Developing and Evaluating a Computerized Adapted Mathematical Test

Jang Schiltz

University of Luxemburg

jang.schiltz@uni.lu

The context

- New orientation procedure for the transition from primary to secondary school
- Mathematical test from November 1996
- Not many multiple choice items, some geometrical construction items
- 81 items
- Population: 3590 mostly 12 year old pupils
- Had to eliminate 7 of these items because they did not verify the assumption of local independence; hence I kept 74 items for a more thorough analysis.

Item calibration

- Done with BILOG-MG
- Because of the lack of multiple choice items, I tried the 1 respectively 2-parameter IRT model
- Likelihood ratio test indicates that twe models are valid
- But the 1-parameter IRT model 26 items with good item fit versus 67 with the 2-parameter IRT.

- Item Response Function:

$$P_j(\theta) = \frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}}$$

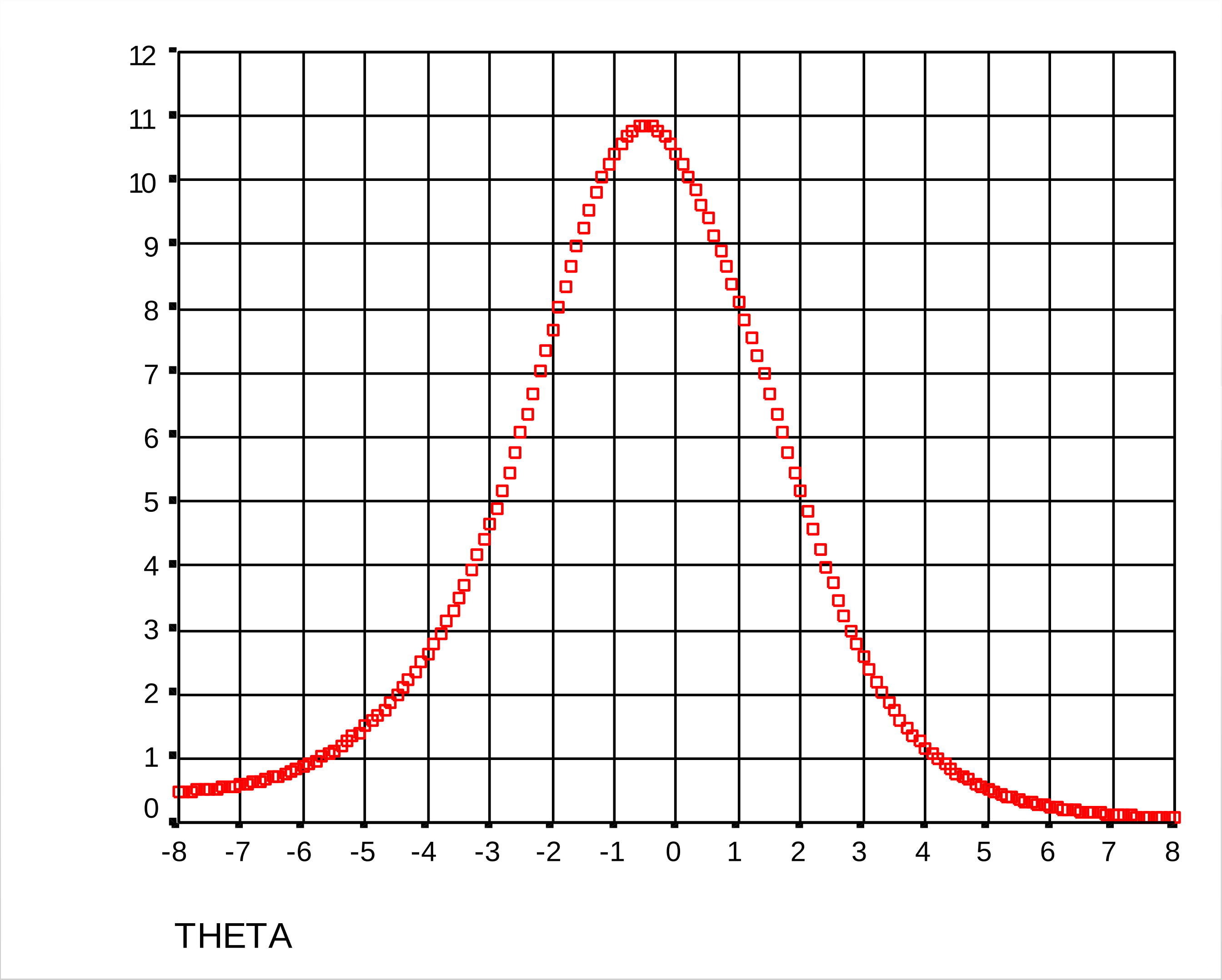
a_j item slope

b_j item treshold

Checking the invariance property

- The next step consisted in eliminating the items that did not satisfy the invariance property.
- I seperated the examinees population into two groups, by splitting them at the median of the mathematical ability.
- I calibrated the items separately for the two groups keeping the same distribution and eliminated the items for which the difference of the thresholds in the two groups was larger than 1,96 σ .
- I kept a final item Pool of 63 items.

Test information function



Construction of the CAT

- I used the multimedia-programming platform Quest Net+ for Windows™.
- Initialisation: An item with mean difficulty (0,078) and not too strong discrimination power (0,618).

