Hvordan er det vi tester?

Inge Henningsen

Dansma 5. oktober 2018

Når man kvantificerer mennesker……

Internationale og nationale læse/regneundersøgelser spiller en væsentlig rolle i den offentlige debat om skolen og har dybtgående indflydelse på indretningen af de nationale skolesystemer

Vigtigt

- at data fra sådanne undersøgelser bliver analyseret professionelt og retvisende.
- at elever, lærere og forældre kan overskue testprocessen.

Nu følger en tur i maskinrummet

Algoritmerne kommer ikke – de er her.

”Med automatiseret beslutningstagning og fortolkning af enorme datamængder kan algoritmerne styre menneskers liv og bestemme hvilke informationer og muligheder, der bliver tilgængelige for borgerne”

- *De skjulte algoritmer 2018*

I dag er algoritmer ikke bare en regneforskrift.

De er systemer til automatisk behandling af data, der er uigennemskuelige og mere eller mindre styret af skjulte antagelser, som konstruktørerne har lagt ind i programmerne

To slags matematiktest

- De ”gammeldags” (MAT og MG og MF) hvor opgaverne er kendte og der tilstræbes en bred faglig dækning. Bedømmelse er (nogenlunde) transparent. Eleven (klassen) i fokus.

- De moderne (PISA og de nationale test). IRM-baserede. Opgaverne er hemmelige og er af beregningsmæssige grunde fagligt indsnævrede. Eleverne får randomiserede opgavesæt (fx adaptive test). Analyse, scoring og vurdering foregår automatisk og er for alle praktiske formål uigennemskuelig for de involverede parter. Evaluering og sammenligning af lærere, skoler og skolesystemer er i centrum.

IRM – et greb i værktøjskassen

Alle de sidste års store internationale studier (PISA, IALS, TIMSS, ALL, PIAAC) af læse- og regnefærdighed er baseret på Item Response Modeller (IRM). Det gælder også de nationale test.

Dette har nogle fordele, men det medfører voldsomme begrænsninger på testbatterierne, fordi det kræver at alle opgaverne skal ”måle det samme”:

“The fundamental assumption built into the IALS design is that proficiency is related in a regular way to item difficulty, a way that **is invariant across language, culture and subculture**.”

(Murray, 1995:11-12)

Raschmodel

Den en-dimensionale Item Response Model (Raschmodellen) kræver at opgaverne (items) i et prøvesæt i matematik alle skal måle ”det samme”. Det betyder at

- Matematikfærdighederne skal kunne repræsenteres ved en enkelt latent skala på intervalskalaniveau.

- Ingen differentiel item funktion (DIF). Items skal fx fungere på samme måde for drenge og piger og for danske og brasilianske børn (bortset fra overordnede niveauforskelle).

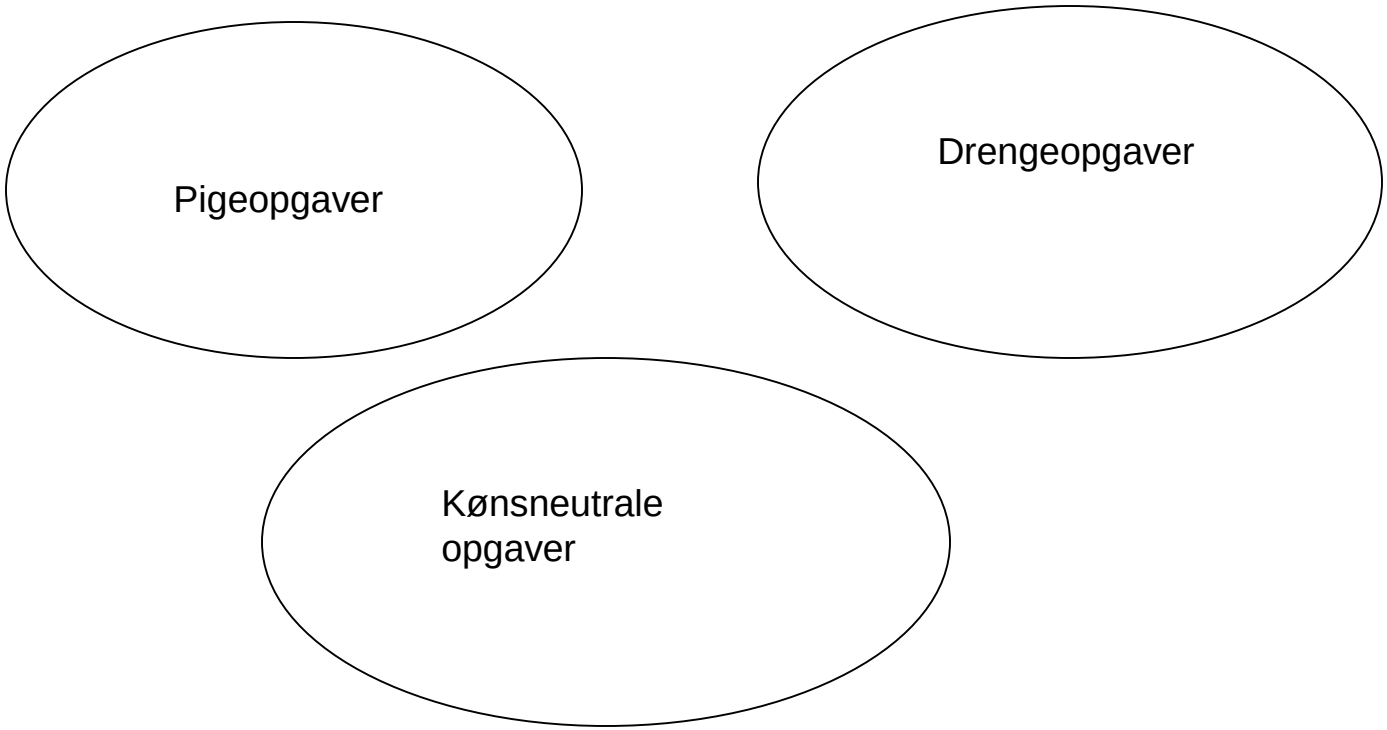
- Sammenligninger mellem elever skal være principielt uafhængige af hvilke delsæt af opgaverne, der inkluderes i prøven.

