devoteam consulting

De nationale it-baserede test i folkeskolen

Rapport fra REVIEW-panelet

29. oktober 2007

CONNECTING BUSINESS & TECHNOLOGY

Indholdsfortegnelse

1. Indledning 5
1.1. Review af de nationale it-baserede test 5
1.2. Reviewets formål og emner 6
1.3. Reviewets tilrettelæggelse 6

2. Evalueringsportalen 8
2.1. It-baserede prøver og test 8
2.1.1. På tærsklen til et gennembrud 8
2.1.2. Udvikling internationalt 8
2.1.3. Muligheder og begrænsninger i dag 9
2.2. Status for Evalueringsportalen 11
2.2.1. Fremme af evalueringskulturen i Folkeskolen 11
2.2.2. Netstedet 11
2.2.3. Vejledninger og fagspecifikke inspirationsmaterialer 12
2.2.4. It-baserede test 12
2.3. Projektets udfordringer 13
2.3.1. Tidsplan 13
2.3.2. Nedbrud 14
2.3.3. Pressen 14
2.4. Den videre udvikling 14

3. Opbygningen af de nationale test 15
3.1. Introduktion 15
3.2. Opgavebanken 15
3.2.1. Forventningerne 15
3.2.2. Størrelsen af den foreliggende opgavebank 16
3.2.3. Opgaveskabeloner 18
3.2.4. Opgavevarianter 20
3.2.5. Registrerede fejl i opgaver 20
3.3. Udvælgelse af opgaver og estimering af dygtighed 21

4. Testenes psykometriske egenskaber 22
4.1. Fundamentet for testene 22
4.1.1. Introduktion 22
4.1.2. Anvendelsen af de psykometriske metoder 22
4.1.3. Opgaver med flere items (polytome opgaver) 23
4.1.4. Vedligeholdelse af opgavebanken 24
4.1.5. Præstationsscorer 25
4.1.6. Prisen for en-dimensionalitet 25
4.2. Panelets analyser af testene i maj-juni 26
4.2.1. Datagrundlaget for panelets psykometriske vurdering 26
4.2.2. Opgavesværheder og homogenitet 26

4.2.3. Forskydning af niveauer fra afprøvningen til maj-juni ... 27
4.2.4. Profilområder ... 28
4.2.5. Analyse af opgaver og besvarelser ... 29
4.2.6. Varighed og statistisk sikkerhed i praksis ... 29
4.2.7. Bemærkelsesværdige forløb ... 31
4.2.8. Overensstemmelsen mellem behovet og opgaver i opgavebanken ... 33
4.2.9. Opgaveskabeloner ... 36
4.2.10. Kan de nationale test vise forskelle på skoler? ... 36
4.3. Vurdering af de nationale test i internationalt perspektiv ... 37

5. Testenes faglige egenskaber ... 40
5.1. Indledning ... 40
5.2. Generelle observationer ... 42
5.2.1. Opgavevarianter ... 43
5.2.2. Teaching-to-the-test ... 43
5.2.3. Brug af flere items pr. opgave ... 44
5.2.4. Antal opgaver og spredning i opgavesværhed ... 44
5.2.5. Elevforløb ... 49
5.2.6. Distraktorer ... 49
5.2.7. Overskrifter ... 50
5.2.8. Illustrationer, sprog og formuleringer ... 50
5.2.9. Scrollning ... 50
5.2.10. Metadata ... 51
5.2.11. Trinmål i de enkelte testforløb ... 51
5.3. Vurdering af Fysik/Kemi ... 53
5.3.1. Sammenfaldende profilområder ... 54
5.3.2. Opgavevariation ... 54
5.3.3. Illustrationer og sprog ... 55
5.3.4. Distraktorer ... 56
5.3.5. Brug af begreber/terminologi ... 56
5.4. Vurdering af Dansk/Læsning ... 56
5.4.1. Afkodning ... 58
5.4.2. Sprogforståelse ... 58
5.4.3. Tekstforståelse ... 59
5.5. Vurdering af Matematik ... 59
5.5.1. Profilområder ... 59
5.5.2. Validitet ... 60
5.5.3. Tilbagemelding på test ... 60
5.5.4. Opgaveskabeloner ... 60

6. Anbefalinger ... 62
6.1. Forskningstilknytning ... 62
6.1.1. Tættere samarbejde med forskere ... 62
6.1.2. Åben dokumentation af de nationale test ... 62
6.2. Psykometrisk opbygning ... 62
6.2.1. Fastlægge itembank-værdier på rigtig tid og population ... 62

6.2.2. Fastlæggelse af præstationsscores ... 63
6.2.3. Sikkerhed på bedømmelser ... 63
6.3. Opgavebank ... 64
6.3.1. Yderligere kontrol af opgavebanken ... 64
6.3.2. Fornyet afprøvning af afviste opgaver ... 64
6.3.3. Inhomogene opgaver kasseres ... 65
6.3.4. Uddrage information af de afviste opgaver ... 65
6.4. Forløb 65
6.4.1. Bedre fordeling af brug af opgaver ... 65
6.4.2. Sikre afvikling af test på 45 minutter ... 65
6.4.3. Tilbagemelding om usandsynlige elevforløb ... 66
6.5. Opgaver ... 66
6.5.1. Revurder brugen af opgaver med flere items ... 66
6.5.2. Anvend kun varianter uden genkendelsesvirkning, ... 67
6.5.3. Aldrig scrollning ... 67
6.6. Tværgående faglige anbefalinger ... 68
6.6.1. Begræns antallet af opgaveskabeloner ... 68
6.6.2. Antallet af opgaver skal mange-dobles ... 68
6.6.3. Mere stringent opgavekonstruktion ... 69
6.6.4. Systematisk anvendelse af metadata ... 69
6.7. Specielt om Dansk/Læsning ... 69
6.7.1. Fasthold profilområder ... 69
6.7.2. To opgavetyper per profilområde ... 70
6.8. Specielt om Matematik ... 70
6.8.1. Fasthold profilområder ... 70
6.8.2. Analyser validitet ... 71
6.8.3. Revurder af opgaveskabeloner ... 71
6.9. Specielt om Fysik/Kemi ... 71
6.9.1. Fasthold profilområder, ændre opgavekonstruktionen ... 71
6.9.2. Inddrag flere emner ... 71
6.9.3. Anvend kun entydige distraktorer ... 71
6.9.4. Opstramning af illustrationer og sprog ... 72

1. Indledning

1.1. Review af de nationale it-baserede test

I sommeren 2006 indgik Undervisningsministeriet aftale om etablering af en evalueringsportal som en del af den indsats, der i de kommende år skal styrke evalueringskulturen i folkeskolen.

Aftalen blev indgået med COWI A/S, der leder et konsortium bestående af bl.a. JCVU, CVU Storkøbenhavn og @ventures.

Som led i evalueringsportalen skal konsortiet udvikle i alt 10 nationale it-baserede test til folkeskolen. Den første version af de tre test i Dansk/Læsning (8. klasse), Matematik (6. klasse) og Fysik/Kemi (8. klasse) blev taget i brug i maj-juni 2007.

For at sikre at erfaringerne med den første runde af testene bliver opsamlet og vurderet, har Skolestyrelsen aftalt med konsortiet, at der skal gennemføres et review af de nationale test.

Reviewet gennemføres af en række forskere og andre, der er eksperter inden for de enkelte fag, pædagogik og test. Ekspertpanelet er udpeget i fællesskab mellem Ministeriet og konsortiet.

Panelets medlemmer er:

- Lektor Helene Sørensen, Fysik/Kemi*
- Fhv. seminarierektor Peter Norrild, Fysik/Kemi
- Lektor Dorthe Klint Petersen, Dansk/Læsning*
- Professor Carsten Elbro, Dansk/Læsning
- Lektor Arne Mogensen, Matematik*
- Udviklingskonsulent Kim Foss Hansen, Matematik
- Adjunkt Lars Stenius Stæhr, Engelsk/Almen pædagogisk tilgang*
- Professor Niels Egelund, Almen pædagogisk tilgang
- Lektor Svend Kreiner, Psykometri*
- Professor Peter Allerup, Psykometri

De med stjerne markerede deltagere har i et vist omfang været involveret i udviklingen af testene som rådgivere for konsortiet, mens de øvrige deltagere tilsvarende i et vist omfang har været involveret som rådgivere for Ministeriet.

Devoteam Consulting A/S har fungeret som sekretariat for panelet og har udarbejdet denne rapport på panelets vegne på basis af panelets undersøgelser og drøftelser.

1.2. Reviewets formål og emner

Review-panelets opgave er todelt:

1. Panelet skal komme med eventuelle forslag til ændringer af de nationale test, som review-panelet måtte nå frem til vil kunne give en bedre overordnet opfyldelse af målet med de it-baserede test
2. Panelet skal bidrage til afdække, i hvilket omfang leverancen af de it-baserede test, herunder opgavebanken, lever op til kravene i kontrakten mellem Ministeriet og konsortiet

Formålet med denne rapport er at redegøre for panelets vurdering af de nationale test og for panelets forslag til forbedringer. Panelets gennemgang af opfyldelsen af de enkelte krav i kontrakten vil blive afrapporteret direkte til styregruppen for projektet.

Reviewet handler om de test-mæssige aspekter ved de it-baserede test, baseret på en vurdering af de faktisk gennemførte test i maj og juni 2007.

Panelet har ikke skullet vurdere den it-mæssige eller den vejlednings-mæssige side af testene. Panelet har heller ikke haft til opgave at vurdere omkostningerne ved - eller effekten af - de nationale test.

1.3. Reviewets tilrettelæggelse

Den overordnede tidsramme har været fastlagt af Skolestyrelsen og konsortiet således, at reviewet skulle være afsluttet med udgangen af oktober måned. Denne tidsramme har sammenhæng med, at observationer og anbefalinger fra reviewet gerne skulle have en effekt i forhold til arbejdet med de kommende test.

Det har været en del af aftalen mellem Skolestyrelsen og konsortiet, at review-panelet har arbejdet uafhængigt af såvel Skolestyrelsen som konsortiet. Panelet har således selv, indenfor rammerne af kommissoriet, tilrettelagt sine undersøgelser og selv prioriteret, hvilke observationer og anbefalinger, der skal fremlægges.

Som grundlag for arbejdet har panelet kunnet rekvirere oplysninger fra Skolestyrelsen og fra konsortiet.

