The Danish National testing system - a tool for improving the culture of evaluation

– Computer Adaptive Testing in the Danish school system as a pedagogical tool

Jakob Wandall,
Chief Adviser,
Skolestyrelsen (Danish National School Agency)

The Danish National testing system

- Background for the initiative
- The Danish testing system
  - Design of the testing system
  - Status for the project
- Adaptive vs. linear testing
- Challenges

Background for the Danish initiatives to improve the culture of evaluation

Facts of the public primary and lower secondary education in Denmark, ”Folkeskolen”:

- 600.000 pupils, 60.000 pr. form
- 1.300 ordinary folkeskoler, 600 other institutions (e.g. for special education)
- School governed by decentralized local governments → significant differences from municipality to municipality
- 4-5 pupils pr. new computer (<3 year)
- highspeed internet-connection almost everywhere

→ **Infrastructure suited for It-based assessment**

Background for the Danish initiatives to improve the culture of evaluation

Teachers in Denmark

- Until 2004: No common objectives between 1. and 9. form in “Folkeskolen” → significant differences from school to school
- Poor tradition for standardized assessment of pupils – e.g. weak tradition for testing and use test results
- Strong culture for independent, self-governing teachers with focus on “soft evaluation methods”
- Weak tradition for leadership in school
- Fear of the control aspect in the test system.

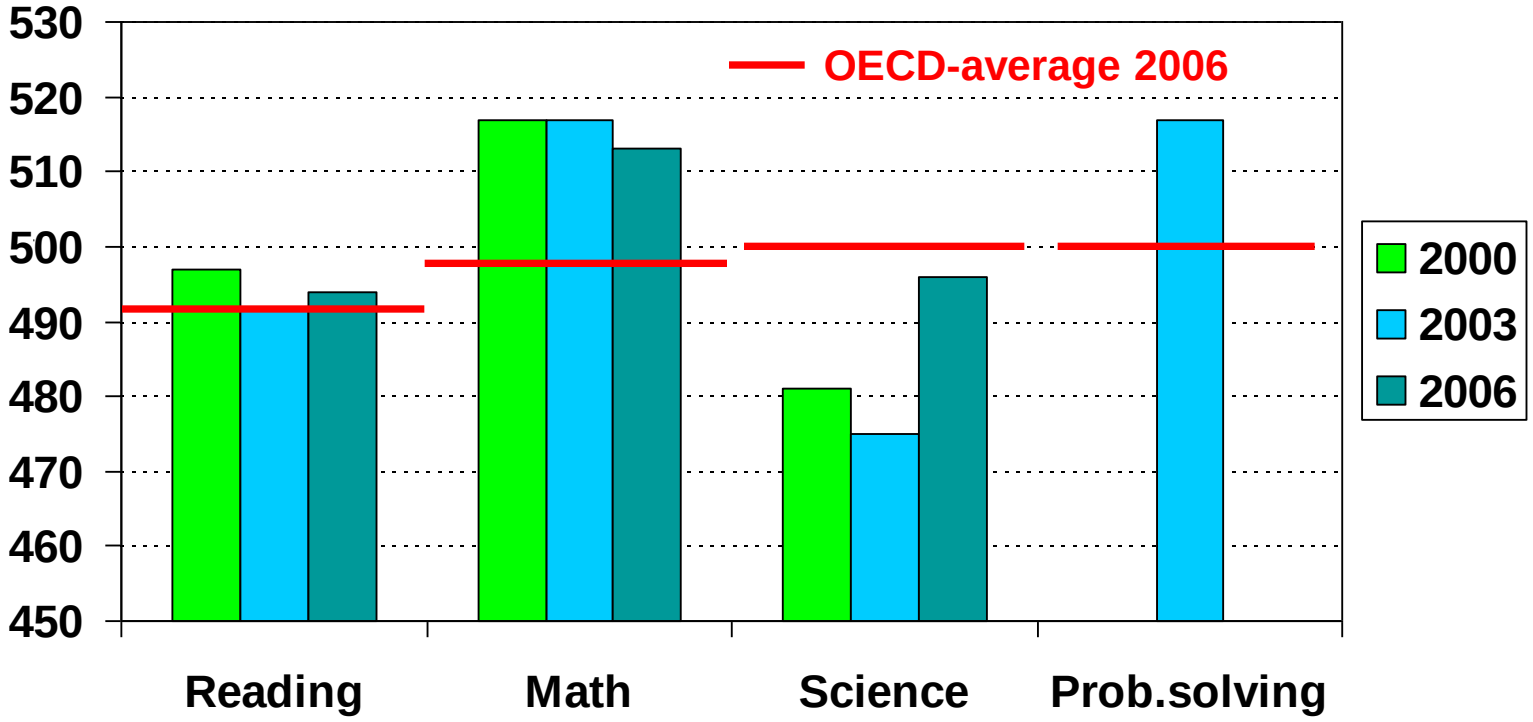
Background for the Danish initiatives to improve the culture of evaluation

1. PISA-surveys (2000-2006)

2. OECD-review on Denmark (2004) and National reports on the ongoing evaluation/assessment in school and differentiation in teaching (2004)

3. Governments initiatives on school-improvement, several changes of the act of the ”Folkeskole” (Primary and lower secondary public school), according to the OECD-Recommendations

The Danish PISA results, 2000 - 2006



Review of National Policies for Education, Denmark, OECD 2004

”Denmark has one of the most expensive education systems in the world, and for years perceived it to be one of the best in the world.

However the disappointing results of recent international tests to measure schooling outcomes confirmed earlier evidence, that the system actually is underperforming.”