- Sådan er matematik generelt ikke – men sådan skal prøvesættet fungere, hvis man skal bruge Raschmodellen.

.

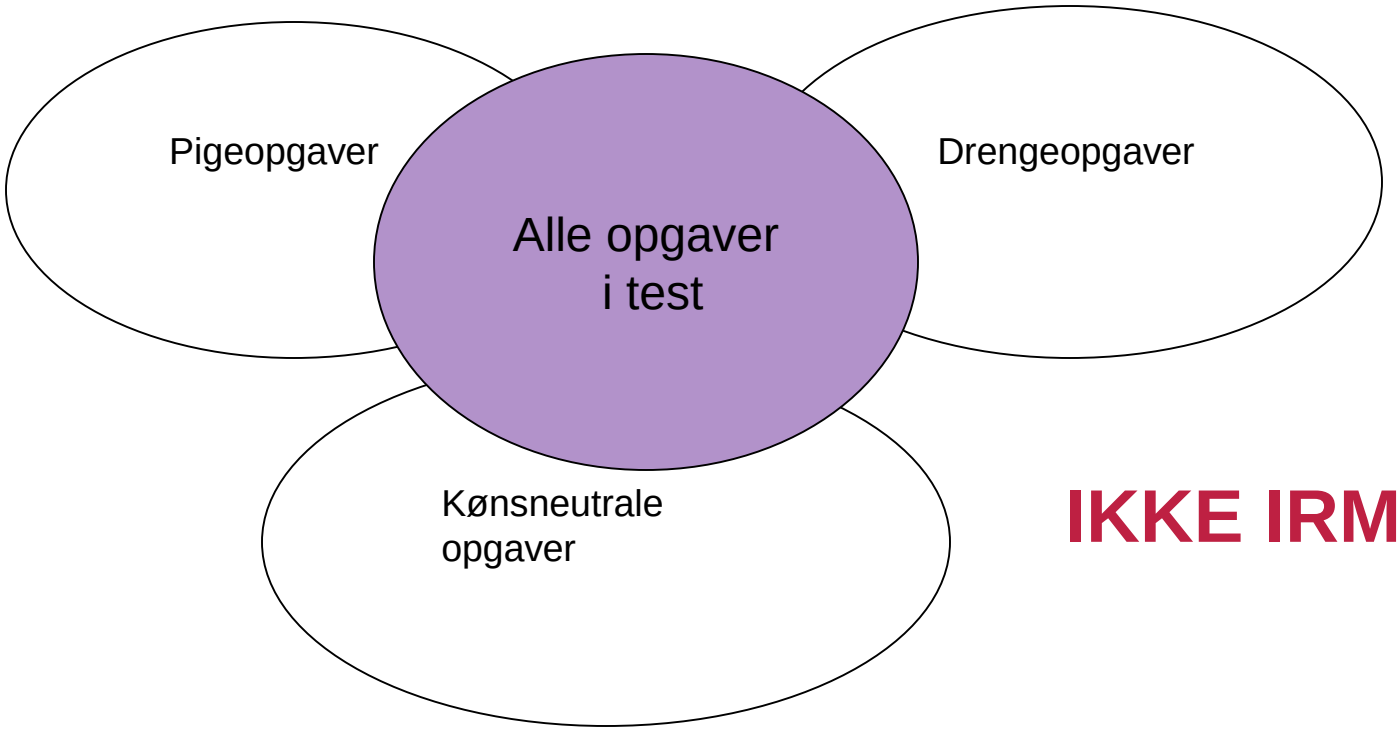
Item Response Model

- Konstruktion af IRM



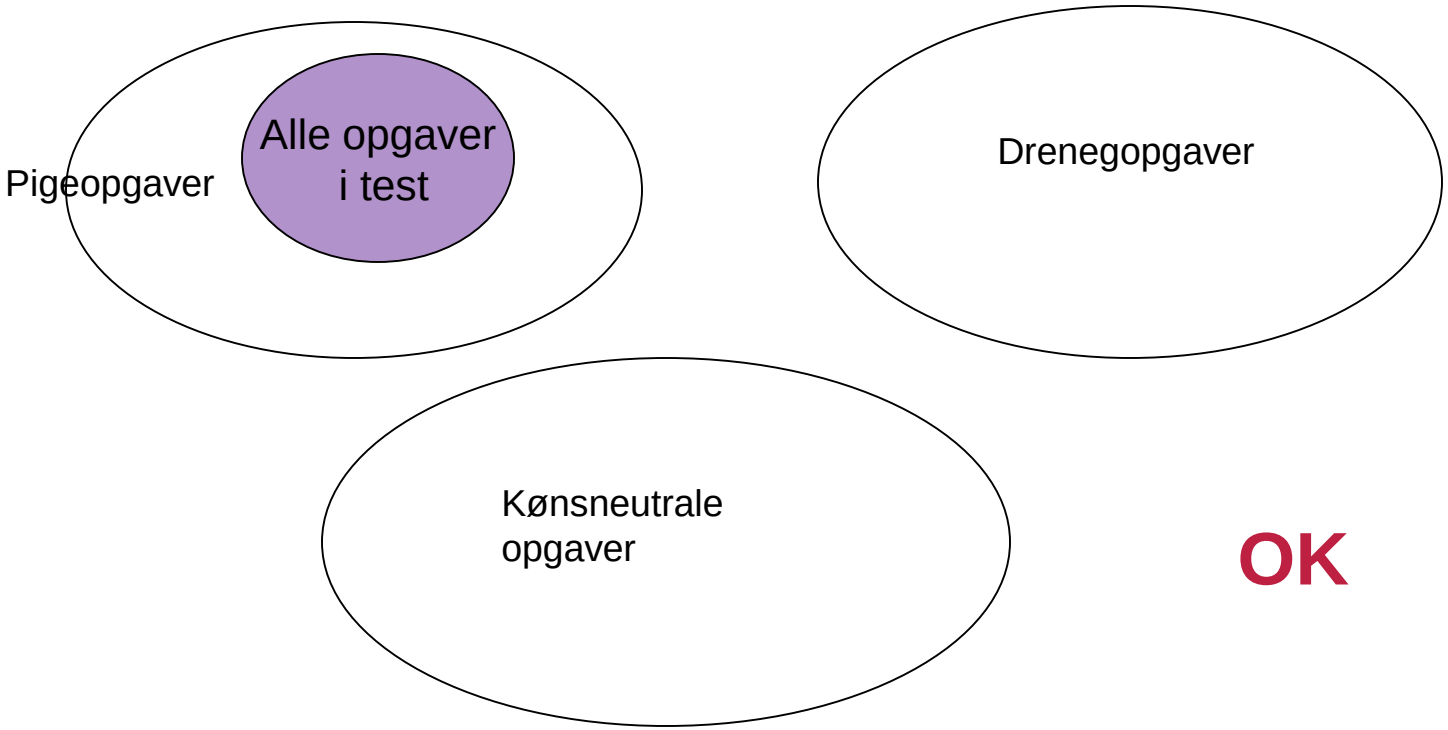
Item Response Model

- Konstruktion af IRM



Item Response Model

- Konstruktion af IRM



Hvordan laves et opgavesæt?

- Et antal elever regner opgaverne.
- Det undersøges statistisk om opgaverne har de krævede homogenitetsegenskaber
- Opgaver (også gode opgaver) skubbes ud, hvis de ikke passer i modellen
- Tilbage bliver kun opgaver, der over hele det betragtede domæne ”måler det samme.”

- Men ved vi hvad de måler?

Flerdimensionale kompetencer

- Men uanset hvad man gør i et prøvesæt, så findes der ikke entydig måde at reducere en flerdimensional virkelighed til en enkelt latent dimension. Problemerne kommer ikke bare fra testkonstruktørernes beslutning om hvilke items der skal inkluderes i hvilke test eller domæner, men også i den efterfølgende tilpasning af oversimplificerede modeller som fører til yderligere selektion og udeladelse af items for at få testet til at passe til et sæt af modelantagelser. Redskaber bør – i passende grad - reflektere den kompleksitet virkeligheden udviser. En-dimensionaliteten, der forudsættes i Rasch-modellen står i stærk kontrast til matematikkens måde at præsentere sig selv på.