Panelet mødtes første gang den 28. juni 2007 og drøftede prioriteringen og fokuseringen af panelets arbejde, herunder spørgsmålet om hvilke redegørelser og andre oplysninger, det ville være realistisk at rekvirere fra konsortiet eller Skolestyrelsen.

Panelet valgte at fokusere sin indsats på følgende to fokusområder:

1. Hvordan fungerer de adaptive test i praksis
2. Kvaliteten af opgavebanken

Som grundlag herfor har panelet rekvireret - og i slutningen af august modtaget - et anonymiseret udtræk af ca. 12.000 testforløb og af alle opgaver. Panelet har endvidere modtaget nogle redegørelser og andet materiale fra konsortiet.

På grundlag af det tilvejebragte materiale har panelet gennemført egne analyser og holdt tre møder i september og oktober 2007.

Forløbet har været presset af hensyn til den overordnede tidsramme for reviewet og de nationale test. Med mere tid til panelets arbejde, og mere tid for konsortiet til at foretage undersøgelser i datamaterialet, som panelet kunne have ønsket sig, kunne nogle af panelets anbefalinger måske være blevet skærpet, og andre emner kunne have været berørt. Panelet vurderer dog, at man har fået belyst og vurderet det væsentligste.

Devoteam Consulting har samlet panelets observationer og anbefalinger i denne rapport, der er endeligt afstemt efter en skriftlig høring.

2. Evalueringsportalen

2.1. It-baserede prøver og test

2.1.1. På tærsklen til et gennembrud

I foråret 2007 har begreber som it-baserede prøver og test ikke givet anledning til en optimistisk tro på nye muligheder for tidssvarende og mere retfærdige vurderinger af resultaterne af uddannelsesforløb. Tekniske problemer ved afviklingen af folkeskolens it-baserede afgangsprøve i naturfag og forsinkelser i tilbagemeldingen fra de nye nationale test har været et gennemgående tema i den pædagogiske debat. Dette har også ført til, at skeptikernes opfattelse er blevet styrket:

- at it-baserede prøver, test og eksaminer kun kan dække paratviden, mens analytisk og vurderende kompetence ikke kan bedømmes.

Børnesygdomme ved den anden afvikling af it-baserede afgangsprøver og indførelsen af et – også på verdensplan – revolutionerende adaptivt testsystem bør dog ikke lede til, at man straks forkaster al nytænkning med hensyn til it-baserede prøver og test.

Disse redskaber har haft begrænset nytte af it-teknologien indtil nu, og papir og pen-bårne opgaver, mundtlige fremstillinger og diskussioner og større skriftlige redegørelser for projekter har været den dominerende vurderingsform. Det må i den forbindelse også erkendes, at formen er faldet godt i tråd med den traditionelle lærebogsorienterede, gruppe- og klassebaserede undervisning, hvor løbende intern evaluering af læringsprogression har været præget af tilfældigheder og manglende skriftlighed.

Der er imidlertid gennem de seneste år sket en udvikling, dels på den psykometriske side, dels på det it-mæssige felt, der gør, at nye og it-baserede test er realistiske og nyttige alternativer.

2.1.2. Udvikling internationalt

På den psykometriske side har bl.a. arbejdet med internationale komparative studier, f.eks. PISA, betydet, at brugen af statistiske metoder[1] har givet en betydelig sikkerhed i afgørelsen af de sværhedsgrader, som test- og prøveitems har.. Man

[1] Først og fremmest med Rasch modeller og IRT-skalering

kan således med stor sikkerhed udlede valide indikationer på en elevs eller en studerendes færdighed ud fra løsningen af et relativt begrænset sæt af opgaver.

De store internationale komparative undersøgelser har også vist, at det er muligt at gennemføre valide vurderinger af åbne spørgsmål, grafiske fremstillinger og tegninger, hvorved det bliver muligt at lade analytisk og syntetiserende evne indgå i en kvantitativ testning. Hvis man anvender multimediebåret simulation, er det også muligt at vurdere originalitet og kreativitet. Vurderinger af åbne spørgsmål og simulation sker på baggrund af en omhyggeligt udformet kodevejledning og efter træning af de personer, der skal kode de åbne opgaver – en testform, der dog ikke indgår i de danske nationale test.

De internationale komparative undersøgelser har været gennemført som papir og pen-test, men der har i PISA2006 været en mulighed for at gennemføre en it-baseret test i naturfag, som imidlertid kun tre lande har betjent sig af, Sydkorea, Island og Danmark – i øvrigt lande, der alle ligger langt fremme, hvad it angår. I PISA2009 vil der indgå en mulighed for at gennemføre en it-læsetest, der gennem en simuleret browser udnytter alle de læsemæssige og tekstmæssige muligheder, it giver. I sommeren 2007 har seksten lande, herunder Danmark, ytret interesse for at gennemføre it-læsetesten. Ud over at give informationer om antallet af rigtige opgaveløsninger rummer de it-baserede test også oplysninger om responstider, der kan give selvstændig information om elevkompetencer. Det er holdningen i det konsortium, der gennemfører PISA, at man fra 2012 helt skal overgå til elektronisk testning. Den væsentligste hindring er det it-udstyr og den båndbredde, der er til rådighed på skolerne i de godt 60 lande, der deltager i PISA.

På den it-mæssige side har udviklingen af adaptive testsystemer givet nye udfordringer og muligheder. Disse er emnet for andre dele af denne rapport.

2.1.3. Muligheder og begrænsninger i dag

På det udviklingsniveau, it-baserede prøver og test har i 2007, vil anvendelse af åbne spørgsmål og simulation ikke være realistisk i større landsdækkende undersøgelser. Det vil på nuværende tidspunkt kræve så store ressourcer og et så stort et tidsforbrug, at det ikke vil være realistisk i forhold til forventninger om hurtig – eller øjeblikkelig – tilbagemelding.

Det er derfor nødvendigt kun at anvende test, der kan vurderes maskinelt, og som derfor består af spørgsmål, hvor svarenes rigtighed er defineret på forhånd, idet der dog kan være tale om, at svar kan være delvis rigtige.

Historisk har nødvendigheden af skrift og tal været en begrænsning ved brugen af it. I dag byder it dog på andre muligheder, og derfor indeholder it-baserede test og prøver da også andre muligheder, f.eks. ”drag and drop” på skærmen. I den forbindelse kan alt indgå i drag and drop – billeder, ord, tal eller bokse eller en kombination af disse. Man kan også lade den testede samle dele til en figur eller til et diagram. Introduktionen af bevægelse på skærmen gør det muligt at teste en mangfoldighed af færdigheder. Ligeledes kan lyde og videoklip indgå i testningen. Optisk læsning af tekster, stemmegenkendelse og syntetisk tale åbner mange muligheder, ikke bare for elever med ordblindhed.

De centrale processer, der kan afprøves med det udviklingsniveau, it i dag har, er:

- tilstedeværelse af konkret viden, herunder evnen til at kunne nævne, genkende, definere, identificere.
- tilstedeværelse af forståelse, herunder at kunne forklare, formulere, beregne, fortolke, opstille og beskrive.
- evnen til at anvende, herunder at kunne vælge, løse, afprøve og behandle.
- evnen til at analysere, herunder at kunne sammenligne, skelne, udvælge, uddrage og opdage.
- evnen til at syntetisere, herunder at kunne generere hypoteser, finde regler, finde relationer og planlægge forløb.
- evnen til at vurdere, herunder evnen til at kunne bedømme, overveje, diskutere, afgøre og forsvare.
- evnen til at kommunikere, herunder at kunne udtrykke sig skriftligt i form af tekst, diagrammer, skemaer, tal og tabeller.

Ovennævnte centrale processer og deres elementer indgår på de allerfleste videnområder, ikke bare i naturfaglige sammenhænge, men også i brugen af sprog, i matematik og i samfundsfaglige og humanistiske sammenhænge.

Den første version af de nationale test bygger imidlertid helt overvejende på de processer, der er nævnt under første punkt ovenfor.

En klar begrænsning er, at it i den form, der findes i dag, ikke kan vurdere sproglige kommunikative kompetencer, kreativitet og originalitet, evne til samarbejde og mere personlige kompetencer. Derfor kan - og skal - it ikke stå alene i vurde-

ringen af kompetencer, men må være et supplement til andre former for vurdering.

2.2. Status for Evalueringsportalen

2.2.1. Fremme af evalueringskulturen i Folkeskolen

I 2006, da Undervisningsministeriet indgik aftale om 'Fremme af evalueringskulturen i folkeskolen', var det med ambitionen om at give evalueringskulturen i folkeskolen et løft og samtidig give lærerne en række værktøjer til at udnytte evalueringen således, at også undervisningen af den enkelte elev påvirkes positivt.

Med aftalen har Undervisningsministeriet skabt grundlaget for udvikling af værktøjer til understøttelse af den løbende interne evaluering og undervisningsdifferentiering i folkeskolen. Evalueringen skal danne grundlag for vejledning af den enkelte elev og for den videre planlægning og tilrettelæggelse af undervisningen samt for underretning af forældrene om elevens udbytte af undervisningen.

Udviklingen af Evalueringsportalen er i dag nået så langt, at en lang række evalueringsværktøjer nu ligger frit tilgængelige på det officielle website. Til gengæld må projektet også konstatere, at det sidste år har været stormomsust, og at projektet igen og igen har befundet sig i mediernes søgelys. I de nedenstående afsnit redegøres kort for status på de enkelte områder.

2.2.2. Netstedet

Netstedet www.evaluering.uvm.dk blev etableret i marts måned som det første synlige bevis på, at en række evalueringsværktøjer var under etablering. Netstedet rummer information til lærere og ledere, til elever og forældre samt til kommuner, skolebestyrelser og pressen.

På netstedet findes vejledninger og inspiration til at finde ud af hvilke evalueringsværktøjer, der passer til hvilken undervisning. Man kan læse artikler om evaluering generelt og se eksempler på, hvordan andre har gjort. Desuden kan man læse om styrker og svagheder ved forskellige evalueringsværktøjer, samt om de it-baserede, adaptive test.

Netstedet er med andre ord en indgang til samtlige evalueringsværktøjer og til information om, hvordan de anvendes. Samtidig rummer netstedet information fra projektet bag 'Fremme af evalueringskulturen i folkeskolen'.

Ud over en fokus på information og værktøjer ifm. evaluering i folkeskolen, indeholder netstedet også information og værktøjer fra en række af Undervisningsministeriets øvrige initiativer. Et eksempel herpå er CEPRA (Videnscenter for evaluering i Praksis), der har opbygget en database med en national oversigt over udbuddet af efteruddannelsestilbud for lærere og ledere inden for evaluering. CEPRA har desuden udviklet 30 helt nye efteruddannelsestilbud til ledere, lærere og pædagoger. De kan også kan findes på netstedet.