OECD-review 2004

Strengths and weaknesses $\Rightarrow$ 35 recommendations

Some weaknesses:

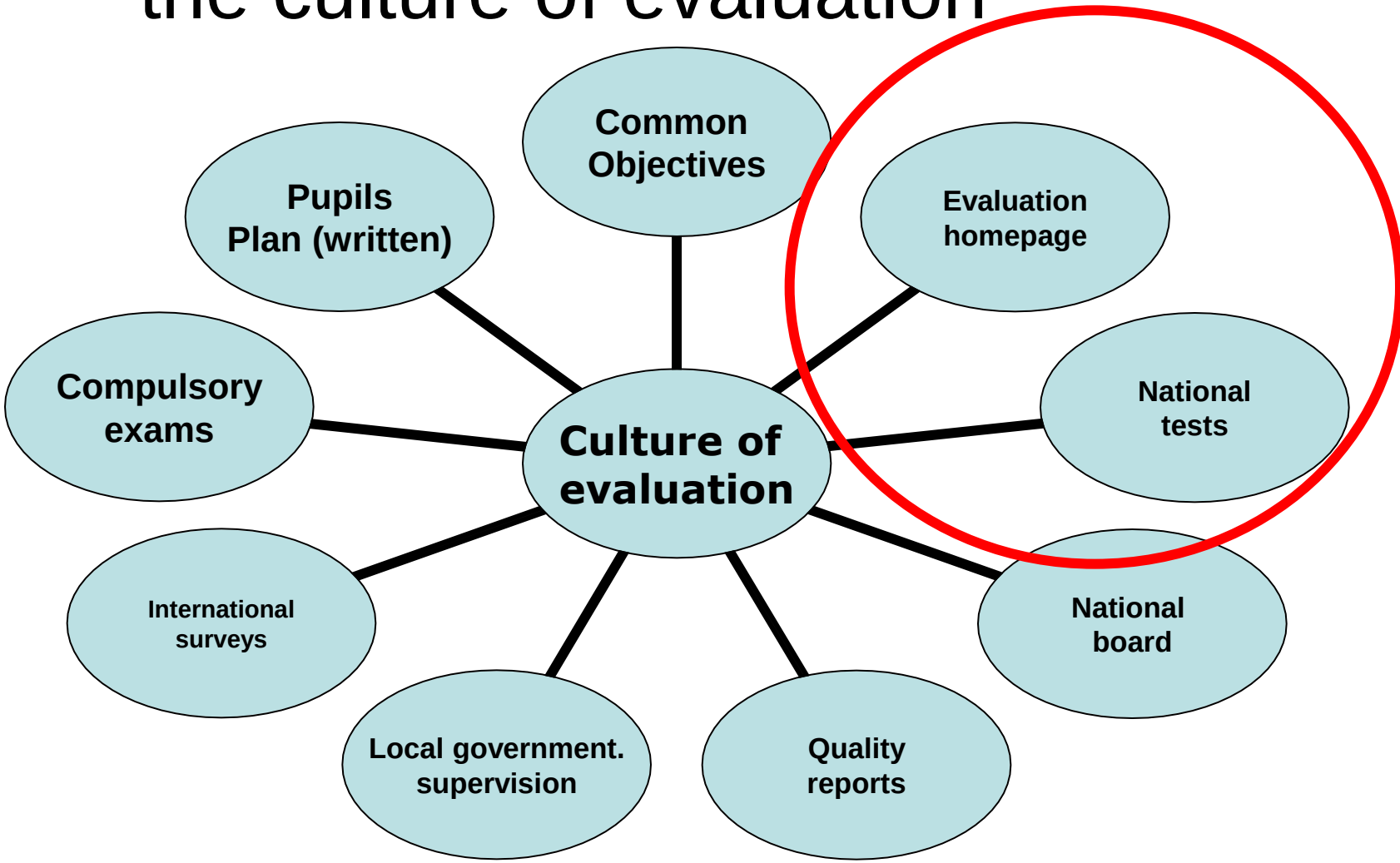
- Poor tradition of pupils assessment
- Lack of written feedback to pupils and parents
- Insufficient teacher qualifications in assessment techniques
- Insufficient mutual experience exchange (“best practice”)

OECD-review 2004

Some recommendations for the Minister:

- Establish an agency to monitor the local governments administration of local schools.
- Establish a national performance profile to monitor the development in the national level in subjects.
- **Initiate development of criteria-based tests.**
- **Initiate evaluation/development of different assessment methods and -materials.**
- **Carry out a policy based on the principle, that test results don't get published for ranking purposes.**
- Consider if it is necessary to change the act of "folkeskolen".

Government initiatives to improve the culture of evaluation



Implementation of OECD's recommendations

- The Government initiatives followed the OECD recommendations with some modifications – e.g. in the testing system
- The OECD-team recommended criteria-based test - recommendation based on a different tradition of testing/evaluation
- Background: The OECD-team came from England, Canada and Finland

Two different traditions for assessment/use of test results

1. The Nordic / continental

Terminology, Nordic/German origin (Danish: Prøve, vurdering, bedømmelse, opgave, karakter)

2. The Anglo-American

Terminology, English origin (Danish: test, score, item)