- *Jævnfør*
- *Kompetencer og matematiklæring (KOM-rapport*) UVM 2002

Kompetencer og matematiklæring

	Problembehandling	Modellering	Ræsonnement og tankegang	Repræsentation og symbolbehandling	Kommunikation	Hjælpemidler
Tal og algebra						
Geometri og måling						
Statistik og sandsynlighed						

2 of 2

UVM 2002

Ikke-spørgsmål

Hvad kan man **ikke** spørge om inden for Rasch-modellen?

Eksempelvis

- Hvad karakteriserer opgaver, som er specielt vanskelige for de svage elever?
- Hvad karakteriserer items, hvor piger scorer (relativt) højt?
- Er der faglige områder, hvor elever med lavtuddannede forældre er specielt stærke/svage?
- Har minoritetselever særlige problemer/styrker

Disse spørgsmål kan principielt ikke stilles fordi alle items skal ”måle det samme” for alle.

Alt i alt

I Raschmodellen kan resultatet reduceres til antal rigtige svar.

- Hvis alle opgaver måler det samme på samme måde for alle, mister man ingen information ved alene at se på antal rigtige svar.
- Men hvordan afgør man lige hvad et opgavesæt måler, når det opfylder kravene i Raschmodellen, især hvis opgaverne er forskellige

- Hvis opgaverne måler noget forskelligt er der i almindelighed information tilbage i opgaverne ud over antal rigtige svar.
- I inhomogene opgavesæt afgør testkonstruktøren, hvem der skal klare sig godt.

PISA

- PISA-programmet (Programme for International Stu- dent Assessment) er etableret som et samarbejde mellem OECD-landene.
- Formålet er at måle, hvor ”godt forberedt de elever, der befinder sig i slutningen af deres undervisningspligtige alder, er til at møde fremtidens udfordringer.”
- Omfatter over 100 lande
- Leverer rangering af landene i matematik, læsning, naturfag
- Systemrettet, ikke elevrettet
- Baseret på IMR

"PISA According to PISA"

Hopmann, Brinek & Retzl 2007 har følgende kritikpunkter

PISA er

- Tilrettelagt kulturelt skævt
- Metodisk indsnævrende
- Ikke dækning for, at PISA-scorerne giver et validt billede af "hvad enhver elev skal kunne".
- De nationale rangordninger er baseret på usubstantierede antagelser om validitet, pålidelighed og homogenitet
- De nationale og internationale PISA- konsortier indgår ikke i en sædvanlig videnskabelig dialog om deres resultater.

- Hopmann, S.T., Brinek,G. & Retzl, M. (Eds.) (2007) *PISA zufolge PISA – PISA According to PISA,* LIT VERLAG, Berlin

Kompetenceniveauer

- Ud over testscorerne opererer PISA for hvert af domænerne med en række kompetenceniveauer. I matematik er der seks niveauer. Det følgende er beskrivelse af niveau4
-
- ”Eleverne skal kunne arbejde effektivt med eksplicitte givne modeller for sammensatte konkrete situationer, der kan være pålagt bånd eller nødvendiggør specifikke antagelser. De kan vælge og integrere repræsentationer udtrykt i symboler og forbinde dem direkte til aspekter i situationer fra “the real world”. Eleverne på dette trin kan ligeledes fleksibelt udnytte veludviklede færdigheder og ræsonnementer med nogen indsigt i situationer. De kan konstruere og kommunikere forklaringer og argumenter baseret på fortolkninger, argumenter og handlinger.” (PISA 2003:49)

”Funktionelle analfabeter”

Kompetenceniveauerne fastlægges statistisk (normbaset). Eksempelvis er niveau 1 i matematik defineret ved at omfatte de 18.8% svageste elever i de oprindelige OECD-lande og karakteriseres således

”På niveau 1 kan eleverne besvare spørgsmål, der indeholder velkendte sammenhænge, hvor alle informationerne er til stede, og spørgsmålene er klart formuleret. Eleverne er i stand til at finde informationer og udføre rutineprocedure efter direkte instruktion i eksplicit givne situationer. De kan ligeledes udføre handlinger, der er tydeligt angivet og følger direkte af de givne stimuli.”

I Danmark blev elever på niveau 1 i 2003 karakteriseret som ”funktionelle analfabeter”, der ikke kunne gennemføre en uddannelse eller passe et job. Og det har hængt ved siden.

Presseklip 1. april 2014

- Det betyder ifølge rapporten, at hver femte unge dansker senere i livet får svært ved at passe et job eller bestå en ungdomsuddannelse. Og det er chokerende nyt, mener Niels Egelund, professor på Aarhus Universitet og formand for den danske PISA-komité.
- Resultaterne peger på, at 20 procent af de unge må fravælge uddannelse og i stedet klare sig i ufaglærte job i fremtiden.
- Og det er et stort problem, når andre lande kun har 5-7 procent, der ikke er gode til problemløsning, siger Niels Egelund.