2.2.3. Vejledninger og fagspecifikke inspirationsmaterialer

Netstedet fungerer som indgang til de centrale og væsentlige dele af den værktøjskasse, som 'Fremme af evalueringskulturen i folkeskolen' udgør.

Der er etableret et vejledningsmateriale, der henvender sig særligt til lærere og ledere. Det omhandler løbende intern evaluering, forskellige evalueringsværktøjer, formidling af evalueringsresultater til elever og forældre og beskæftiger sig også med evalueringens betydning for den videre planlægning og tilrettelæggelse af undervisningen.

Derudover er der etableret fagspecifikke inspirationsmaterialer rettet mod alle folkeskolens obligatoriske fag og emner samt tilbudsfag. Materialerne indeholder konkrete anvisninger og eksempler, der kan støtte lærerens valg af evalueringsværktøjer og valg af måder at formidle evalueringsresultater på. Samtidig findes her også viden om, hvordan man følger op på evalueringsresultaterne gennem den videre tilrettelæggelse af undervisningen.

2.2.4. It-baserede test

De it-baserede test er dét enkeltstående evalueringsværktøj, der er blevet ofret mest opmærksomhed i det forgangne år. Det skyldes ikke mindst, at sådanne adaptive test ikke tidligere er set i de danske folkeskoler. Det er samtidig det evalueringsværktøj, der fylder mest i folkeskole-debatten.

De it-baserede test blev lanceret i fagene Matematik, Fysik/Kemi og Dansk/Læsning på hhv. 6. og 8. klassetrin. De folk, der var ansvarlige for testene, oplevede ved lanceringen en del udfordringer, herunder at både elever og lærere for første gang skulle gennemføre et it-baseret testforløb, at opgaverne skulle gennem deres ilddåb, og at tilbagemeldingerne til lærerne blev forsinkede.

2.3. Projektets udfordringer

Projekt 'Fremme af evalueringskulturen i folkeskolen' har i det hele taget haft en del at slås med undervejs, Herunder udfordringer, der ikke mindst er opstået som følge af projektets rammer. Hele baggrunden for projektet er, at der skal skabes fokus på, hvordan folkeskolen implementerer en god evalueringskultur, men samtidig er der krav om, at der skal skabes resultater inden for en forholdsvis kort periode.

2.3.1. Tidsplan

Projektets tidsplan har været ambitiøs. Man er gået efter leverancer på forskellige tidspunkter af skoleåret, og tidsplanen har på visse strækninger også været meget stram. Det skyldes ikke mindst ambitionen om på ganske kort tid at skabe store resultater. Med resultater menes, at der ligger en ambition i projektet om, at de enkelte folkeskoler får implementeret en god evalueringskultur allerede i skoleåret 2006/2007. Projektets formål er at understøtte implementeringen og at tilvejebringe de rigtige evalueringsværktøjer.

Netop tidsplanen har været drøftet på det seneste og er netop blevet justeret på en række punkter. Formålet med denne justering har været, at tidsplanen i sig selv ikke må stå i vejen for, at der skabes gode evalueringsværktøjer af høj kvalitet. Ændringen af tidsplanen skal dermed skabe større sikkerhed og kvalitet i leverancen af de nationale test.

I indeværende skoleår ser tidsplanen efter justeringen således ud:

- *oktober 2007 – januar 2008:* Frivillig anvendelse af test for folkeskoler og frie grundskoler. Der vil være mulighed for at anvende de tre test, som folkeskolerne gennemførte i maj og juni 2007: Matematik 6. klasse, Dansk/Læsning 8. klasse og Fysik/Kemi 8. klasse. Booking af tider til frivillige test er mulig fra 15. august 2007.
- *marts - april 2008:* Obligatoriske test i: Matematik, 6. klasse, Dansk/Læsning, 8. klasse og Fysik/Kemi, 8. klasse.
- *maj - juni 2008:* Obligatoriske test i: Dansk/Læsning, 2., 4. og 6. klasse, Matematik, 3. klasse,

2.3.2. Nedbrud

I maj måned lancerede Skolestyrelsen folkeskolens afgangsprøve i biologi i digital form Afviklingen af prøven løb ind i massive tekniske problemer, hvilket medførte aflysning af den ligeledes digitale afgangsprøve i geografi.

Aflysningen skete med udgangspunkt i, at Skolestyrelsen ikke fandt det forsvarligt at udsætte elever og lærere for yderligere frustrationer. Da den digitale prøve i geografi var planlagt til at blive afviklet dagen efter den digitale prøve i biologi, kunne det ikke garanteres, at fejlene ikke ville opstå igen.

Nedbruddet kom på et yderst uheldigt tidspunkt for projektet, idet det blev sat i forbindelse med afviklingen af de nationale test. Og dermed skabte stor usikkerhed om de nationale tests formåen.

2.3.3. Pressen

'Fremme af evalueringskulturen i folkeskolen' har i det forgangne skoleår ofte været nævnt i pressen. Det skyldes ikke mindst de ovenfor nævnte udfordringer. Det gennemgående i presseomtalen har været bekymringen for, om testene nu også virker, som de skal, om eleverne skal udsættes for dem - og om de overhovedet kan give det positive bidrag til undervisningen, som det er meningen, de skal.

2.4. Den videre udvikling

Projektet befinder sig stadig i en hektisk udviklingsfase. Der udestår at blive leveret test i de resterende fag/trin, ligesom vejledninger og fagspecifikke inspirationsmaterialer også skal etableres for disse fag/trin.

De mange tilbagemeldinger fra folkeskolerne og de mange analyser og undersøgelser, der er gennemført i projektet på baggrund af maj/juni testen, skal indarbejdes i det grundlag, der udgør testene. Herunder også, at resultaterne fra denne rapport skal indgå i den videre udvikling.

3. Opbygningen af de nationale test

3.1. Introduktion

De nationale test er it-baserede, selvscorende og adaptive. Centrale elementer i realiseringen af disse test er

- en *opgavebank*
- en *adaptiv* metode til udvælgelse af passende opgaver fra banken til den enkelte elev
- en metode til *estimering af elevens dygtighed* på grundlag af elevens besvarelse af opgaverne.

Det er disse elementer, som skal vurderes i dette review.

De nationale test består også af en række andre elementer, såsom de it-mæssige løsninger, der gør det muligt at afvikle testen over internettet, og de vejledninger til lærere m.fl., der knytter sig til testene. Disse elementer er ikke omfattet af dette review.

De nationale test skal være valide og reliable, hvilket stiller krav til den måde, opgavebanken og de enkelte opgaver er udformet på, ligesom det stiller krav til de anvendte metoder. Forventningerne til opgavebanken og dens aktuelle indhold beskrives i det følgende afsnit 3.2, og forventningerne til den adaptive metode og dygtighedsestimeringen beskrives i afsnit 3.3. Panelets vurdering af testene beskrives i kapitel 4 og 5.

3.2. Opgavebanken

3.2.1. Forventningerne

Opgavebanken skal rumme et antal opgaver for hvert fag med forskellige emner og forskellige sværhedsgrader.

Som grundlag for udviklingen af opgavebanken er der blandt fagets trinmål[2] udvalgt de trinmål eller dele heraf, der egner sig til it-baserede test. Opgaverne i banken skal relatere sig til disse trinmål, og skal inden for disse i videst muligt omfang inddrage fagets kundskaber og færdigheder[3].

[2] Trinmål anvendes her som betegnelse for de slut- og trinmål, der indgår i de bindende fælles nationale mål i form af fagformål, centrale kundskabs- og færdighedsområder (slutmål) og trinmål.
[3] De udvalgte trinmål fremgår af afsnit 5.2.4

Inden for hvert fag er der defineret tre profilområder, der omfatter forskellige områder af faget og/eller forskellige aspekter af elevernes dygtighed[4]. Bedømmelsen af eleven skal gives som en *profil* med en uafhængig bedømmelse af præstationen inden for hvert af de tre profilområder på en skala fra 1-5, f.eks. en profil med scoren 4-3-5. Skalaen fra 1-5 skal fastlægges således, at 5 svarer til de bedste præstationer, og således at eleverne fordeler sig procentvis på trinene med: 10-25-30-25-10.

Da hver elev løser et individuelt udvalg af opgaver, er det af største vigtighed, at opgaverne i opgavebanken er homogene, dvs. at opgaverne alle tester den samme færdighed (inden for hvert profilområde) og kun denne. Nedenfor i kapitel 4 diskuteres dette nærmere.

Homogeniteten af opgavebanken og de enkelte opgavers sværhed er blevet fastlagt efter en såkaldt *afprøvning* i december 2006 og februar/marts 2007. Til afprøvningen blev opgaverne samlet i test-sæt bestående af opgaver svarende til ca. 30 minutters test. Grupper af 5 test-sæt blev derefter afprøvet på en gruppe af ca. 600 elever. I alt er der blevet løst opgaver af ca. 11.000 elever på ca. 230 skoler landet over. På grundlag af disse elevers besvarelser har det ved hjælp af psykometriske (statistiske) metoder (den såkaldte Rasch-model) været muligt at udvælge de opgaver, der opfylder kravene til opgavebanken, og at vurdere opgavernes relative sværhedsgrad.

Den praktiske tilrettelæggelse af denne afprøvning ligger uden for kommissoriet for reviewet, men de anvendte psykometriske metoder vurderes nedenfor i kapitel 4.

Resultatet af afprøvningen er en opgavebank med opgaver, der er godkendte som homogene. Opgavebanken vurderes ligeledes i kapitel 4 - samt i kapitel 5.

3.2.2. Størrelsen af den foreliggende opgavebank

Der blev i december 2006/februar-marts 2007 afprøvet i alt godt 2.000 opgaver med tilsammen godt 4.000 spørgsmål (items).

[4] Profilområderne fremgår af Tabel 3.1

Afprøvningen viste imidlertid, at kun 760 opgaver med ca. 2.000 spørgsmål (items) kunne godkendes som homogene inden for deres profilområde og indgå i opgavebanken. For profilområderne Dansk/Læsning-Afkodning og Matematik-Tal og algebra blev så få opgaver godkendt, at det blev besluttet ikke at lade disse profilområder indgå i testene før end februar 2008.