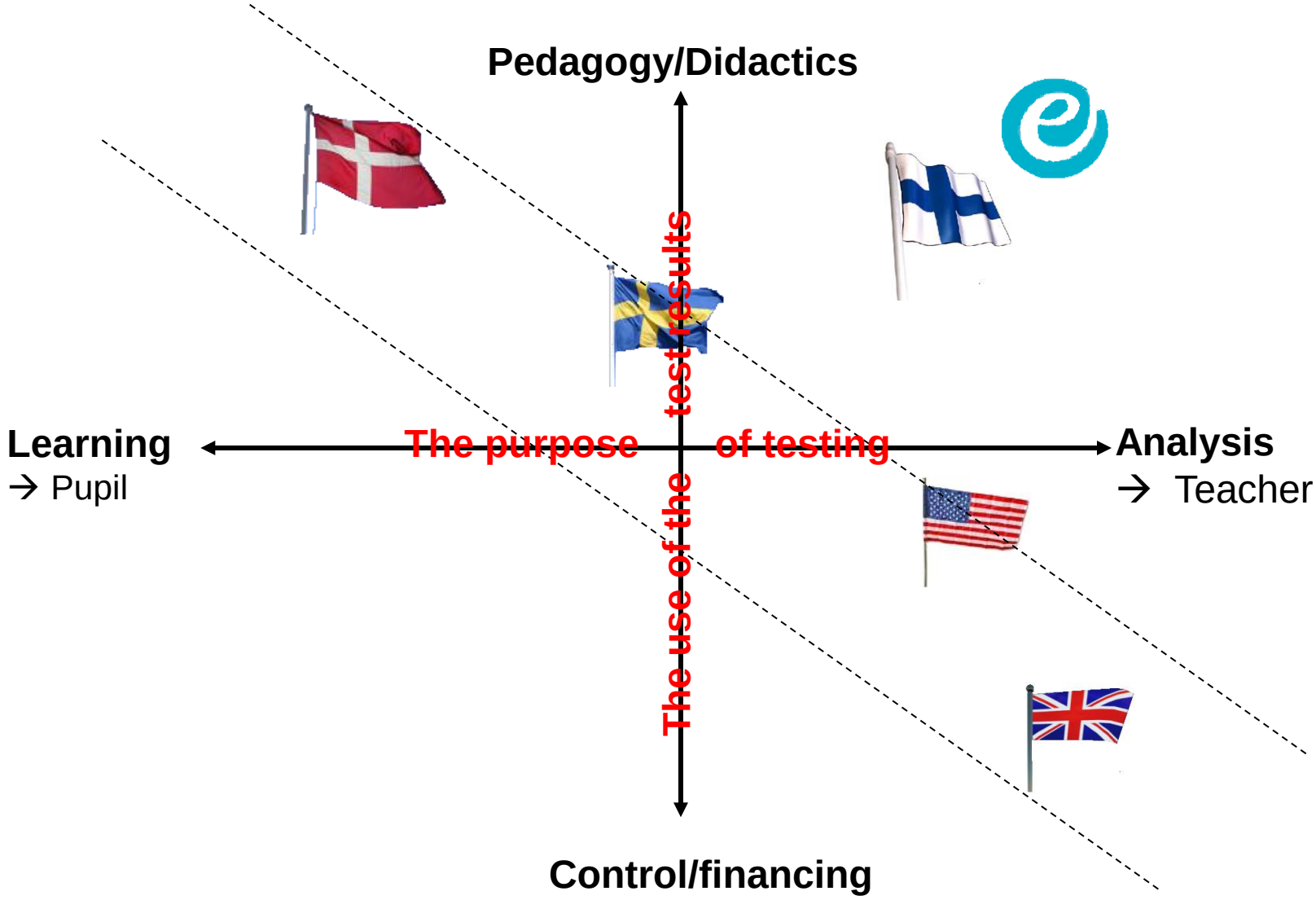
Traditions for assessment/testing,

Why testing? What are the results used for?

Traditions:	Nordic/continental	English/American
Use of results	Pedagogy	Control of outcome
	Didactics / Teaching	Financing / grants
	Formative / Low-stake	Summative / High-stake
Purpose	Learning → Pupil	Analysis → The "system"/Professionals
	Focus on equality and solidarity	Focus on ambitions and elite
	Solicitude	Fairness

Traditions for testing,

Why test pupils – What are the results used for



Criteria for choice of strategy

- **Test in combination with other assessment tools**
  - Easy to use for the teacher
  - Flexible systems from the schools point of view
  - Low/no cost for the school
- **Priority to pedagogic purposes – Formative – Low stake**
  - For teachers assessment of the pupils
  - For the headmasters monitoring the pupils results
  - For the local government monitoring the schools results in tests
  - For monitoring progress in the school system - only on national level
- **Effective Self-correcting tests**
  - valid,
  - reliable
  - detailed results
  - **Max. 45 minutes** (1 lesson)

→Central administered internet-based Computer adaptive testing - **with focus on pedagogy**

Result of a EU-tender:
A private contractor delivers the tests and other assessment tools

COWI A/S

-with subcontractors:

- **@ventures (Public it-enterprise, e-learning)**
- **CSC Danmark A/S**
- **JCVU**
- **CVU-Storkøbenhavn**
- **University of Bergen**
- **Public Communication A/S (information bureau)**
- **DPU (Danish Pedagogic University)**

Homepage containing:

- Guidance of assessment - methods (<50 articles)
- Examples of best practice (<200 examples)
- Tools of assessment (27 methods, e.g. tests)
- Access to the test system
    - 12 test – (< 500 items each)

The Danish testing system

The national tests

Form / Subject	1	2	3	4	5	6	7	8	9
Danish / reading		X		X		X		X	
Math			X			X			
English							X		
Geography								X	
Biology								X	
Physics/chemistry								X	
Danish as 2'nd language									

