Optimal Calibration Designs for Computerized Adaptive Testing

Angela Verschoor, Cito
angela.verschoor@cito.nl

- Development phases of a test
- Some common designs
- Homogeneous designs
- Simulations & live results
- Conclusions

Development of a CAT

- Specifications: Purpose, Blueprint, etc.
- Item development
- Gathering data about items: parameters
- Design of the CAT
- Test administration

Gathering Data

- Estimation of item parameters through a sample of the population: errors!

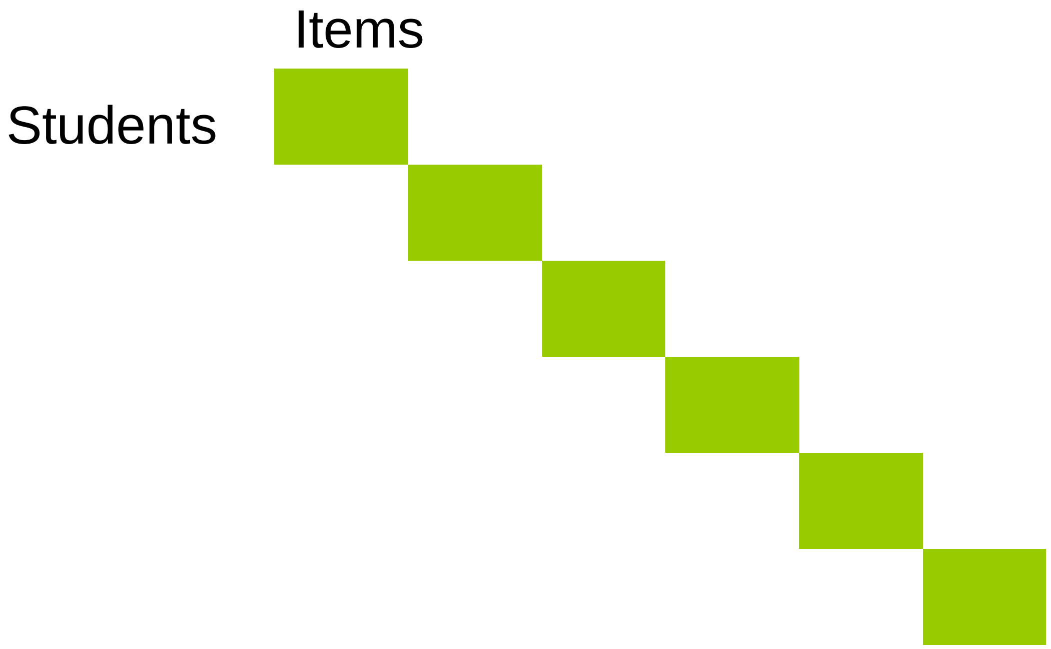
- CAT pretesting or linear pretesting?

- More items need to be analyzed than test booklet length:
- Incomplete design: not every student in sample will take all items

- **How to divide the items over the test booklets?**

Some Common Designs (1)

- Unlinked:



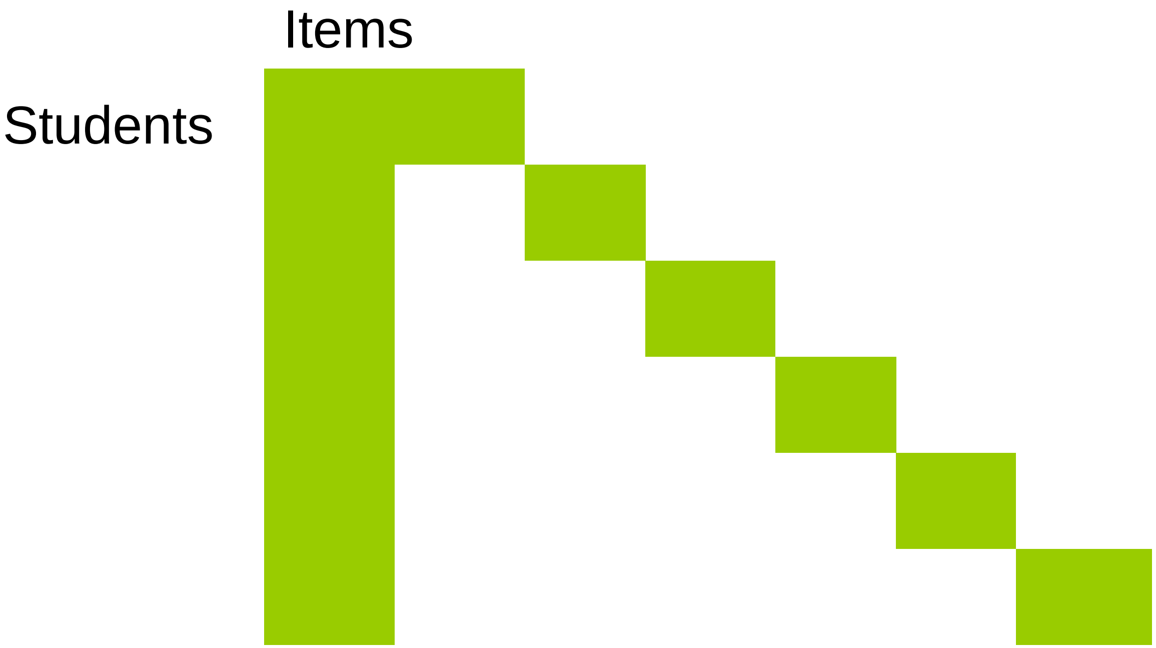
Items

Students

- How can we compare item difficulties over different test booklets?
- Item difficulties – Student abilities

Some Common Designs (2)

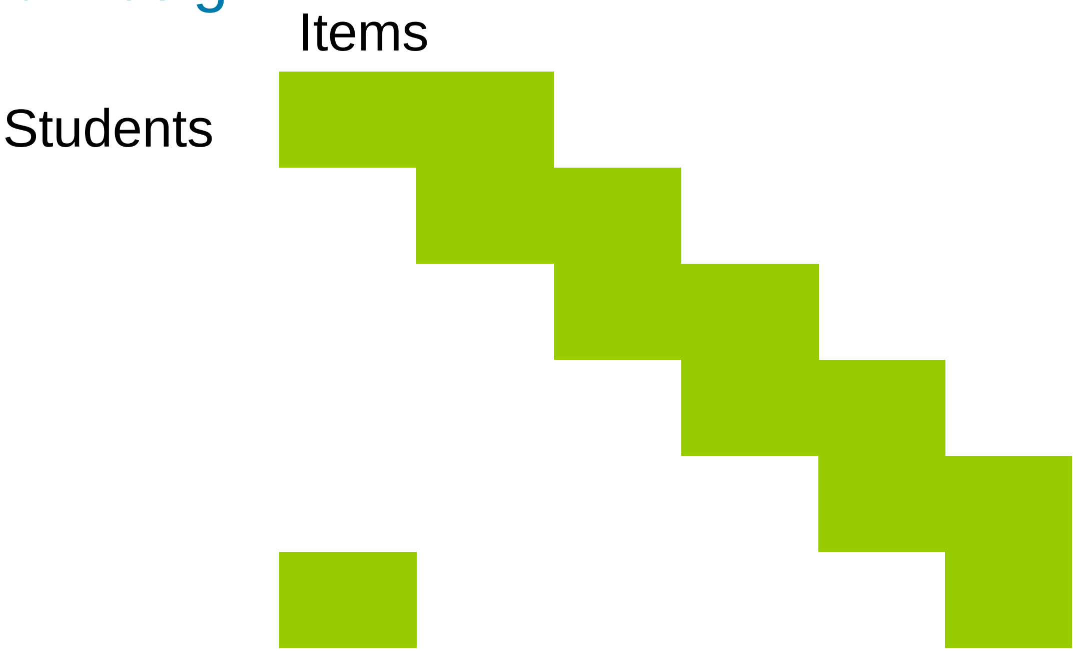
- Central Anchor:



- Some item parameters are estimated more precisely than others.
- Is this efficient?

Some Common Designs (3)

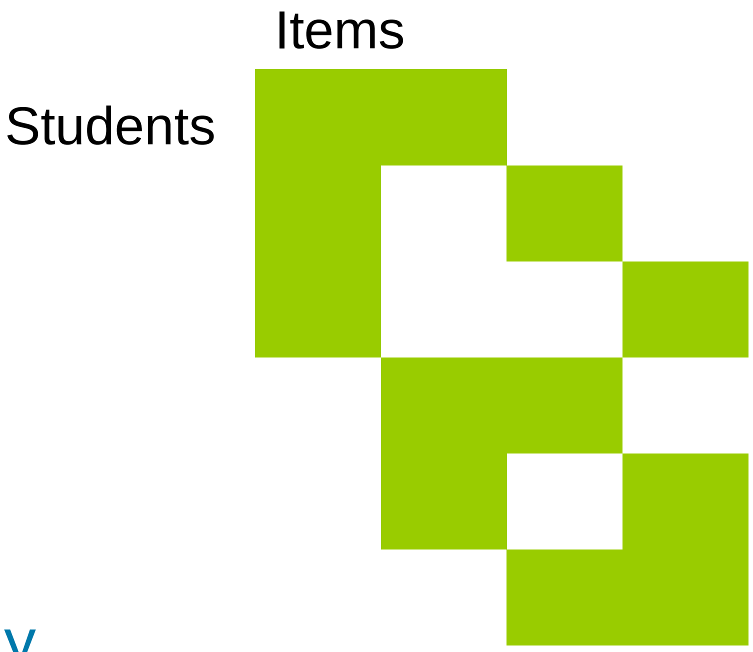
- Block interlaced Design:



- All items are observed (about) equally

Some Common Designs (4)

- Balanced Block Design:



- All items observed equally
- All item pairs observed:
  detection of misfit (dependency!)

Optimization

- **Can we exploit the advantages of BB while keeping the logistics manageable?**

- Maximize number of item pairs
- Subject to maximum number of test booklets
- Subject to other constraints

- **Homogeneous Designs:**
- Overlap between test booklets as regular as possible

Experiments

- Simulations

- Rasch model
- Items: b ~ N(0,1)
- Population: theta ~N(0.2,1)

- a constant number of observations per booklet, and per item

Simulations (1)

- 3 item pools, 3 designs for each pool:
- 150 items, 30 items per booklet, 10 booklets
- 180 items, 30 items per booklet, 12 booklets
- 160 items, 20 items per booklet, 16 booklets

- Homogeneous, BI, BB
- 45, 66, 120 booklets (BB)
- 2250, 2640, 3600 students
- 450, 440, 450 observations per item
- Overlap (Hom.) 4-5, 2-3, 1-2

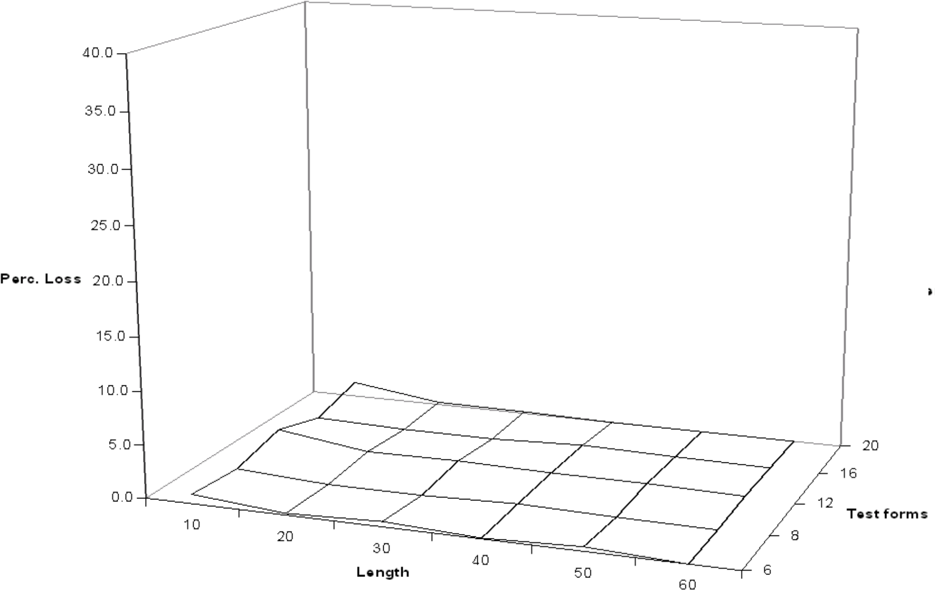
Simulations (2)

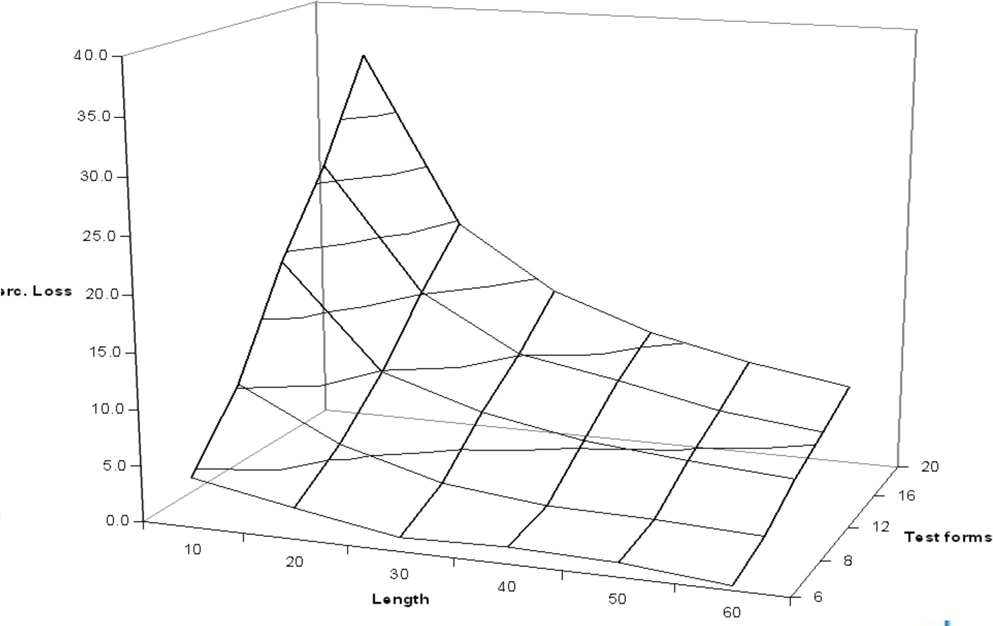
- Average Standard Error of b:

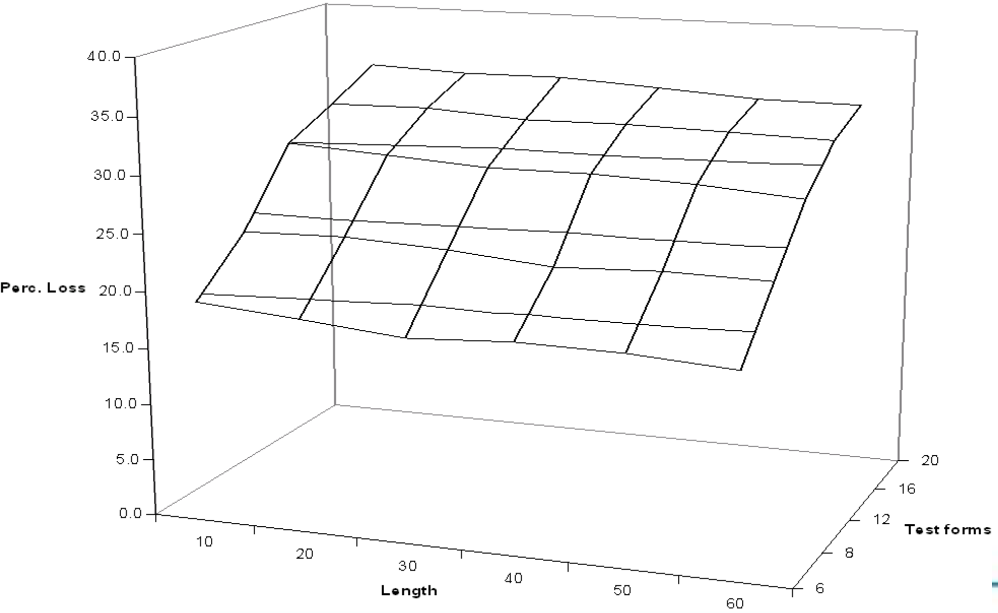
	Hom	BI	BB
150	0.114	0.121	0.114
180	0.114	0.122	0.114
160	0.117	0.134	0.117

- Reduction of 6 – 12%
- Reduction of 12 – 24% of sample size

Simulations (3)







Simulations (4)-Misfit

- Multidimensionality:
- Pool 150 items, booklet length = 30
- 10 items 2nd trait, uncorrelated

Item Fit Test: (p-value)

Item	Hom.	BI	BB
141	0.000	**0.306**	**0.106**
142	0.000	0.003	0.028
143	0.015	**0.485**	0.024
144	0.000	0.003	0.000
145	0.000	**0.601**	**0.979**
146	0.000	0.000	0.001
147	0.000	0.046	**0.069**
148	0.000	**0.077**	0.035
149	0.000	**0.097**	0.049
150	0.000	0.015	0.007

Simulations (5)

- describe a perfect world
- Can we find similar advantages in the real world?
- Entrance test (11 yr. olds):

  approx. 130000 students per year

  120 items Arithmetic, 2 PL

Arithmetic

Length 20, 3168 students sampled – 528 per item

100 repl.	Hom	BI	BB
Booklets	12	12	66
se(b)	0.116	0.127	0.116
sd(b)	0.115	0.130	0.117

Length 30, 2240 students sampled – 560 per item

100 repl.	Hom	BI	BB
Booklets	8	8	28
se(b)	0.110	0.113	0.111
sd(b)	0.109	0.115	0.110

Conclusions

- Establish overlaps as regular as possible between **all** test booklets
- Or, at least as many test booklets as possible

Thank you

?

angela.verschoor@cito.nl

now you know