Antallet af opgaver og items fremgår af Tabel 3.1

Tabel 3.1Antal opgaver og antal items i opgavebanken med fordeling på profilområder

Profilområde	Antal opgaver	Antal items	Svartid (min)*
Sprogforståelse	118	232	28,9
Afkodning	0	0	0
Tekstforståelse	100	236	87,6
Dansk/læsning i alt	218	468	116,5
Tal og algebra	0	0	0
Geometri	127	322	93,2
Matematik i anvendelse	63	199	94,6
Matematik i alt	190	521	187,8
Energi og energiomsætning	112	331	71,1
Stoffer og materialer	150	446	98,1
Anvendelse og perspektiver	90	270	59,1
Fysik/kemi i alt	352	1047	228,4
Opgaver i alt	760	2036	532,7

* Gennemsnitlig observeret svartid i afprøvningen

Det kvantitative udgangspunkt for opbygningen af opgavebanken er et krav om, at der for hvert fag skal være godkendte opgaver med en samlet løsningstid på mindst 675 minutters løsningstid (for en middelhurtig elev). Dette minimum skal være fordelt ligeligt på profilområder og på fem sværhedsklasser, svarende til den ovennævnte indplacering af eleverne på niveau 1-5. Størrelsen af opgavebanken skal blandt andet sikre, at der er opgaver nok af den rette sværhedsgrad for elever, der tager flere test i løbet af året.

Tabel 3.1 viser også erfaringerne med løsningstiderne for opgaverne i afprøvningen. Opgjort således indeholder opgavebanken mellem en femtedel og en tredjedel af det forventede antal opgaver. Panelet har ikke haft mulighed for at undersøge, om løsningstiderne under testene i maj-juni afveg fra tiderne i afprøvningen, eller i hvilket omfang, opgavebanken opfylder det kvantitative krav til størrelse.

I afsnit 4.2.8 vurderes, hvilken betydning opgavebankens faktiske størrelse har haft for testene i maj-juni.

3.2.3. Opgaveskabeloner

De enkelte opgaver er udformet i et antal forskellige opgaveskabeloner, det giver opgavekonstruktøren forskellige muligheder for at vælge på hvilken måde, billeder og tekst indgår, og hvordan eleven skal svare (f.eks. ved at klikke med musen eller at skrive svaret i en svarrude).

Forventningen er, at de forskellige opgaveskabeloner dels understøtter de forskellige opgavetyper, der er relevante inden for de forskellige emner i testene, dels medvirker til at skabe variation i testene og i testsituationen for den enkelte elev.

Anvendelsen af opgaveskabeloner inden for profilområder fremgår af Tabel 3.2

Tabel 3.2 Opgaver fordelt på testopgavetyper (skabeloner)

		Profilområde						
Testopgavetype	Antal opgaver i alt	Sprog-forståelse	Tekst-forståelse	Geometri	Matematik i anvendelse	Energi og energiom-sætning	Stoffer og materialer	Anvendelse og perspektiver
Indsættelse	58	8	9	6	3	14	13	5
Indtastning	47			24	23			
Kemireaktionsligning	2					2		
Markering af tekst	5	5						
Match	22	2		5	1		12	2
Multiple choice i skemaform	232	14	18	33	15	48	57	47
Multiple choice Løbende valg	23	3	20					
Multiple choice med flere korrekte svar*	359	86	51	55	18	48	65	36
Ret en tekst	2		2					
Rigtig form	10			4	3		3	
	760	118	100	127	63	112	150	90

*) Omfatter også opgaver med 1 korrekt svar

Det ses, at der i den foreliggende opgavebank i alt anvendes opgaver af 10 forskellige opgavetyper. Forskellige udformninger af multiple choice anvendes i 614 ud af de 760 opgaver. Følgelig er multiple choice også den dominerende opgavetype inden for hvert profilområde.

Beskrivelse af opgavetyperne fremgår af Figur 3.1. Af hensyn til hemmeligholdelsen af opgaverne er det desværre ikke muligt at bringe illustrationer af de forskellige opgavetyper i denne rapport.

Figur 3.1 Beskrivelse af testopgavetyper (skabeloner)

Testopgavetype	Kort beskrivelse	Antal item
Multiple choice med flere korrekte svar	Opgavetypen består af et spørgsmål og nogle svarmuligheder, hvoraf der er flere korrekte. Der kan sættes nøjagtig lige så mange krydser, som der er korrekte svar.	Antal korrekte svarmuligheder
Multiple choice i skemaform	Opgavetypen består af en instruktionstekst efterfulgt af et skema med en Multiple choice pr. række. Hver Multiple choice har præcist et korrekt svar.	Antal rækker med svarmuligheder
Multiple choice Løbende valg	I Opgavetypen skal eleven udpege ét valg blandt fire i en løbende tekst, f.eks. det rigtige udtryk.	Antal valgmuligheder i den løbende tekst.
Indsættelse	Opgavetypen består af en instruktionstekst efterfulgt af en tekst med tomme felter og en række indsættelseselementer, som vha. træk og slip kan placeres i de tomme felter.	Antal indsættelsesord.
Match	Opgavetypen er en instruktionstekst efterfulgt af to søjler, hvor elementerne i rækkerne i den første søjle skal Parres/matches med elementerne i den anden søjle. Elementerne i anden søjle præsentere i forskellig rækkefølge fra gang til gang. Vha. træk og slip skal elementerne parres/-matches.	Antal parringsmuligheder
Indtastning	Opgavetypen består af en instruktionstekst efterfulgt af en venstresøjle med tekst (f.eks. foreslået stavemåde, et regnestykke eller en formel) og en højresøjle, hvor de korrekte svar skal indtastes. Det korrekte svar skal skrives korrekt med store og små bogstaver og talangivelser de korrekte steder.	Antal korrekte svarmuligheder
Rigtig form	Rigtig form består af en instruktionstekst efterfulgt af en tekst med elementer som mangler og skal indtastes. Ved hvert indtastningsfelt kan skrives en ledetekst til det, der skal indtastes.	Antal ord, der skal indtastes
Ret en tekst	Opgavetype, hvor en fejlbehæftet tekst skal rettes.	Antal forkerte ord, der skal rettes
Markering af tekst	Marker bestemte ord i en tekst (f.eks. navneord, ord i flertal eller lign).	Antal ord, der skal markeres
Kemireaktionsligning	I tomme felter på hver side af et lighedstegn skal indsættes kemiske tegn, der gør ligningen korrekt. De kemiske tegn vælges i en mængde af tegn.	1

Anvendelsen af opgavetyper i testene vurderes særskilt for hvert fag nedenfor i kapitel 5.

Testene er opbygget således, at alle opgaver har et eller flere spørgsmål. Hvert spørgsmål skal være udformet således, at der kun er to muligheder: Enten er elevens svar *rigtigt,* eller også er det *forkert*. Et sådant spørgsmål kaldes et *item.*

Tabel 3.3 viser, at ca. halvdelen af opgaverne er udformet med udelukkende *et* item, mens resten af opgaverne har op til 10 items. Andelen af opgaver med flere items er relativt mindst i Dansk/Læsning og relativt størst (ca. 2/3) inden for profilområderne Matematik i anvendelse og Stoffer og materialer (Fysik/Kemi).

Tabel 3.3 Antal opgaver fordelt efter profilområde og antal items i opgaven

	Opgaver	Antal items pr opgave									
Profilområde	i alt	1	2	3	4	5	6	7	8	9	10
Sprogforståelse	118	87	5	6	10	3	2				5
Tekstforståelse	100	59	6	7	9	13	3	1	1		1
Geometri	127	63	16	9	11	16	10	1		1	
Matematik i anvendelse	63	20	1	21	8	4	4	3	1	1	
Energi og energiomsætning	112	53	7	7	8	20	11	3	3		
Stoffer og materialer	150	66	4	8	30	28	10	4			
Anvendelse og perspektiver	90	36	3	5	20	24	1	1			
Opgaver i alt	760	384	42	63	96	108	41	13	5	2	6

Antallet af items pr. opgave har betydning for flere forhold:

- løsningstiden for hver opgave og hvert item, og dermed for hvor mange items, der kan indgå i en test på 45 minutter.
- den faglige spredning af opgaverne, der kan blive tid til inden for en test på 45 minutter
- værdien af den information om dygtigheden, som opgaven giver
- muligheden for at udvælge opgaver med passende sværhedsgrad til hver elev

Vurderingen af disse forhold gennemgås i afsnittene 4.1.3 og 5.2.3.

3.2.4. Opgavevarianter

En række af opgaverne i opgavebanken er udviklet som varianter af en grundopgave. Grundlaget kan være en fælles tekst eller tegning, hvor temaer og spørgsmål varieres.

I et vist omfang fremgår dette forhold direkte af opgaveidentifikationen. Ca. 590 opgaver er angivet som varianter af 210 grundopgaver, mens 170 opgaver ikke optræder med varianter i den foreliggende opgavebank. I andre tilfælde optræder den samme tekst f.eks. i flere opgaver, uden at det kaldes en variant.

Anvendelsen af varianter i opgavebanken vurderes i afsnit 5.2.1.

3.2.5. Registrerede fejl i opgaver

På grundlag af indberetninger fra lærere mv. til Skolestyrelsen er der efter det oplyste identificeret ni opgaver med fejl, heraf en inden for Dansk/Læsning, seks inden for Matematik og to inden for Fysik/Kemi.

3.3. Udvælgelse af opgaver og estimering af dygtighed

De nationale test vedrørende et fag er beregnet på at kunne afvikles på 45 minutter. I den tid skal eleven igennem de tre profilområder, og på grundlag heraf skal der kunne gives en bedømmelse med indplacering på niveau 1-5 for hvert profilområde, f.eks. med profilen 4-3-5.

Testene skal være adaptive, dvs. tilpasse sig den enkelte elev, således at hver elev skal præsenteres for et udvalg af opgaver fra opgavebanken, der passer i sværhedsgrad til ham eller hende. Forenklet beskrevet skal der udvælges en sværere opgave, hvis eleven har svaret rigtigt, og en lettere opgave, hvis eleven har svaret forkert på den forrige opgave.

Elevens dygtighed skal derfor estimeres ud fra det udvalg af opgaver, som eleven har været præsenteret for, om der er svaret rigtigt eller forkert, og hvor svære de har været. Dette sker ved hjælp af psykometriske metoder.

Metoderne til udvælgelse af opgaver skal endvidere tage højde for, at der skal veksles mellem profilområderne under testen, at opgaverne skal bruges jævnt, og at eleven ikke må præsenteres for opgaver, som han eller hun tidligere har set.

Panelets vurdering af de anvendte metoder fremgår af kapitel 4.