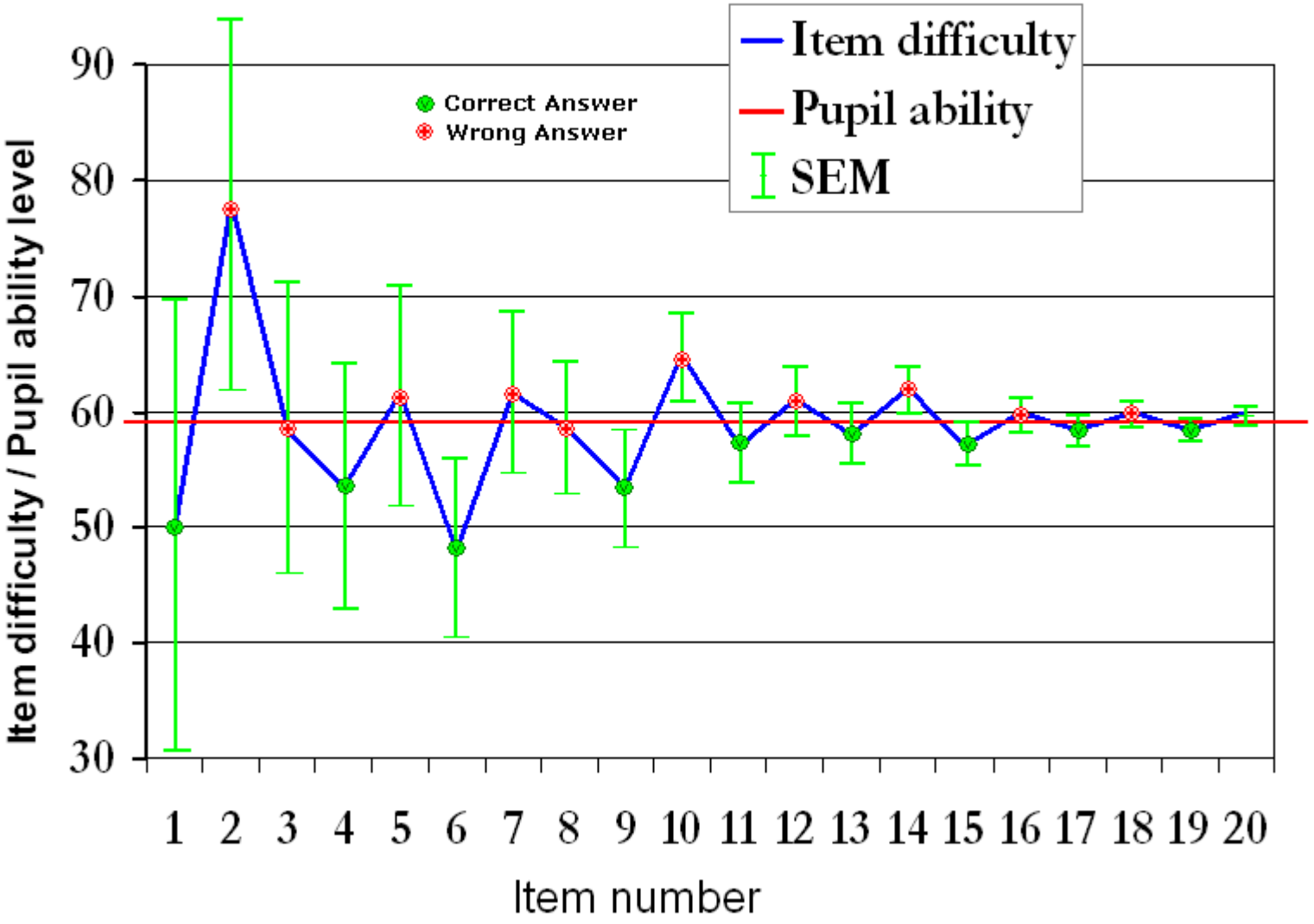
Forms where tests can be used

X 10 compulsory tests

Main features of the Danish national tests

- IT-based
- Provided freely by The Ministry of Education (The National School Agency)
- Automatically scored – the teachers do not have to correct and analyze the test
- Large database with quality controlled and standardized items
- Feedback on various levels (e.g. to the teacher)
- The tests are based on an Adaptive principle

Illustration of an adaptive test



CAT (Computer Adaptive Testing)

- Adaptive = Adapts to the individual pupils ability

  The principle is simple:

  Right answer ➔ More difficult questions
  Wrong answer ➔ Easier questions

  The test is most efficient when item difficulty = pupil ability

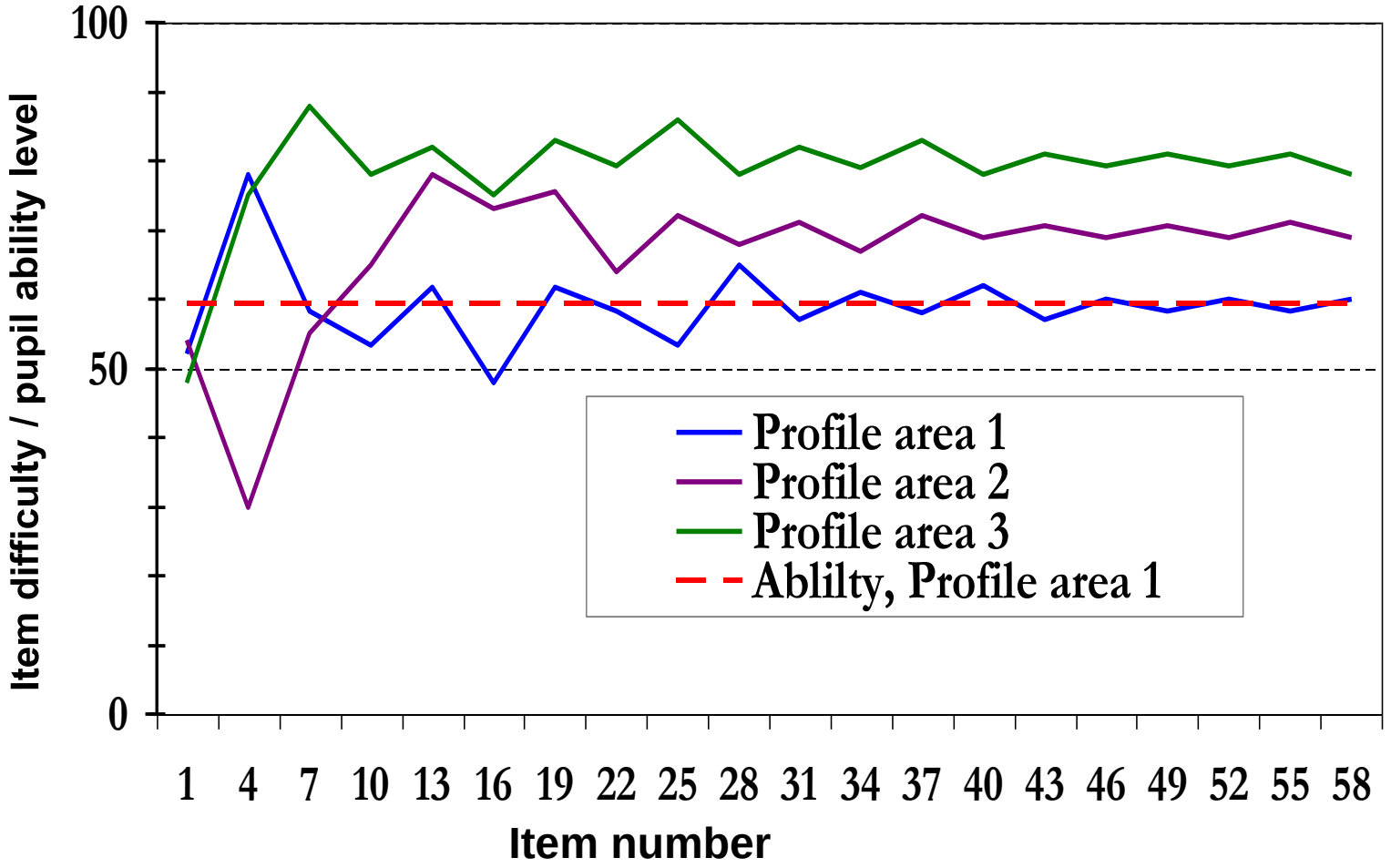
- Effective testing ➔ more detailed results:

  Adaptive within 3 “profile areas” ➔ 3 tests in 1

  (E.g. Math: Numbers & Algebra, Geometry and Math in context)

- Simple principles, but a few tricky conditions:
  - Extensive demands to the technology – both capacity and stability.
  - Very large Item-banks with the exact right mix of high quality items

Adaptive test – 3 profile areas



Different types of questions/items

- Three main types (with a lot of varieties)
  - Multiple choice
  - Insert text/number
  - Drag and drop
- Both dichotomous and polytomous items are used
- Assessment of item response based on a partial credit model
- Polytomous items/partial credit provides new opportunities within the Rasch model. E.g. we are working on inclusion of the time dimension in the estimation of the pupils reading ability.

Controlling the items

Condition for adaptive testing is:

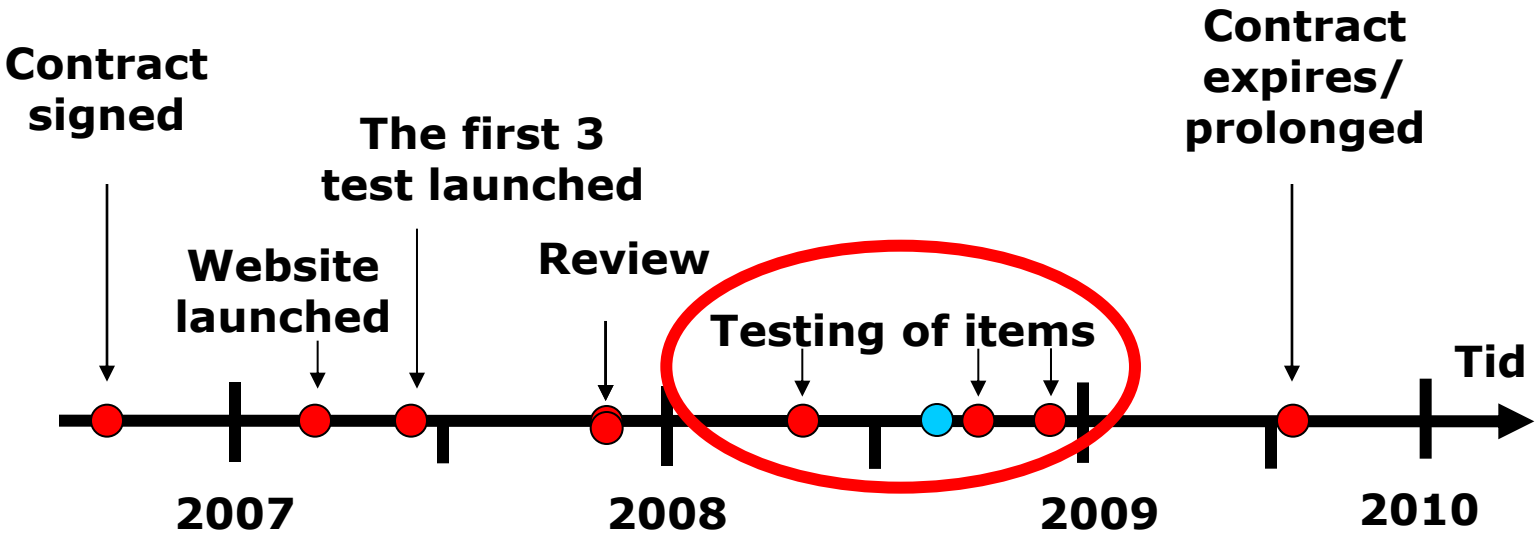
- That the difficulty of the items are well defined and stabile over time (homogeneity).
- No differential item function.
- Sufficient number of items, so that even the fastest pupils don’t run out of items
- That the items are evenly distributed on item-difficulty → challenges for all pupils.

