let's try one:
 - $\sigma_1 = 1$ is constant across grades
 - The σ -parameters for cards 2 and 3 are free to vary across grades

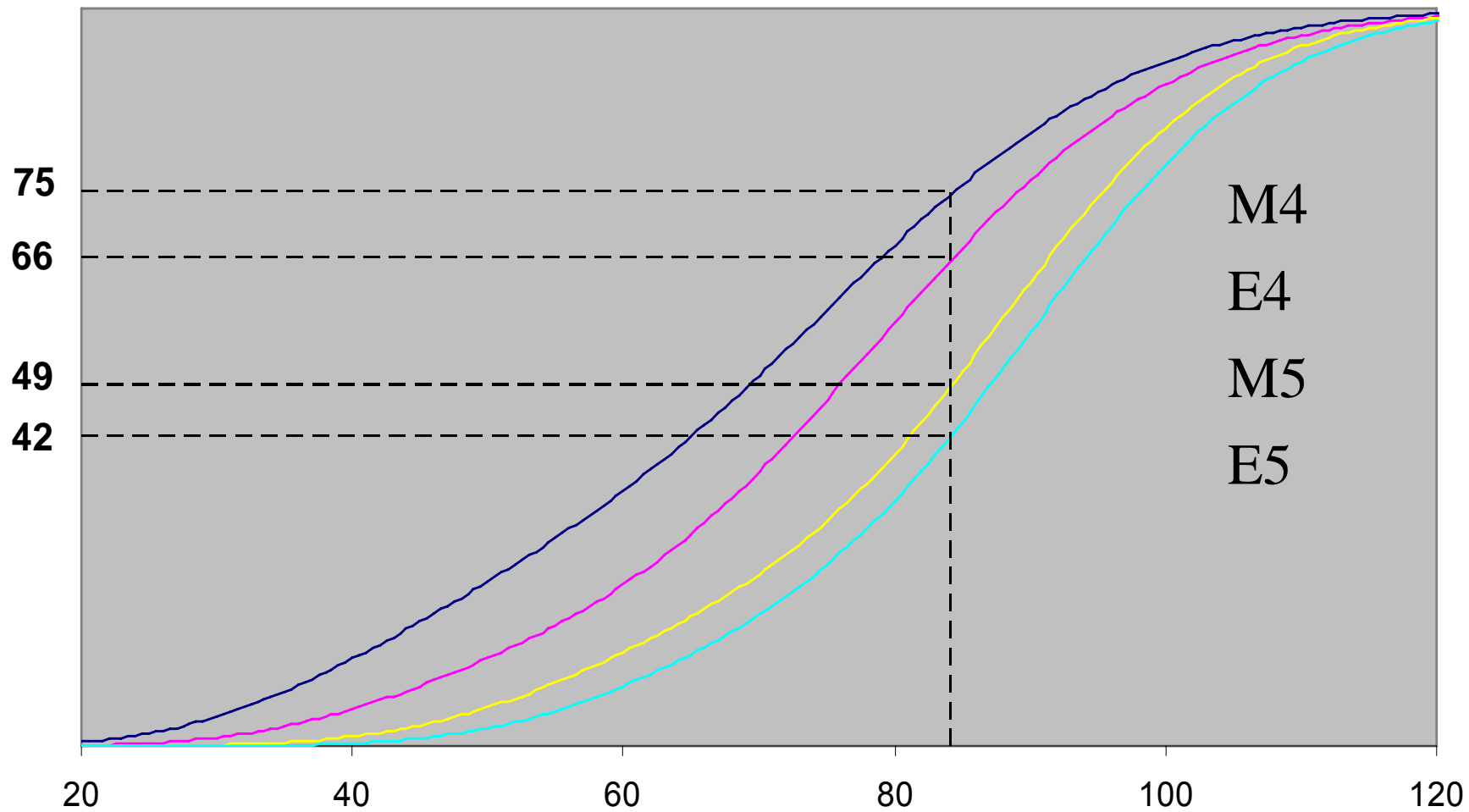
A clear demonstration of DIF



Validation

- Was even better than under the original model.
- So, we are done, and have nice possibilities of comparison