  - DR nyheder

Presseklip 1. april 2014

- Mange unge med indvandrerbaggrund forlader folkeskolen uden tilstrækkelige funktionelle matematikfærdigheder. Mange af dem vil ifølge PISA få problemer med at gennemføre en ungdomsuddannelse eller klare kravene på dagens arbejdsmarked. *(PISA etnisk)*

- »Det er et stort problem, at så mange elever med anden etnisk baggrund end dansk forlader folkeskolen med så lave kompetencer. Det betyder, at de har svært ved at klare sig i uddannelse eller i arbejde efterfølgende«, siger undervisningsminister Christine Antorini (S).

  - Politiken

PISA og tosprogede elever

”Tosprogede drenge og mænd læser elendigt
Det står skidt til for tosprogede drenge og mænd. Det viser nye tal, som forskerne bag den danske del af OECD’s Pisa-undersøgelser har lavet for Politiken og DR P4 København. … Tallene viser blandt andet, at 55 pct. af københavnske tosprogede drenges læsefærdigheder er så dårlige efter 9. klasse, at de betragtes som funktionelle analfabeter.”

Politiken 20.11.2012

”Tosprogede elever halter stadig bagefter
Omkring 50 procent af de tosprogede elever forlader folkeskolen uden at beherske de basale færdigheder til at kunne fortsætte på en ungdomsuddannelse, viser undersøgelsen (PISA 2010), der offentliggøres til april.”

Information 25.03.2011

”Funktionelle analfabeter" får studenterhuer og svendebreve

59 procent af de 9. klasses elever, der ifølge PISA 2007-testen ikke havde funktionelle læsekompetencer, populært kaldet funktionelle analfabeter, havde tre et halvt år senere taget en ungdomsuddannelse eller var i gang med at tage en.

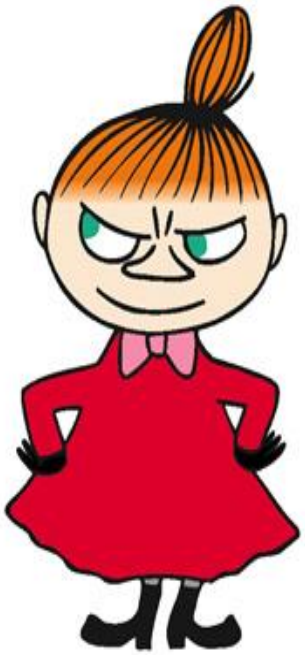
Inddrages socioøkonomiske baggrundsfaktorer viser det sig, at PISA-scorerne hverken i læsning eller matematik kan bruges som indikator for, om de unge falder fra en ungdomsuddannelse.

(*Allerup et al 2012)*

.

Altid 18,8 % dårlige til matematik i PISA

Hvis alle OECD-landene forbedrede deres matematikresultater ville der stadig være 18,8 % funktionelle analfabeter. De ville bare være bedre til at regne



PISA-konsortiet har ikke empirisk belæg for påstanden om, at elever der falder under niveau 1 ikke kan gennemføre en ungdomsuddannelse. Det er en ren teoretisk konstruktion

Peter Allerup har allerede vist at det ikke forholder sig sådan.

Nationale test

- De nationale test er ligesom i PISA baseret på IMR-teknologi dvs. det antages at på hvert profilområde (Tal og algebra, Geometri og måling og Statistik og sandsynlighedsregning) og hvert klassetrin måler alle opgaver ”det samme”. ‘
- Testene er adaptive, dvs. tilpasses elevens svar. Antal rigtige svar bestemmer elevens score.
- Elevens resultater formidles til forældre enten på en norm- eller en kriteriebaseret skala med henholdsvis 5 og 6 trin
- Opgaverne er hemmelige.10% skal udskiftes hvert år. *)
- De nationale test skal skabe ”testkultur” og danne baggrund for tilbagemelding til eleverne

- *) Bliver tilpasning til Rasch-modellen kontrolleret?
- Hvorfor bliver de udskiftede opgaver ikke offentliggjort?

Målsætning

Resultaterne fra de nationale test anvendes også til at følge den faglige udvikling på landsplan. Der er opstillet følgende resultatmål:

- Mindst 80 % af eleverne skal være gode til at læse og regne i de nationale test
- Andelen af de allerdygtigste elever i dansk og matematik skal stige år for år
- Andelen af elever med dårlige resultater i de nationale test for læsning og matematik skal reduceres år for år.

Det nationale mål om at ”Folkeskolen skal mindske betydningen af social baggrund i forhold til faglige resultater” indgår ikke i de national test

”Teaching to the test”

De svageste skal blive bedre år for år.

- Men hvordan måles bedre?
- Via de nationale test (der kun dækker en del af pensum).

Teaching to the test må være en pligt, når mål er formuleret i forhold til testresultaterne.