Therefore following requirements has to be met:

- Minimum 540 items pr test (180 pr. Profile area) – evenly distributed on difficulty levels
- All items are tested on a large number of pupils (600-700)
- The items are required to fit a Rasch model
- Not more than 3 runs pr pupil/test (including compulsory test)

Status on the implementation

- The first 3 test was launched in may/June 2007, with reduced item banks
- There has been conducted an expert review in oct. 2007 that showed, that
  - the psychometric standard in the National Danish test is at a very high level, but
  - the quality of the items and size of the item banks were not sufficient
- Development and testing of items in 2008 – includes 900 schools

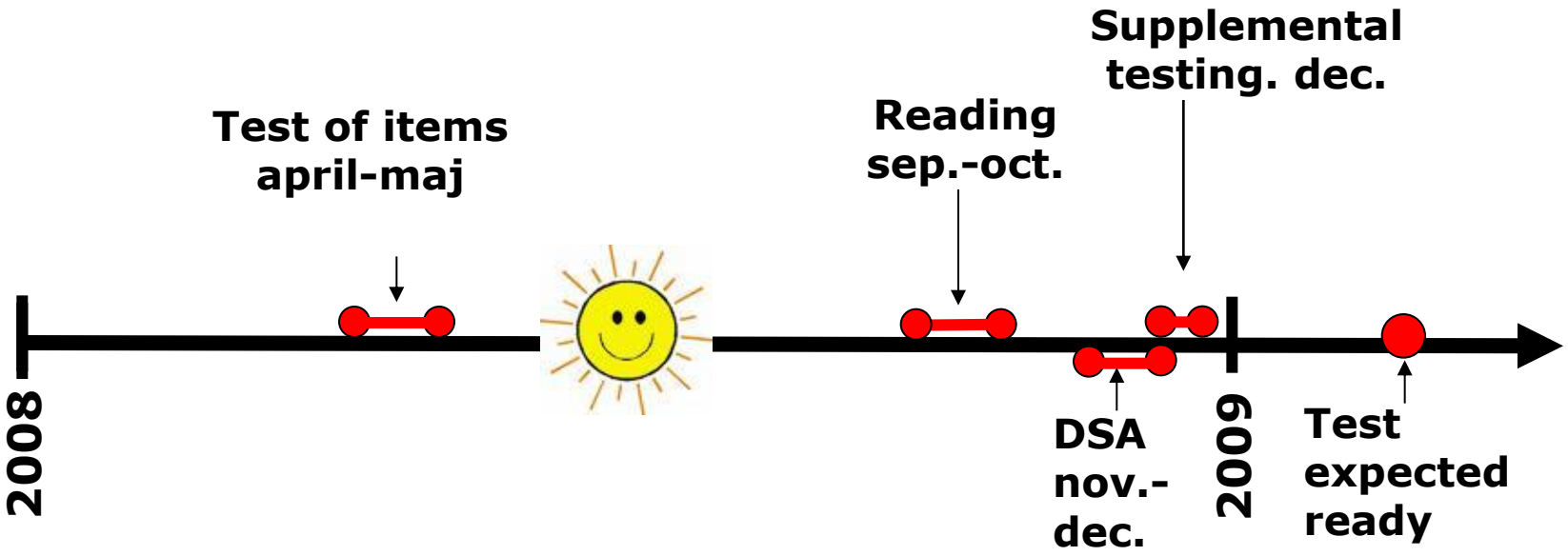


Plan for test of items

- **7. april-2. maj 2008: Test of items for Matematik3 & Matematik6, English7, Geography8, Biology8 and Physics /Chemistry8. (450 schools).**

- **29. sep.- 24. oct. 2008: Test of items for Danish/Reading (2, 4, 6 and 8. form), (320 schools)**
- **17. nov.- 9. dec. 2008: Danish as 2'nd language (DSA, 5 og 7 form.)- (130 schools)**
- **8. dec. – 19.dec. 2008: Supplemental testing of items (if necessary).**

- **The test expected ready in the spring 2009**



Status of the national tests in Denmark

- The second generation of the items are being designed and tested
- More items at a high standard – and better distribution of item difficulty – are coming up
- Preliminary results from Rasch-analysis looks promising

Linear vs. adaptive setup

How the testing of items are organized

- Schools participate with all pupils in one or two forms (not pupils with special need).
- All pupils get the same set of items, but not in the same order.
- Every pupil answers 5 set of items. Each set contains 20-50 items (max. 30 minutes for the average student).
- Results are calculated and given to the schools (no stake → low stake)
- The items are tested on several forms, so that item difficulty should match pupils ability as precisely as possible.
- The testing of items is administered in a linear (non-adaptive) setup.