4. Testenes psykometriske egenskaber

4.1. Fundamentet for testene

4.1.1. Introduktion

Det er panelets opfattelse, at det er grundlæggende rigtigt at udforme de nationale it-baserede test som adaptive test baseret på en opgavebank med homogene items godkendt efter en Rasch-model.

Det er naturligt at tage udgangspunkt i de såkaldte Rasch-modeller i forbindelse med udvikling og afprøvning af nye test, fordi modellen kan ses som en formel repræsentation af optimale og ideelle test. Hvis Rasch-modellen holder, kan målingerne ikke blive bedre.

Ifølge Rasch afhænger sandsynligheden for et korrekt svar på en opgave af to og kun to forhold: personens færdighed og opgavens sværhedsgrad. Der findes en solid teori, der godtgør, at anvendelse af en Rasch-model fører til færdighedsmålinger, der både er begrebsvalide, objektive, statistisk sufficiente og – hvad der ikke er uvæsentligt i forbindelse med adaptive test – relativt enkle at have med at gøre.

Objektiviteten – og specielt kravet om at målingen ikke afhænger systematisk af, hvilke items, der benyttes – er af central betydning i forbindelse med adaptive test, hvor forskellige personer skal besvare forskellige opgaver. En situation, hvor visse personer har chancer for at placere sig forholdsvis godt, hvis der bruges et bestemt sæt opgaver, men ringere chancer for at placere sig godt, hvis der bruges andre opgaver, er naturligvis uacceptabel.

4.1.2. Anvendelsen af de psykometriske metoder

Panelet har modtaget en redegørelse for de psykometriske metoder og principper, der har været lagt til grund for afprøvningen af opgaver til de nationale test.

Panelet vurderer, at disse metoder er velvalgte, at de er stærkere end i de fleste store internationale test-projekter og at de - på visse punkter - også er stærkere end de krav, der har været defineret som grundlag for de nationale test.

Kontrollen af Rasch-modellen og estimeringen af opgavernes sværhedsgrader efter afprøvningen er foretaget ved hjælp af metoder, som - i lyset af det meget store datamateriale - vurderes at være tilstrækkeligt gode.

Kontrollen af den anvendte model har været forholdsvis grundig og har forholdt sig mere kritisk til tegn på problemer, end man normalt ville gøre i forbindelse med afprøvning af opgaver, idet opgaver er blevet forkastet ved ethvert tegn på afvigelser fra Rasch-modellens krav uden en vurdering af opgavens indhold.

På enkelte områder kan modelkontrollen dog forbedres ved gennemførelse af yderligere analyser, f.eks.

- Vurderinger af responsmønstre for enkelte elever
- Grafiske kontroller

Estimeringen af dygtigheden er i forbindelse med *afprøvningen* af opgavebanken foretaget ved hjælp af optimale metoder Til anvendelse i forbindelse med *afviklingen* af de adaptive test er kun implementeret den næstbedste metode, der er nemmere at implementere, og som vurderes at være fuldt tilstrækkelig netop i forbindelse med adaptive test.

4.1.3. Opgaver med flere items (polytome opgaver)

Som omtalt er en stor del af opgaverne i opgavebanken polytome, dvs. opbygget med flere items, som hver især besvares med korrekt/ej korrekt svar. Derefter tælles antallet af korrekte items sammen, og hvis der f.eks. er M items, opnår eleven dermed en opgave-score med mulighederne X=0,1,2,3,…. M. Denne rigide måde at *registrere* elevens respons på i denne type opgaver kræver, at der er en vis orden i den måde, hvorpå eleverne kan svare/besvarer de enkelte del-items.

Ideelt set burde de items, der ligger til grund for registreringen af X (= antal korrekte delbesvarelser), selv være underkastet en vurdering af Rasch-homogenitet. Kun i dette tilfælde svarer stigende eller faldende værdier af X til stigende eller faldende værdier af den færdighed, som måles ud fra mængden af items. En lærer vil i denne sammenhæng ud fra høje eller lave værdier af X (jf. tilbagemeldingen af elevresultatet) få nyttig information med henblik på at forbedre den videre undervisning af eleven. Men hvis der er mærkværdigheder i items (med lokale afhængigheder mv.), vil lærerne ikke kunne anvende indholdet af item-svarene.

Såvel konsortiets opgørelser fra afprøvningen af itembanken som panelets selvstændige analyser af udtrækket giver indikationer af, at den simple Rasch-model i mange (sandsynligvis de fleste) af de polytome opgaver ikke holder over for rækken af items.

I nogle tilfælde er der indlysende grunde til dette, idet en successiv besvarelse af del-items f.eks. kan medføre, at man ikke kan lave kun én fejl (har man 4 rigtige har man automatisk også den sidste rigtig – og har dermed 5 rigtige). Der kan være andre begrundelser for sådanne typer af lokal item-afhængighed, som vanskeliggør en dagligdags operationalisering af lærerens tilbagemelding til eleven. Læreren har ikke grundlag for at 'forstå' de mekanismer bag ved de items, der øger/sænker værdien af X, eftersom items er filtret ind i hinanden på en sådan måde, at en simpel sammentælling af antal rigtige svar ikke giver mening.

Anvendt rigtigt kan polytome opgaver imidlertid give et større bidrag til afklaring af elevens dygtighed end et tilsvarende antal items i dikotome opgaver, forudsat at de har en passende sværhedsgrad i forhold til eleven. Anvendes opgaverne på en sådan måde, at de enten er for lette eller for svære for eleven, kan de derimod være mindre effektive end dikotome opgaver i denne henseende. Anvendelsen af dikotome opgaver er derfor stærkt afhængige af en velfungerende adaptiv procedure.

Panelet anbefaler derfor, at der foretages omhyggelige analyser af den måde, de polytome opgaver fungerer på i forbindelse med de adaptive test, ud fra såvel et fagligt som et psykometrisk synspunkt, inden det besluttes i hvilket omfang polytome opgaver skal konstrueres og indgå i opgavebanken.

Panelet anbefaler, at polytome opgaver underkastes særlige analyser af lokal afhængighed, inden de accepteres i opgavebanken.

4.1.4. Vedligeholdelse af opgavebanken

Panelet peger på, at det er vigtigt, at der installeres faste rutiner til sikring af, at opgaverne i opgavebanken vedvarende er homogene. Det gælder ved forhold, hvor banken konstrueres på ét tidspunkt på året, mens 'trækket', dvs. den praktiske afvikling af testene, finder sted på et andet tidspunkt, idet opgavernes relative egenskaber mht. til sværhed kan skifte mellem de to tidspunkter. Men det gælder også under forhold, hvor bankens opgaver løbende skiftes ud.

Hvis man udskifter bankens opgaver med nye opgaver, og alene benytter de 'dynamiske' elevspecifikke data, vil det - hvis man ikke tager aktive skridt til det modsatte - være en stadig vanskeligere opgave, at holde fast i, at opgaverne i item-banken skal være homogene.
Et eksempel på et sådant skridt er fra tid til anden at foretage en indsamling af data ved ikke-adaptive afprøvninger, som det skete i december 2006. I dette tilfælde skal en sådan opdatering tjene to formål: Fastholdelse af homogeniteten og opdatering af normerne.

4.1.5. Præstationsscorer

Når det gælder beregning af præstationsscorer, der kan bruges til at præsentere resultater vedrørende elevernes dygtighed, anvendes flere forskellige repræsentationer på en usystematisk måde, bl.a. således at elevens dygtighed og opgaven sværhed ikke fremstilles på samme måde.

Der skal tages stilling til, om man ønsker præstationsscorer, der kan sammenlignes over tid, eller præstationsscorer, der refererer specifikt til et bestemt prøvetidspunkt. Panelet har således ikke haft mulighed for at vurdere dette yderligere.

4.1.6. Prisen for en-dimensionalitet

Kravet om, at de nationale test skal give en valid og reliabel indplacering af eleverne på en en-dimensional skala (inden for hvert profilområde), indebærer i sig selv den begrænsning, at kun de færdigheder, der lader sig indplacere på en og samme skala, kan indgå i testen.

Dermed udelukkes måling af en række af mulige andre færdigheder hos eleven, og dermed reduceres til en vis grad testenes værdi som diagnostisk test.

I princippet er det ligegyldigt hvilke konkrete opgaver, eleven har besvaret rigtigt eller forkert, for de måler alle det samme (omend med forskellig sværhed). De opgaver, der måtte have målt andre aspekter af faget og elevens færdigheder, er sorteret væk under afprøvningen. Det gælder også opgaver, som meget specifikt kunne følge op på konkrete emner i undervisningen.

I de nationale test har læreren adgang til at se elevens konkrete svar på de opgaver, der har været udtrukket. Heraf kan læreren evt. uddrage mere viden om ele-

vens færdigheder, end han eller hun kan på basis af den profil, der kommer ud af testen.

4.2. Panelets analyser af testene i maj-juni

4.2.1. Datagrundlaget for panelets psykometriske vurdering

Panelet har ikke fundet anledning til at gennemføre en selvstændig gentagelse af analysen af data fra afprøvningen i oktober-december måned 2006. Den allerede gennemførte analyse har ført til godkendelse af opgavebanken og estimering af opgavesværheder mv. Det formodes, at der næppe vil dukke nye inhomogeniteter op ved en fornyet analyse af data fra opbygningen af opgavebanken.

Til brug for egne analyser har panelet rekvireret og modtaget et udtræk af alle test fra 110 tilfældigt udvalgte skoler. Udtrækket omfatter 11.430 (anonymiserede) testforløb fordelt på ca. 4.000 test for hver de tre fag. Udtrækket omfatter i alt ca. 632.000 opgavebesvarelser.

Endvidere har panelet kunnet trække på resultater fra visse analyser udført af konsortiet bag de nationale test.

4.2.2. Opgavesværheder og homogenitet

Rasch-analyse af alle data fra testene i maj-juni 2007 har dels afsløret, at visse af opgaverne fungerede på en anden måde i maj-juni, end de gjorde ved afprøvningen, og dels, at der ser ud til at være problemer med homogeniteten i nogle af opgaverne.

Nedenstående tabel opsummerer resultaterne. Den første søjle af tabellen indeholder oplysninger om korrelationen mellem estimaterne af item-parametrene i december 2006 og i maj-juni 2007. Denne korrelation skal være tæt på 1, fordi parametrene er bestemt med forholdsvis stor sikkerhed. Hvis dette ikke er tilfældet, må det skyldes, at opgavernes indbyrdes sværhedsgrader må have ændret sig i den mellemliggende periode.