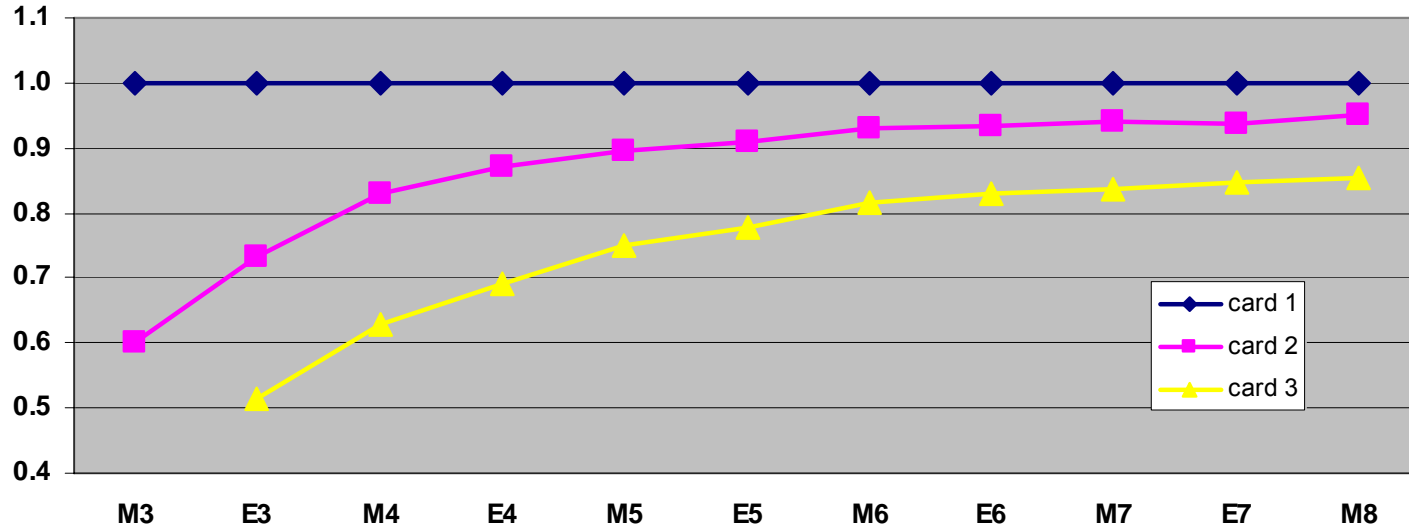
delta-1 fixed to 1



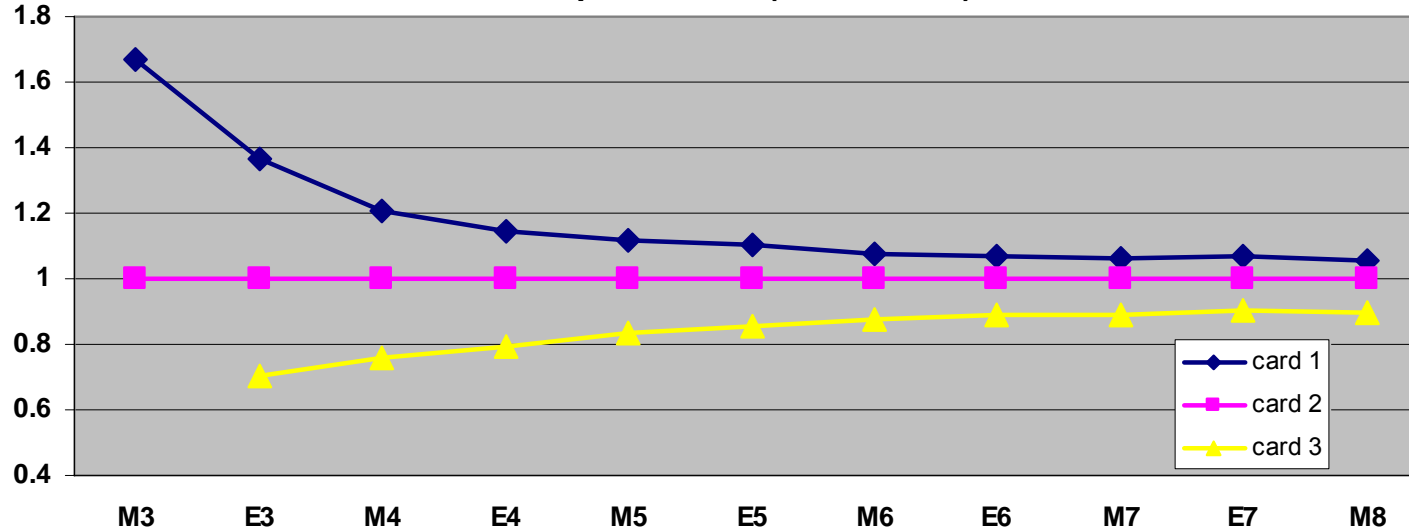
Other possibilities

- We can choose card 2 as the anchor across grades.
- Indicate the new scale by an *
- The new outcomes are completely predictable from the old ones.