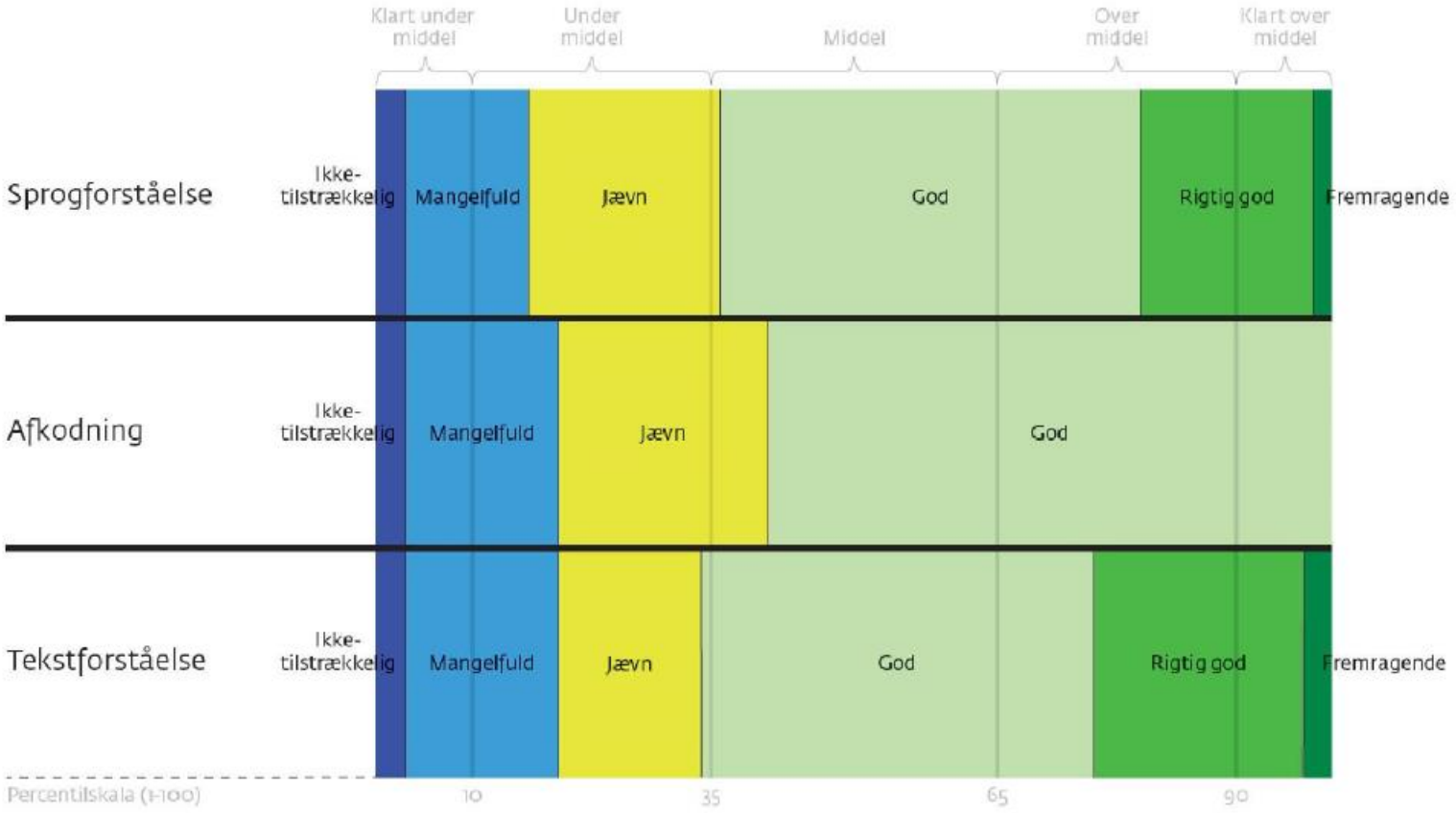
- Målsætningen kan opfyldes, hvis mere og mere tid afsættes til de dele af pensum som kan testes.
- Kan resultaterne sammenlignes år for år?

Norm- og kriteriebaseret tilbagemelding

Normbaseret:	Kriteriebaseret:
En del over gennemsnittet Over gennemsnittet Gennemsnittet Under gennemsnittet En del under gennemsnittet.	Fremragende præstation Rigtig god præstation God præstation Jævn præstation Mangelfuld præstation Ikke tilstrækkelig præstation.

Tilbagemelding til forældre kan fore gå norm- eller kriteriebaseret

Norm- og kriteriebaseret tilbagemelding



Norm- og kriteriebaseret tilbagemelding

Normbaseret

- Elevresultaterne (baseret på antal rigtige) indplaceres på skala bestemt ud fra resultaterne i et bestemt basisår (2015? 2010?).
- En præstation med bedømmelse ”middel” kan være både over og under middel i det aktuelle år. Hvis målsætningen om fremgang lykkes vil den normbaserede tilbagemelding eksempelvis overestimere elevens placering i forhold til aktuelle population.