Den anden søjle indeholder oplysninger om, hvor mange opgaver, der tilsyneladende fungerer heterogent i maj-juni.

Tabel 4.1Korrelation mellem sværhedsgrader ved afprøvningen og maj-juni, samt antal heterogene opgaver i maj-juni

Profilområde	Korrelation mellem sværhedsgrader i december 2006 og maj-juni 2007	Antal heterogene opgaver
Sprogforståelse	0.74	18
Matematik - geometri	0.81	0
Matematik - anvendt	0.93	21
Fysik/Kemi - energi	0.93	22
Fysik/Kemi - stof	0.91	16
Fysik/Kemi - anvendt	0.92	14

Korrelationen er generelt høj, men trods alt ikke høj nok for sprogforståelse og geometri. Inden for disse områder synes der at have været tale om forskydninger i sværhedsgraderne i perioden mellem afprøvningen og den obligatoriske prøve i maj-juni 2007. Også på de øvrige områder optræder der opgaver, hvor sværhedsgraderne har ændret sig, selvom det er mindre udtalt for disse områder.

Som konsekvens heraf bør item-parametrene og tærskelværdierne i opgavebanken erstattes af værdier, der er estimeret ud fra det samlede materiale fra maj-juni. Det skal sikre, at den fremtidige estimering af dygtigheden ikke skævvrides af parameterværdier, der ikke er gyldige på tidspunktet for de obligatoriske test.

Analyseresultaterne for de opgaver, der synes at fungere heterogent i maj-juni, bør studeres nøjere. Supplerende analyser kan resultere i to forskellige årsager til misfit i forhold til Rasch-modellen:

- For det første vil der være opgaver, der diskriminerer for dårligt, fordi de ikke måler det, som testen drejer sig om.
- For det andet kan der være tale om items, der diskriminerer bedre end Rasch-modellen tillader.

I begge tilfælde bør opgaverne udgå af opgavebanken.

4.2.3. Forskydning af niveauer fra afprøvningen til maj-juni

Med udgangspunkt i niveau-indplaceringer af elever kan man opstille nedenstående oversigt over fordelingen af eleverne på profilværdier (baseret på udtrækket).

Tabel 4.2 Fordeling af elever på trin for profilområder (udtræk af test i maj-juni)

	Antal test	Profilværdi				
		1	2	3	4	5
Sprogforståelse	3.710	11,4%	24,6%	29,9%	24,0%	10,1%
Tekstforståelse	3.710	13,4%	26,6%	28,7%	23,8%	7,4%
Geometri	4.090	12,1%	26,3%	30,9%	21,8%	8,8%
Matematik i anvendelse	4.090	10,7%	24,1%	33,1%	23,4%	8,7%
Energi og energiomsætning	3.630	11,2%	25,8%	29,3%	24,5%	9,2%
Stoffer og materialer	3.630	11,8%	25,5%	29,9%	23,5%	9,3%
Anvendelse og perspektiver	3.630	10,9%	26,5%	29,7%	23,4%	9,5%

Det ses, at der synes at være flere elever på de lavere trin end i den forudsatte fordeling.

Denne forskydning kan næppe skyldes, at eleverne er mindre dygtige end eleverne i december. I stedet må den skyldes, at populationen i december ikke var repræsentativ for elevpopulationen i de landsdækkende test, eller at homogeniteten af opgavebanken har forskudt sig.

Dette forhold indikerer ligeledes, at item-parametrene og tærskelværdierne i opgavebanken bør erstattes af værdier, der er estimeret ud fra det samlede materiale fra maj-juni, for at sikre, at den fremtidige estimering af dygtigheden ikke skævvrides af parameterværdier, der ikke er gyldige på tidspunktet for de obligatoriske test

4.2.4. Profilområder

Det er forventningen, at skalaerne for de forskellige profilområder inden for et fag måler forskellige færdigheder.

Panelet har undersøgt om dette er tilfældet inden for testene i Matematik og Fysik/Kemi.

Analyserne viser, at profilområderne Geometri og Matematik i anvendelse er forskellige dimensioner af faget. Ligeledes udgør Perspektiver og anvendelser en særlig dimension inden for Fysik/Kemi, mens de øvrige profilområder inden for dette fag ikke er to forskellige profildimensioner (Den faglige vurdering er, at begge profilområder, således som opgaverne er udformet, repræsenterer paratviden og ikke en faglig forståelse. Se afsnit 5.3.1). Men analyserne viser også, at variationerne mellem de nævnte profilområder er så beskeden, at den næppe har den store praktiske betydning i psykometrisk henseende..

Det følger som en konsekvens heraf, at det har mening at beregne et samlet præstationsmål for fagene, hvilket er en fordel. Det står imidlertid også klart, at opdelingen på profilområder i realiteten ikke har nogen betydning for indplaceringen af eleven.

Det er muligt, at der kan defineres andre profilområder, der i højere grad beskriver forskellige færdigheder. Det vil forudsætte en didaktisk analyse af faget, men her kan man muligvis hente inspiration i en gennemgang af de afviste opgaver - sammenholdt med de godkendte.

4.2.5. Analyse af opgaver og besvarelser

Panelet har undersøgt, om opgaverne har fungeret som forventet i udtrækket, og om svarmønstrene fra konkrete elever har været sandsynlige i forhold til de forventninger, der stilles via Rasch-modellen.

Resultatet af disse analyser er, at mængden af svarsekvenser, der har en *lille sandsynlighed*, er al for stort; i fagene Dansk/Læsning og Matematik over 25 % og i faget Fysik helt op til 40 % , afhængig af, hvor lange opgavesekvenser, man ser på.

Det giver en kraftig markering af, at noget er galt. Analyserne kan ikke vise, hvad det er, men der kan peges på nogle mulige årsager:

- at der er afhængigheder mellem besvarelserne af visse items, dvs. at eleven 'lærer' noget i løbet af besvarelsen af opgaverne, som kan benyttes, når eleven kommer frem til de sidste opgaver.
- at eleverne ikke har taget opgaverne alvorligt, men præsteret svar, der er styret af andet end opgavesværhed og dygtighed.

Hvis det er tilfældet, at eleverne i et større antal test ikke har taget opgaverne alvorligt, så kan det være en medvirkende forklaring til, at nogle opgaver tilsyneladende fungerer heterogent i maj-juni, som omtalt i afsnit 4.2.2. Det bør under alle omstændigheder undersøges.

4.2.6. Varighed og statistisk sikkerhed i praksis

Den metode til afvikling af de nationale test, der har været anvendt i maj-juni, havde ingen mekanisme til vurdering af sikkerheden på estimatet af elevens dygtighed. I stedet opererede den med et kriterium baseret på antal opgaver, således at

bedømmelsen ansås for sikker efter 20 opgaver inden for profilområdet, hhv. usikker efter 10 opgaver. Ved færre opgave blev der ikke givet en bedømmelse.

Nedenstående tabel viser, at i Dansk/Læsning og Fysik/Kemi er næsten alle forløb i udtrækket afsluttet med en sikker bedømmelse, mens det i Matematik gælder, at 150-200 forløb ikke er afsluttet med en sikker bedømmelse ud fra denne definition.

Det er dog oplyst over for panelet, at konsortiet efterfølgende har vurderet, at sikkerheden for 99,88% af de gennemførte forløb i maj-juni har været inden for de grænser, man fremover vil benytte.

Tabel 4.3 Antal sikre, usikre og ikke bedømte pr. profilområde

Profilområde	Elever i alt	Bedømmelse		
		Sikker	Usikker	Ingen
Sprogforståelse	3710	3694	4	12
Tekstforståelse	3710	3700	2	8
Geometri	4090	3937	29	124
Matematik i anvendelse	4090	3875	127	88
Energi og energiomsætning	3630	3611	1	18
Stoffer og materialer	3630	3615		15
Anvendelse og perspektiver	3630	3610	2	18

Nedenstående tabel viser, hvor mange af udtrækkets test-forløb, der er blevet afsluttet inden for de forudsatte 45 minutter med en sikker bedømmelse. Det ses, at en meget stor del af testforløbene har varet mere end 45 minutter. Specielt for testene i Matematik, hvor 80 % har varet længere end 45 minutter, og gennemsnittet for alle test på dette felt er 64 minutter.

Tabel 4.4 Varighed af test pr. fag

Test	Inden for 45 minutter	Antal elever	Andel	Gnsn. varighed (min)
Dansk	Ja	2.350	64%	39
Dansk	Nej	1.333	36%	53
Dansk i alt		3.683	100%	44
Fysik/kemi	Ja	2.031	57%	40
Fysik/kemi	Nej	1.533	43%	56
Fysik/kemi i alt		3.564	100%	47
Matematik	Ja	689	20%	41
Matematik	Nej	2.793	80%	69
Matematik i alt		3.482	100%	64
Der er set bort fra ca. 500 forløb, der har været afbrudt og ca. 160 uden sikker bedømmelse				

Testenes varighed kan hænge sammen med, at der ikke var indbygget en stopmekanisme ved testene i maj-juni, ligesom der er mange forløb, hvor eleverne har løst langt flere end 20 opgaver pr. profilområde.

Panelet må konstatere, at det ser ud til, at eleverne ikke har kunnet gennemføre testene inden for de forudsatte 45 minutter. Det bør udløse alvorlig bekymring og bør analyseres nærmere.

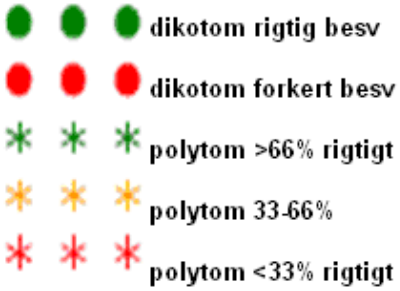
4.2.7. Bemærkelsesværdige forløb

Panelet har på forskellig vis konstateret, at der har været problemer med den praktiske anvendelse af den adaptive rutine, herunder ikke mindst med de meget lange testtider.