How the testing of items are organized
Example: Danish/Reading – 6. form

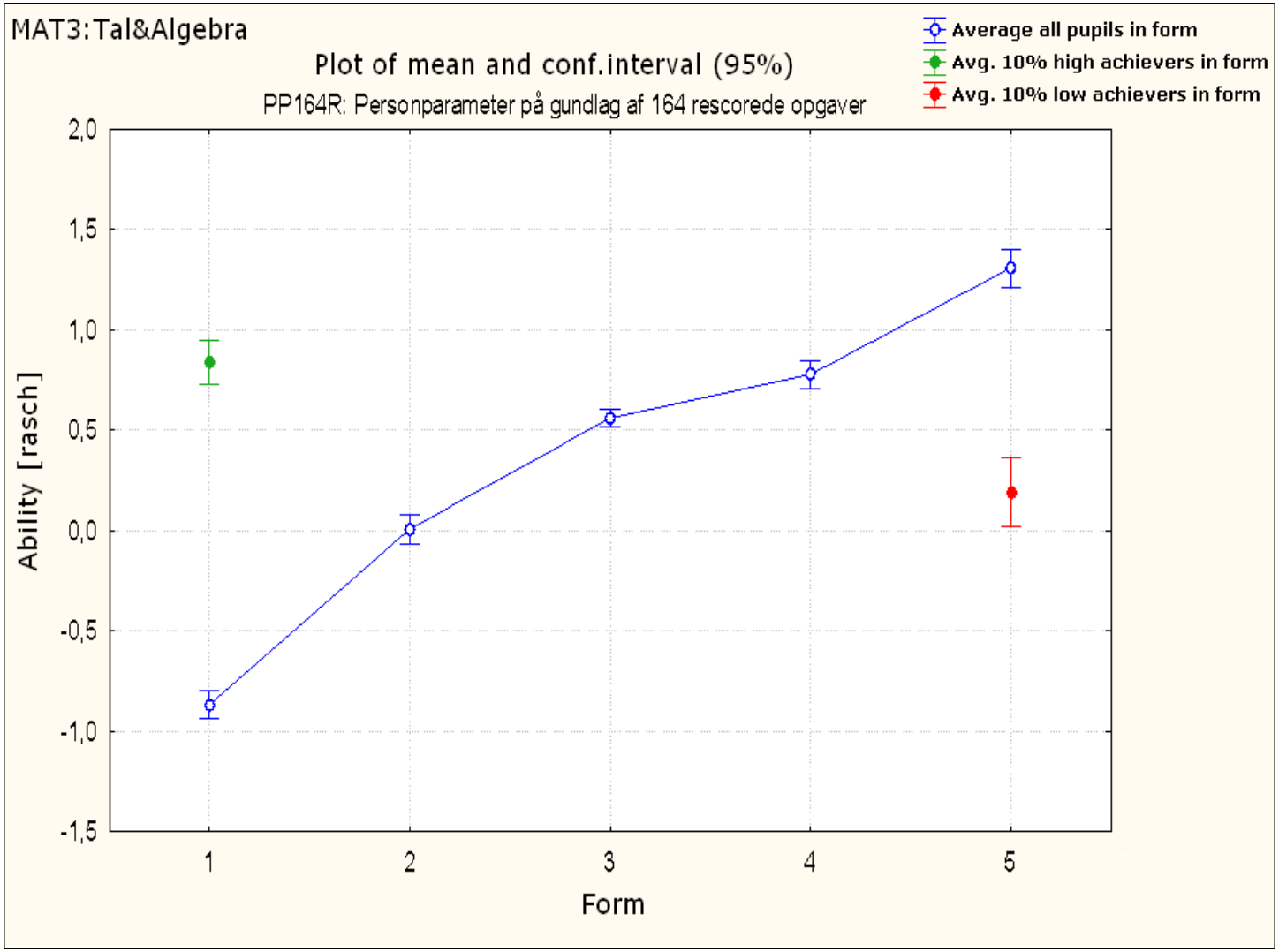
	5. (the form below)	6.	7. (the form above)
	Number of pupils testing each item		
The 20% least difficult items	**350**	**350**	**0**
Medium difficulty items	**150**	**400**	**150**
The 20% most difficult items	**0**	**350**	**350**

Danish experiences: linear vs adaptive

- The adaptive model performs relatively best in the extremes (measuring proficiency of high and low achievers).
- The adaptive setup is pretty efficient: Reduction of SEM to 0,3 with 6-12 items (linear setup app. 20 items).
- The adaptive test is generally a good experience for the low achievers, but can be a frustrating experience for high achievers (no free rides).
- The greater the differences in student ability within the tested group of pupils, the greater the advantage of adaptive testing.

Danish experiences: linear vs adaptive

Progression and differencies between pupils in the same form



Challenges for Denmark

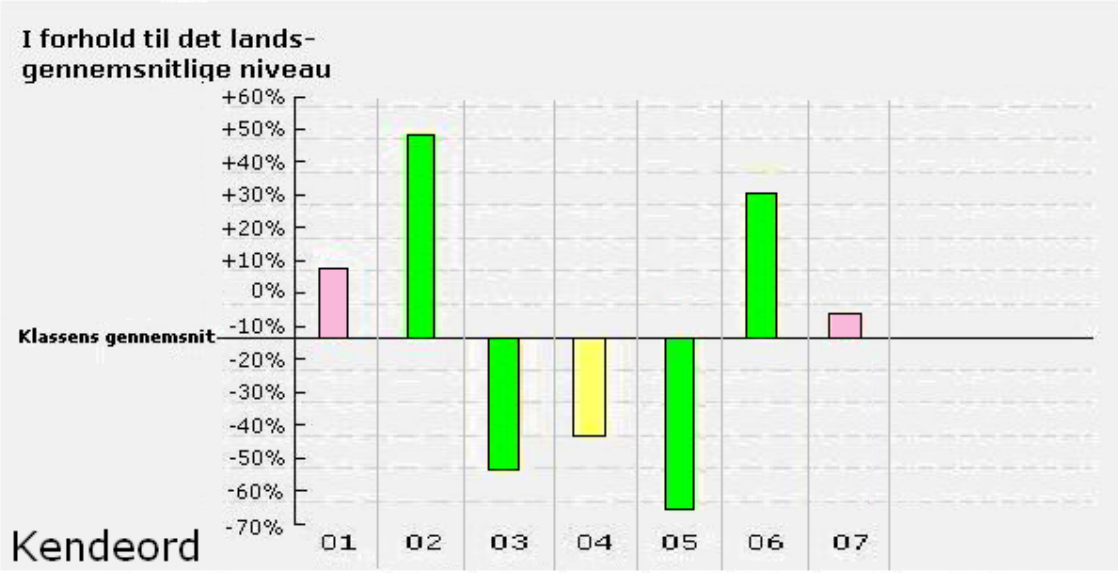
- To make the it-systems work (booking, test and results) flawlessly
- To create and maintain high quality item banks
- To develop user friendly ways to present the results
- To convince the teachers/Schools to use the information – and describe how to do it (guidance in good practice) !!