- The validation procedure (prediction of the observed score distribution) yields identical results under both models
- But...

delta parameters (card 1 fixed)



delta parameters (card 2 fixed)



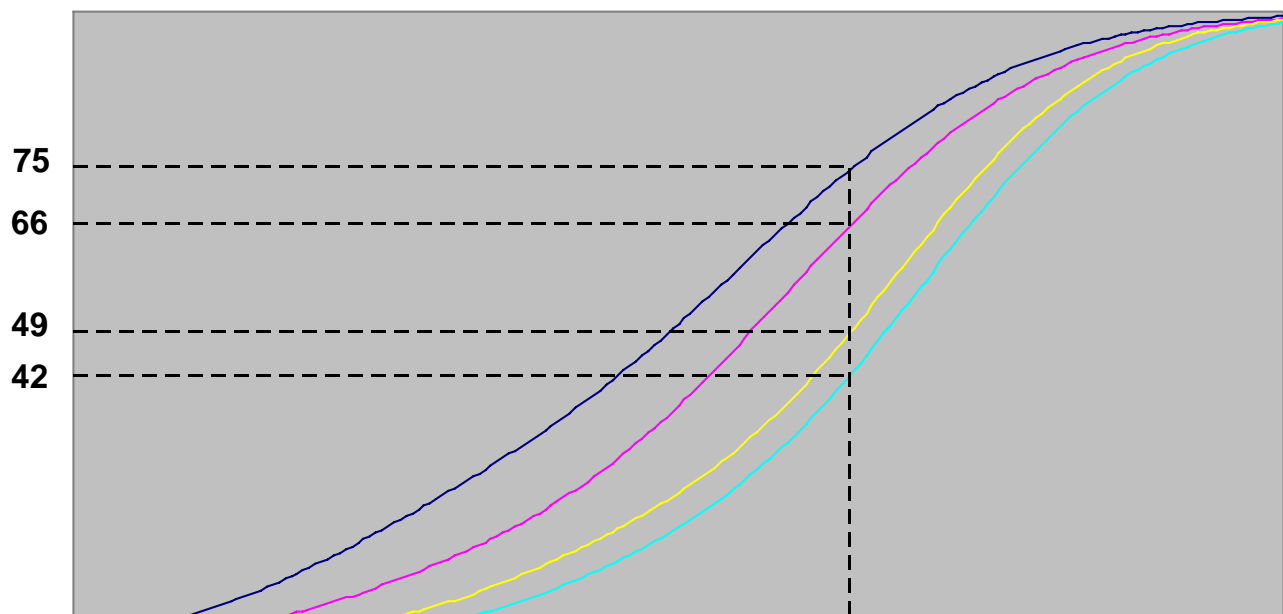
Sigma parameters card i in group g : $\sigma_{ig}^* = \frac{\sigma_{ig}}{\sigma_{2g}}$