Kriteriebaseret

- På grundlag af antal rigtige svar i den aktuelle test estimeres, hvordan eleven ville have svaret på en række ”kriterieopgaver” og elevens niveau beregnes ud fra dette. Er stærkt afhængigt af at alle opgaver opfylder IMR-betingelserne.

Elevernes testresultat i de obligatoriske nationale test i matematik sammenholdt med karakteren fra folkeskolens prøve i matematiske færdigheder efter 6. klasse. Andel elever (pct.)

Prædiktiv værdi af nationale test

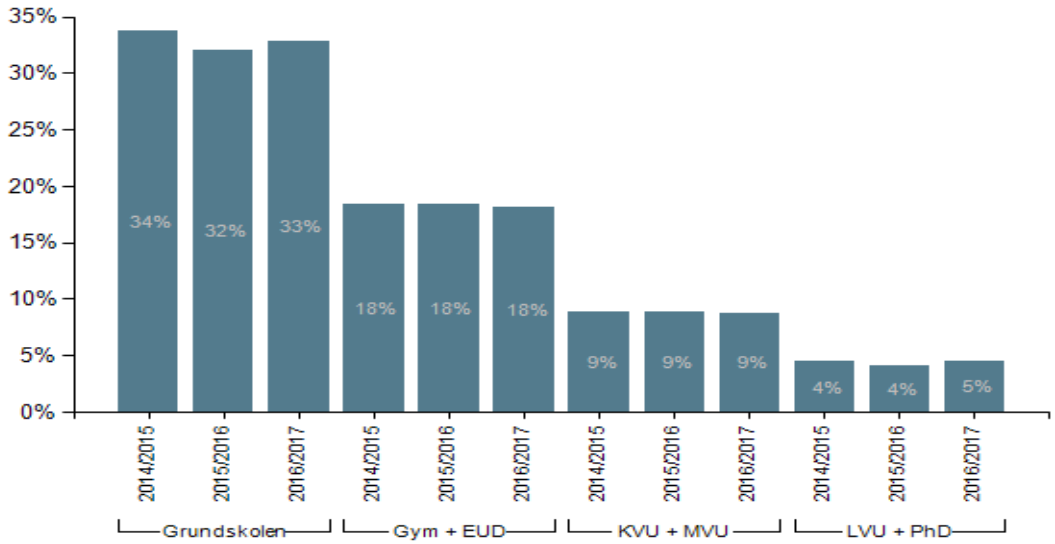
Profil- område		Karakter							
	Testresultat	-3	0	2	4	7	10	12	I alt
Tal og al-	Ikke tilstrækkelig	0	15	33	31	16	4	1	100
gebra	Mangelfuld	0	5	23	35	26	9	2	100
6. klasse	Jævn	0	2	12	29	34	17	6	100
	God	0	1	7	20	34	26	12	100
	Rigtig god	0	0	3	12	28	32	24	100
	Fremragende	0	0	4	12	27	28	29	100
Geometri	Ikke tilstrækkelig	0	13	36	31	15	4	1	100
og måling	Mangelfuld	0	6	24	36	26	7	2	100
6. klasse	Jævn	0	2	13	31	34	17	4	100
	God	0	0	4	15	32	30	19	100
	Rigtig god	0	1	0	13	34	30	21	100
	Fremragende	0	0	12	9	18	35	26	100
Matematik	Ikke tilstrækkelig	0	14	33	33	15	5	0	100
i	Mangelfuld	0	5	24	37	26	7	1	100
anvendelse	Jævn	0	2	15	31	33	15	3	100
6. klasse	God	0	1	7	23	37	23	9	100
	Rigtig god	0	0	2	9	29	35	24	100
	Fremragende	0	0	0	3	13	35	49	100
Samlet	Ikke tilstrækkelig	0	18	37	30	12	3	0	100
6. klasse	Mangelfuld	0	7	28	38	21	5	1	100
	Jævn	0	2	15	34	33	13	3	100
	God	0	1	7	23	37	24	8	100
	Rigtig god	0	0	2	10	30	35	23	100
	Fremragende	0	0	1	5	18	33	42	100