Estimation of the math skill

- I used the Bayesian modal estimation method, supposing a normal distribution of θ.
- Posterior function of likelihood :

$$L(\theta) = \frac{\exp\left[\sum_{i=1}^{k-1} a_{j_i}(b_{j_i} - \theta)(1 - u_{j_i})\right]}{\prod_{i=1}^{k-1} 1 + \exp -a_{j_i}(\theta - b_{j_i})} g(\theta).$$

Maximum value for

$$-\sum_{i=1}^{k-1} a_{ji}(1 - u_{ji}) + \sum_{i=1}^{k-1} \frac{a_{ji} e^{a_{ji}(b_{ji} - \theta)}}{1 + e^{a_{ji}(b_{ji} - \theta)}} - 2\theta = 0,$$

Estimation of the math skill

```
float algo (float x_0)
   {float res, a, b;
     a=x_0-2.0;
     b=x_0+2.0;
     do {
               if (funct(a)*funct(b)<0)
               {if (funct(a)*funct((a+b)/2.0)<0)  b=(a+b)/2.0;
               else a=(a+b)/2.0;}
               else a=a-1.0, b=b+1.0;
          }
     while (b-a>0.005);
     res=a;
   return res;}
```

Choice of the next item

- Maximum information strategy:
- Item information function: $I(\theta) = a_j^2 P_j(\theta) \;\; 1 - \mathrm{P_j}(\theta)$.

```
for (i=1 ; i<=63 ; i++)
        {info[i-1]=I(theta, i);}

for (i=1 ; i<=nom ; i++)
    {info[ens[i-1]-1]=0.0;}

float sup;
        int item=1;
        sup=info[0];
        for (i=2; i<=63 ; i++)
        {if (info[i-1]>sup) {sup=info[i-1];
                        item=i;}}
```

Empirical Evaluation Study

- We constructed an experimental form of our CAT and a parallel paper-and-pencil test and presented the two forms to a representative sample of 123 pupils belonging to the first year of a classical secondary school.

- To neutralize the effect of the order of presentation, the participants were distributed by randomization to two experimental groups. Group A began with the CAT and group B with the paper-and-pencil test.

- Two months later, the experimental conditions were inverted.

Criterion one: Performance scores and success ratio

- There appears no significant difference, nor between the means of the CAT note ($M = 0.55$, $SD = 0.47$) and paper-and-pencil scores ($M = 0.61$, $SD = 0.65$) ($t = 1.578$, $p > 0.05$), nor between the failure ratio, which is 13.8% for the paper-and-pencil condition and 8.9% for the CAT ($\chi^2 = 2.65$, $p > 0.05$).

- Thus, the CAT is comparable to a classical paper-and-pencil test by its ability to measure the performances of pupils.

Criterion two: Ability to rank pupils according to their level of performance

- The correlations between the scores on the CAT and on the paper-and-pencil test are significant ($r = 0.593$, $p < 0.01$), but medium-sized.

- ## Criterion three: Transfer and apprenticeship from one experimental condition to the other

- CAT (subgroup A: $M = 0.5674$, $SD = 0.408$; subgroup B: $M = 0.5420$, $SD = 0.497$; $t = 0.285$; $p > 0.05$)
- Paper-and-pencil test (subgroup A: mean = 1.0093, $SD = 0.558$; subgroup B: mean = 0.4668, $SD = 0.560$; $t = 5.100$; $p < 0.001$).
- Thus, there has been a positive effect of apprenticeship from the CAT to the paper-and-pencil situation and not vice versa

Criterion five: Relations with general intelligence

Correlations

		NOTE	NOTEPAP	GL	QU3_4	QU7_10	QU13_14	QU15
NOTE	Pearson	1	.493**	.362**	.337**	.393**	.035	.187*
	Sig. (2-tailed)	.	.000	.000	.000	.000	.705	.039
NOTEPAP	Pearson	.493**	1	.319**	.280**	.379**	.171	.131
	Sig. (2-tailed)	.000	.	.000	.002	.000	.060	.149
GL	Pearson	.362**	.319**	1	.541**	.722**	.476**	.338**
	Sig. (2-tailed)	.000	.000	.	.000	.000	.000	.000
QU3_4	Pearson	.337**	.280**	.541**	1	.461**	.353**	.197*
	Sig. (2-tailed)	.000	.002	.000	.	.000	.000	.030
QU7_10	Pearson	.393**	.379**	.722**	.461**	1	.406**	.119
	Sig. (2-tailed)	.000	.000	.000	.000	.	.000	.192
QU13_14	Pearson	.035	.171	.476**	.353**	.406**	1	.280**
	Sig. (2-tailed)	.705	.060	.000	.000	.000	.	.002
QU15	Pearson	.187*	.131	.338**	.197*	.119	.280**	1
	Sig. (2-tailed)	.039	.149	.000	.030	.192	.002	.

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Correlations between the CAT and paper-and-pencil test scores and the general intellectual efficiency test scores.

NOTE: CAT score

NOTEPAP: paper-and-pencil score

GL: general intellectual ability

QU3_4: reasoning

QU7_10: spatial representation

QU13_14: perceptual speed

QU15: computational skill