Mixture parameter in group g : $\pi_g^* = \pi_g$

Shape parameter in group g : $\alpha_g^* = \alpha_g$

Scale parameter in group g : $\beta_g^* = \frac{\beta_g}{\sigma_{2g}}$

delta-1 fixed to 1



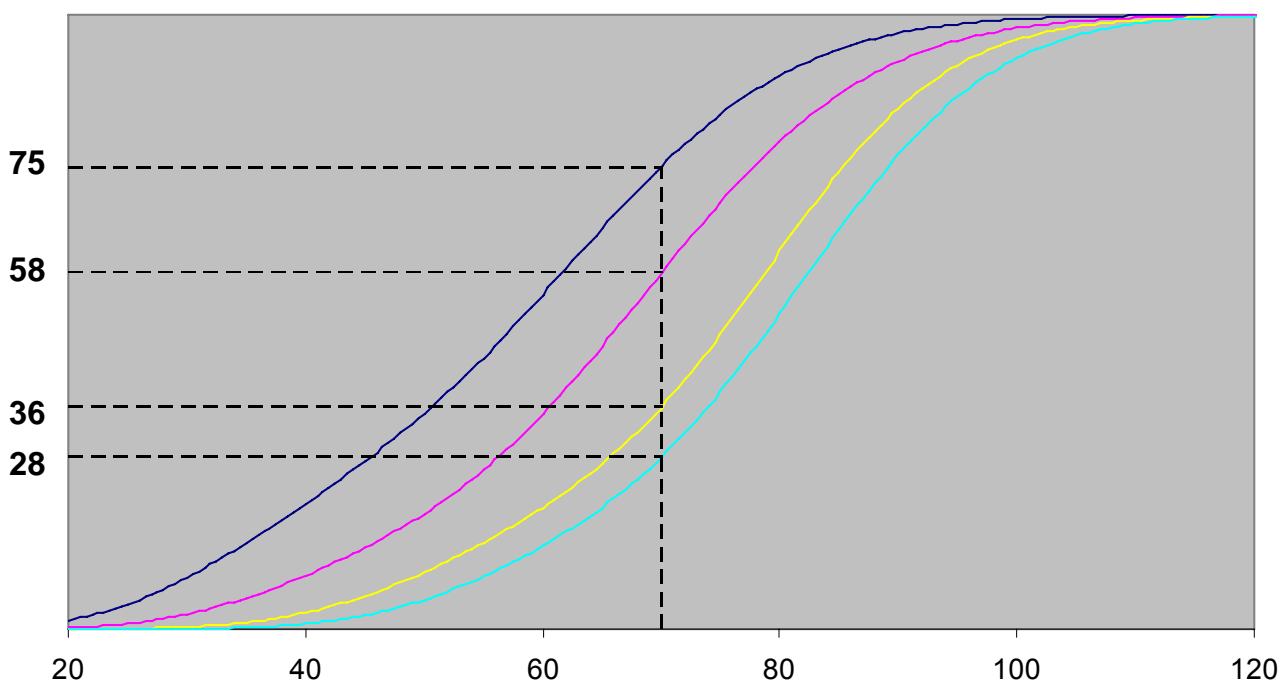
M4

E4

M5

E5

delta-2 fixed to 1



75

58

36

28

20

40

60

80

100

120

P75 in M4 corresponds to

	θ	θ^*
E4	P66	P58
M5	P49	P36
E5	P42	P28

Therefore

- The scale θ^* is **not** a monotone transformation of the scale θ .
- Since the choice of the anchor is arbitrary, the outcome is arbitrary as well, or, in other words, θ means nothing
- What are we left with?

What we are left with

- A set of three cards (produced by Cito, but otherwise arbitrary)
- An observed score distribution for all grades, reasonably stable (sample size) and representative (sampling scheme)
- An ingenious technique to smooth irregularities in the observed distribution
- And that is all!

Sorry, but

- We could have obtained the same result 60 years ago (Gulliksen, 1952)

Tomorrow's full title

DIF and beyond: the Pisa case
a tinge of optimism