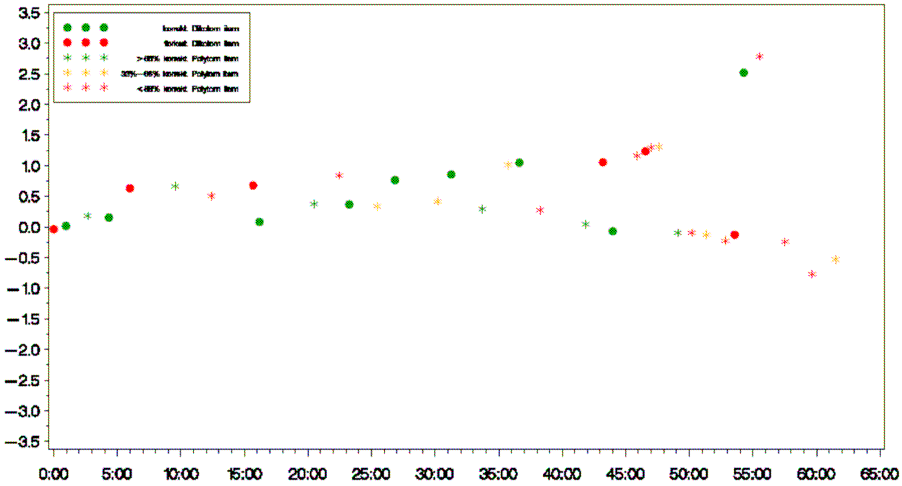
Illustrationerne nedenfor er blot eksempler. De gengiver skridt for skridt sværhedsgraden af de opgaver, som eleven er blevet præsenteret for (Y aksen), mens tiden målt i minutter er gengivet på X-aksen. Punkterne i grafen repræsenterer flere 'lag' af information mht. til typen af opgave og mht. rigtighed af besvarelsen, som det fremgår af signaturen.

Panelet formoder, at de konstaterede problemer har med den begrænsede opgavebank og de meget lange testforløb at gøre. Problemet bør undersøges nærmere.

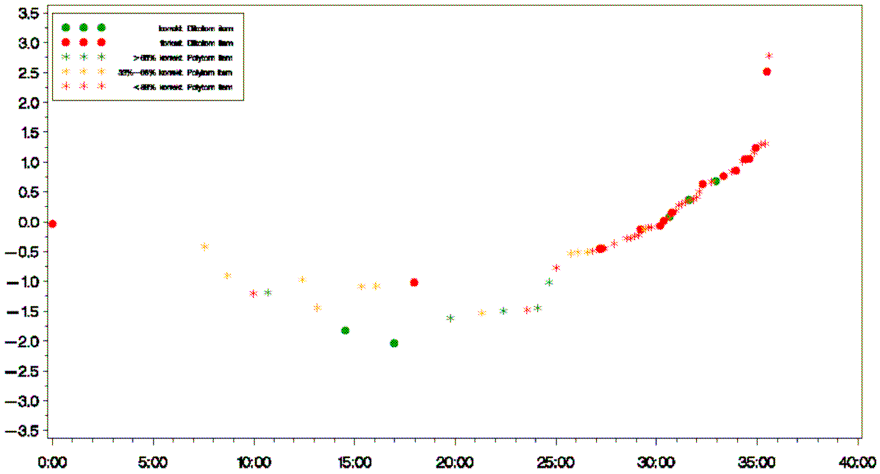
Figur 4.1 Signaturer anvendt i forløbsdiagrammer



Figur 4.2 Uforståeligt forløb med flere divergerende processer på én gang (her med for lang testtid)



Figur 4.3 Uforståelige skift mellem lettere og sværere opgaver, manglende konvergens – inklusiv en mystisk afslutning



I forlængelse heraf har panelet fundet, at der er mange elever, der har løst 100 eller flere opgaver. Det er panelets opfattelse, at der i vidt omfang er tale om forsøg på at tømme opgavebanken, mere end om seriøse testforløb. Der bør indføres mekanismer, der stopper sådanne initiativer.

4.2.8. Overensstemmelsen mellem behovet og opgaverne i opgavebanken

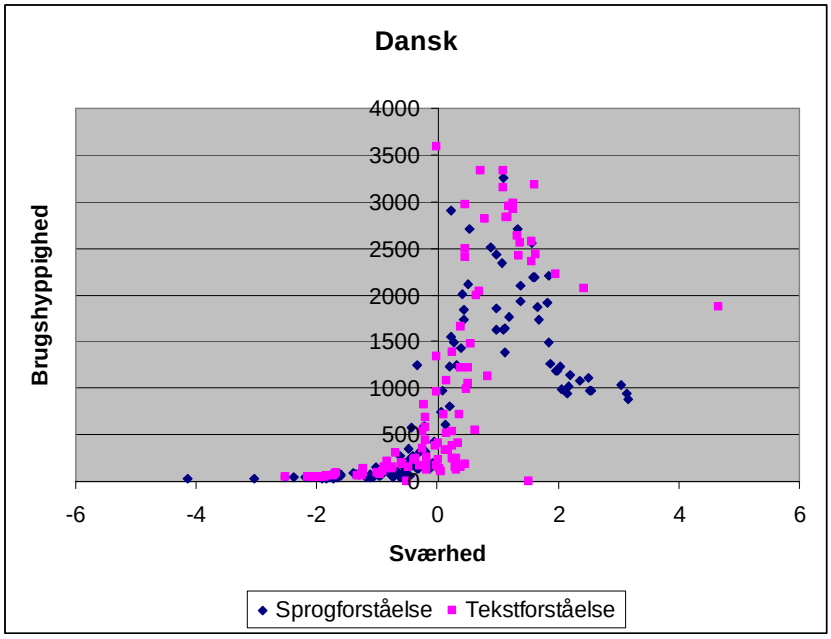
Man kan sammenligne opgavebankens udbud af opgaver med elevernes efterspørgsel af opgaver med forskellige sværhedsgrader. I nedenstående figurer vises forholdet mellem sværhed og det antal gange, en opgave har været besvaret (brugshyppighed) i udtrækket.

Da minimumsforventningen til opgavebanken er, at der skal være lige så mange opgaver til de 10% svageste elever som til de 10% stærkeste elever og til de 30% elever i midten, ville man forvente en klokkeformet brugshyppighedskurve. De 30% forventes at trække tre gange så meget på middelsvære opgaver, som de 10% stærkeste trækker på de sværeste opgaver.

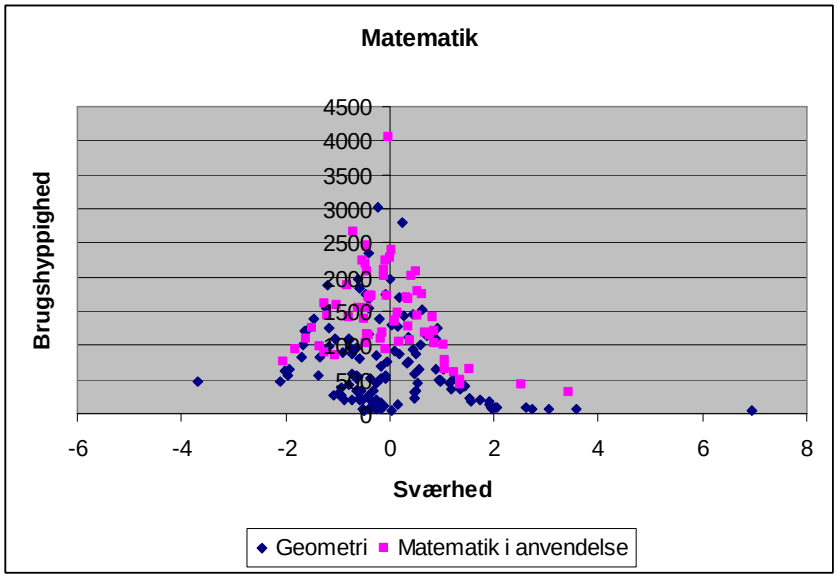
Derudover ville man forvente, at alle opgaver ville blive brugt lige hyppigt - givet deres placering i sværhedsfeltet.

Kigger man på figuren for Dansk/Læsning nedenfor, ser man da også omtrent det forventelige mønster. Figuren tyder dog på, at der er for få svære opgaver til at tilfredsstille efterspørgslen.

Figur 4.4 Brugshyppighed i forhold til sværhed – Dansk/Læsning

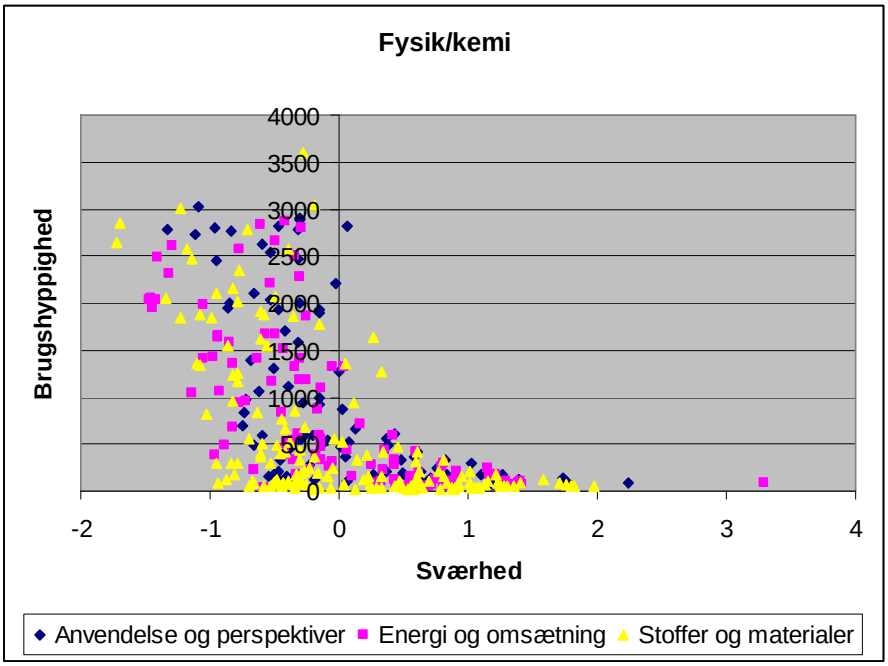


Figur 4.5Brugshyppighed i forhold til sværhed - Matematik



På figuren for Matematik ovenfor genkendes den forventede klokkeform, men der synes at være uforståeligt mange opgaver midt i feltet med relativ lav brugshyppighed, og måske for få lette og for få svære opgaver .

Figur 4.6Brugshyppighed i forhold til sværhed - Fysik/Kemi



Figuren for Fysik/Kemi ovenfor viser et ekstremt mønster. Der er tilsyneladende en stor mængde opgaver inden for Stoffer og materialer af alle sværheder, der bliver brugt i relativt ringe grad, mens et mindre udvalg af opgaverne bliver brugt mere, men også i en uforståelig varieret grad. En delårsag er formentlig, at der er for få lette fysikopgaver i forhold til elevernes dygtighed, hvorfor de relativt lette opgaver efterspørges af alt for mange elever, selv af elever for hvem, de er for svære (fordi de letteste hurtigt bliver brugt op).