Oplysninger om testopgaver og testresultater er fortrolige, jf. folkeskolelovens § 55 b. Uberettiget videregivelse eller udnyttelse af testresultater og testopgaver vil derfor kunne straffes i medfør af straffelovens regler om tavshedspligt.

Denne rapport er dannet af: nina3116. Skole: Evalueringsskolen | Test: **Matematik 6. kl** | Testtype: Obligatorisk | Testperiode: 2006-2007 | Rapportgruppe: **6 x** | Lærer: Henrik

01: Tal og algebra Vis | 02: Geometri Vis | 03: Matematik i anvendelse Vis

Profilområde: 01: Tal og algebra



01: Addition
02: Subtraktion
03: Mulitplikation
04: Division
05: Brøk- og decimalregning
06: Ligninger og uligheder
07: Andre kendeord

Det faglige niveau er bestemt på et:
- statistisk sikkert grundlag
- statistisk mindre sikkert grundlag
- statistisk usikkert grundlag

Brief explanation of the statistical foundation

Forklaring til rapporten
Til hver opgave er tilknyttet oplysninger om, hvor i undervisningen (belyst ved trinmål) der kan undervises i emner, der giver kundskaber og færdigheder, der skal til for at løse opgaven. Klassens testresultater i opgaver med forskellige trinmål vises i forhold til henholdsvis klassens gennemsnit og landsgennemsnittet på tværs af opgaver med trinmålet tilknyttet.

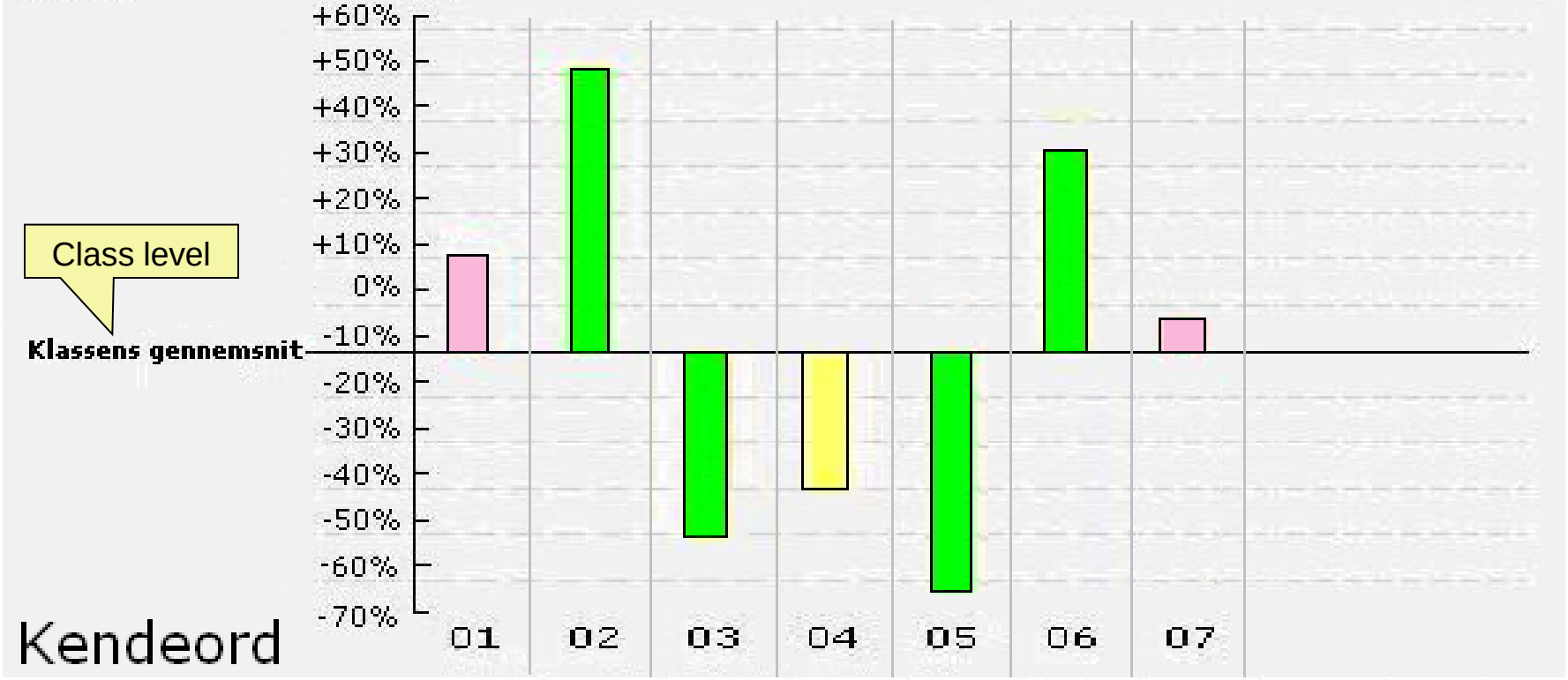
Desuden vises ved farverne på søjlerne hvor sikker målingen er. Resultatet bliver mere sikkert jo flere opgaver der er besvaret, jo flere elever der har besvaret opgaverne, jo mere homogen besvarelsen er.

01: Tal og algebra Vis | 02: Geometri Vis | 03: Matematik i anvend

Profilområde: 01: Tal og algebra

Compared to national level

I forhold til det lands-gennemsnitlige niveau



01: Addition
02: Subtraktion
03: Mulitplikation
04: Division
05: Brøk- og decimalregning
06: Ligninger og uligheder
07: Andre kendeord

Det faglige niveau er bestemt på et:

- statistisk sikkert grundlag
- statistisk mindre sikkert grundlag
- statistisk usikkert grundlag