De svageste elever

Nationalt mål

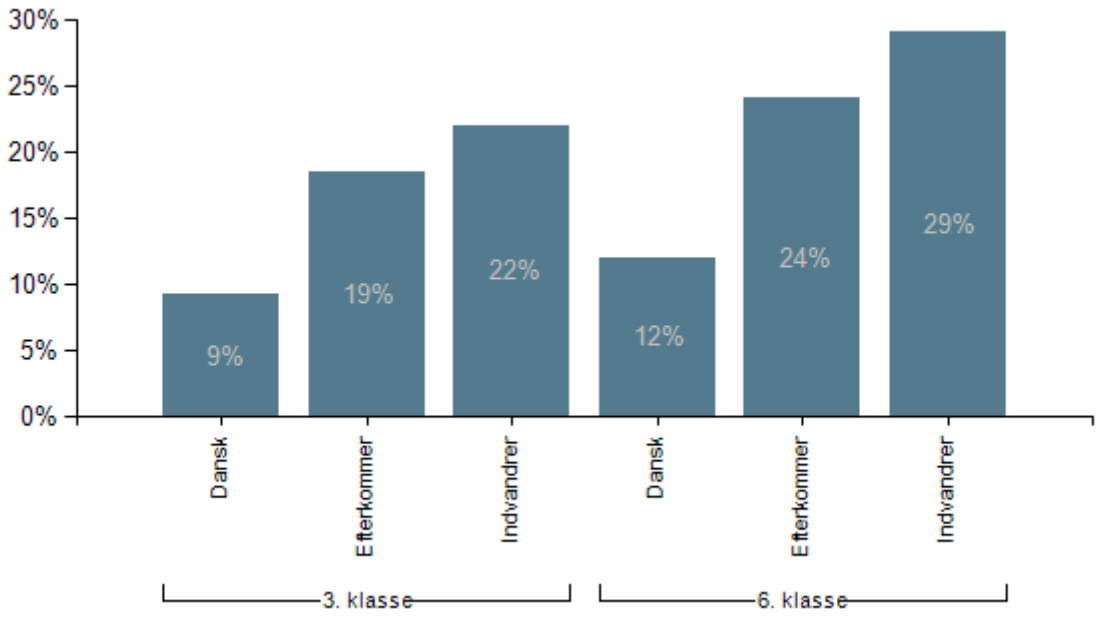
- ”Folkeskolen skal mindske betydningen af social baggrund i forhold til faglige resultater.”
- Indgår ikke i målsætningen knyttet til Nationale test.
- Hvorfor ikke?

- *Andel elever med dårlige resultater i matematik, 6. klasse*



De svageste elever - herkomst

Andel elever med dårlige resultater i matematik, hele landet fordelt på herkomst



Diverse

Validitet, reliabilitet etc:

.

Ved gentagen testning er teoretisk korrelation ca. 0.8. I praksis betydeligt lavere. Uforklaret variation

Adaptive test– ingen elever får det samme test. Testene tilpasses elevens niveau. Smart, men fremmedgørende.

Ambitionen om at få et sikkert resultat giver for de bedste og de dårligste 3-4 timer lange prøver. Er det nødvendigt eller skyldes det uhensigtsmæssig formalisering.

MAT

- Opgaverne er kendte
- Der tilstræbes en bred faglig dækning.
- Bedømmelse er (nogenlunde) transparent.
- Eleven (klassen) i fokus.

- Afrapportering i C-skala med 11 trinr og system med (-,0,+)

- C-skalaen bygger på antagelse om, at "elevers matematik-færdigheder er normalfordelte" og er empirisk i forhold til basisår. Afgrænsning i form af antal rigtige svar (for hvert år og hvert fagligt område) skal i basisåret give (1%, 3%, 7%, 12%, 17%. 20% , 17%, 12%, 7%, 3%, 1%) af eleverne på hvert trin i C-skalaen.
- (-, 0, +)-skalaen skal give (10%, 15%, 75%) på hvert trin.

Eksempel på C-skala

Mat1 hovedområde 1 C-skala

Antal rigtige	C-værdi
0-20	C0
21-42	C1
43-58	C2
59-66	C3
67-72	C4
73-76	C5
77-79	C6
80-82	C7
83-84	C8
85	C9-C10

Elever på niveau C0-C2 (alvorlige indlæringsvanskeligheder) kan besvare over halvdelen af ogaverne rigtigt

C-skalaen

- C0-C2: Tyder på alvorlige indlæringsvanskeligheder
- C3: Usikkert indlært
- C4: Standpunktet er under middel
- C5: Standpunkt er middel
- C6: Standpunkt er over middel
- C7-C10: Standpunktet er betydeligt bedre end almindeligt for klassetrinnet

2 slags udsagn:

- Absolutte C0-C3
- Relative C4-C10

Hvorfor svarer (-.0.+) opdelingen ikke til C-grænserne? (10%, 15%, 75%) og C0-C2: 11%. C3: 12% C4-C10: 77%. Subtile afvejninger eller…?

Mere om C-skalaen

- C-skalaen antager ikke, at alle opgaver måler det samme.
- Det kan derfor have mening at se, hvilke opgaver der ikke er regnet.
- Det kan have mening at se på sammenhæng mellem opgaver og dygtighed.

- Men summen bliver principielt et dårligere mål.
- Kan ikke bruges hvis man vil rangordne eller sammenligne.
- Hvad hvis/når stest bliver high-stakes?