Det er fælles for de mønstre, der fremgår ovenfor, at der i alle fag er opgaver, der har været brugt over 3.500 gange, hvilket svarer til, at de er indgået i næsten alle test i udtrækket. Endvidere er der i alle fagene opgaver med bemærkelsesværdigt lille brugshyppighed, selv i det midterste sværhedsområde. I Dansk/Læsning er der to opgaver, der overhovedet ikke er blevet brugt. Dette tyder på, at der kan være problemer med, at den adaptive procedure, der har været anvendt i maj-juni, ikke i tilstrækkelig grad har spredt anvendelsen af opgaverne.

Panelet har ikke haft mulighed for at fastslå forklaringerne på disse mønstre, men finder, at det bør undersøges nærmere.

Panelet har også set på forholdet mellem dikotome og polytome opgaver.

For Fysik/Kemi udgør forekomsten af dikotome items i itembanken en forholdsmæssig større andel end den, der reflekteres af efterspørgslen via den adaptive rutine. Mens items af denne type er repræsenteret med ca. 40 % i opgavebanken, udgør de kun ca. 20 % i de faktiske adaptive forløb.

Noget kunne tyde på, at der er en relation mellem typen af item og sværheden af items, idet man kan observere gennemsnitlige sværhedsgrader for dikotome items på -0.38, mens det tilsvarende tal for polytome opgaver/items er -0.81. De gennemsnitlige sværhedsgrader er således højere for dikotome opgaver/items, sammenlignet med de polytome, og antyder dermed, at dikotome opgaver i gennemsnit er fremstillet til at være 'sværere' end de polytome.

4.2.9. Opgaveskabeloner

Panelet har undersøgt, om anvendelsen af forskellige testopgavetyper (skabeloner) har indflydelse på elevernes sandsynlighed for at besvare en opgave korrekt.

Der er fundet en signifikant skabelon-effekt (kontrolleret for de øvrige variable).

Testopgavetype 41, Markering af tekst, øver *mindst* indflydelse, dvs. sænker chancerne for rigtigt svar mindst muligt, mens testopgavetype 38, Kemireaktionsligning, er det format, der – alt andet lige – i det største omfang sænker chancerne for et korrekt svar. Det sædvanlige multiple choice-format, testopgavetype 1, Multiple choice i skemaform, øver næstmindst (negativ) indflydelse på chancerne for at svare korrekt.

Disse observationer bør give anledning til overvejelser om brugen af skabeloner og om virkningerne heraf.

4.2.10. Kan de nationale test vise forskelle på skoler?

Panelet har undersøgt, om de nationale test kan måle en forskel mellem skoler.

I forhold til præstationerne inden for profilen 'Anvendelser og perspektiver' i Fysik/Kemi er der ikke stor forskel på fordelingerne fra skole til skole. Niveauerne er meget ens og giver ikke anledning til en tydelig rangordning. Dette er usædvanligt i betragtning af, at man normalt ser effekten af skole slå tydeligt igennem, i PISA og IEA's undersøgelser for eksempel. Det betyder, at de muligheder, man har for at rangordne skoler ud fra gennemsnitskarakterer, ikke kan overføres til

tilsvarende manøvrer på 'karakterer' givet på profilområdet 'Anvendelser og perspektiver'.

For flere af de andre profiler gælder tilsvarende konklusioner, men for profilen Matematik i anvendelse ses en større variation fra skole til skole. I disse tilfælde finder der således en vis mulighed for at diskriminere på skoleniveau.

4.3. Vurdering af de nationale test i internationalt perspektiv

I forbindelse med etableringen af de nationale test er der blevet lagt vægt på, at *"Testene skal være ... baseret på forskning og de nyeste nationale og internationale erfaringer."*

Samtidig blev der som baggrund for etableringen af testene henvist til, at *"... de danske resultater i en række internationale sammenligninger af det faglige niveau i folkeskolen, jf. den seneste PISA-undersøgelse, absolut ikke er tilfredsstillende i de helt centrale fag og færdigheder som dansk, matematik og naturfag."*

Det er panelets opfattelse, at de citerede afsnit ikke alene skal tages som udtryk for, at testene skal realiseres på et højt forskningsmæssigt løsningsniveau. Referencen til f.eks. PISA skal også forstås som en kvalitativ markering af, at implementering og den praktiske elevevaluering ud fra de nationale test skal foregå efter principper og standarder, der er sammenlignelige med de internationale komparative undersøgelser som PISA, IEA (TIMSS PIRLS mv.) samt NAPE – testene i USA.

Panelet vurderer, at der i implementeringen og den praktiske elevevaluering i de danske test, som påpeget i dette kapitel og kapitel 5, har været en række konkrete mangler, der betyder, at de i den undersøgte periode i praksis ikke når samme niveau som f.eks.. PISA.

Det skal dog erindres, at sigtet med PISA er et andet, at PISA kun omfatter et lille antal (ca. 4.500) 15-årige, og at PISA f.eks. ikke giver tilbagemelding til lærerne.

For så vidt angår det forskningsmæssige grundlag på den psykometriske dimension, vurderes de nationale test at være på fuldt på højde med eller bedre end international praksis.

Den danske tradition for Rasch-analyser, der går helt tilbage til begyndelsen af 60'erne, er mere kompromisløs og krævende end internationale traditioner. Internationale praktikere er mere tilbøjelige til at have opfattelser som, at "modelkontrollen er ikke så afgørende, det går nok alligevel", og at "det ser jo ud til at fungere i praksis". Sådanne synspunkter kunne bruges som argument for alligevel at bruge nogle af de opgaver, der blev afvist i forbindelse med afprøvningen.

Rasch-analyserne i de danske test har været mere grundige, og man har været mindre parat til at se gennem fingrene med fejl, end man er andre steder i verden.

De internationale erfaringer med adaptive test kan illustreres med dette citat: "in many applications of IRT to predetermined domains of items, it has been found that the Rasch-model does not provide a good fit to the data"[5]. Citatet stammer fra en af hovedpersonerne inden for teorien for adaptive test, og refererer til de erfaringer man har gjort på The Educational Testing Service (ETS) i USA, der er den førende organisation, når det drejer sig om pædagogiske test. Dette synspunkt har haft den konsekvens, at teorien for adaptive test udelukkende drejer sig om to modeller, der ligner Rasch-modellen, men som omtales som henholdsvis 2- og 3-parameter modeller.

Dette udsagn synes dog ikke at kunne overføres på de danske nationale test. For det første kan det jo konstateres, at det rent faktisk i de nationale test er lykkedes at identificere et meget stort antal opgaver, der rent faktisk ser ud til at passe til Rasch-modellen. ETS-erfaringerne er altså ikke generelle, og i hvert fald ikke relevante for de nationale test i Danmark.

For det andet indeholder 2- og 3-parameter modellerne nogle egenskaber, der er åbenlyst meningsløse og ødelæggende for adaptive test. I flg. disse modeller vil rækkefølgen af opgaverne mht. sværhedsgraden blandt de dygtigste elever være den stik modsatte af sværhedsgraden af de samme opgaver blandt de svageste elever. Et sådant forhold er selvsagt absurd og ødelæggende, når der tales om test, der skal fungere og bruges i begge ender af dygtighedsskalaen.

Teorien for adaptive test er helt central for udviklingen af de nationale test, og kan bruges også over for Rasch-modellerne. Men de erfaringer med brugen af teorien, som litteraturen om adaptive test omtaler, kan ikke bruges.

---

[5] Howard Wainer (red): Computerized Adaptive Testing: A Primer. Second Edition. Lawrence Erlbaum Associates, 2000, p. 65.

En markant forskel mellem de danske test og internationale test udgøres af de polytome opgaver. På dette punkt er der i den nuværende opgavebank nogle problemer med lokal afhængighed mellem items, som giver vanskeligt fortolkelige resultater. Det skønnes, at de nationale test kan løse dette problem ved opdatering af opgavebanken, hvorunder en række af opgaverne udskiftes med nye.

5. Testenes faglige egenskaber

5.1. Indledning

Som del af grundlaget for reviewet har panelet haft adgang til at se samtlige 760 opgaver i opgavebanken. Panelet har ligeledes haft adgang til et dataudtræk på 632.000 besvarelser i anonymiserede testforløb, samt adgang til at se de enkelte besvarelser i et forløb (igen anonymiseret).

Vurderingen af opgaverne har desuden haft det følgende udgangspunkt:

Udviklingen af de it-baserede test er sket på baggrund af et veldefineret aftalegrundlag mellem kunde og leverandør. Aftalegrundlaget definerer udgangspunktet for, hvordan de it-baserede test skal udvikles, herunder bl.a. for den faglige del:

- testene skal udarbejdes med udgangspunkt i de **trinmål**, der gælder for faget på det pågældende trin. For faget dansk dog kun trinmål for læsning.
- de trinmål eller dele heraf, der egner sig til it-baserede test, udvælges således at **faget dækkes så bredt** som muligt
- hver test skal inddeles i **faglige profilområder**, der er veldefinerede i forhold til, hvilke trinmål eller dele heraf der indgår i hvert profilområde.
- profilområderne skal i videst muligt omfang **afspejle forskellige aspekter** af faget,
- profilområderne skal være sammensat på en måde, således at **tilbagemeldingen til lærerne** bliver så anvendelig som mulig ifm. planlægning af undervisningen.

Udgangspunktet for udviklingen af testene er således relativt veldefineret. Udviklingen skal tage udgangspunkt i to centrale elementer: trinmål og profilområder. Som tidligere beskrevet er der for hvert af de tre fag udvalgt tre profilområder.

Fag	Profilområde 1	Profilområde 2	Profilområde 3
Fysik/Kemi	Energi og energiomsætning	Stoffer og Materialer	Anvendelse og Perspektiver
Matematik	Tal og algebra	Geometri	Matematik i anvendelse
Dansk/Læsning	Tekstforståelse	Sprogforståelse	Afkodning

Det faglige review af opgaverne konkluderer, at de valgte profilområder fagligt set dækker det enkelte fag bredt og afspejler forskellige aspekter af faget. Profilområderne opfylder det ønskede hovedformål med opdelingen, nemlig at give læ-

reren en tilbagemelding på en måde, som læreren kan relatere sin undervisning hertil.

Inden testene blev afviklet i maj-juni 2007, havde de enkelte opgaver gennemgået en omfattende afprøvning, hvorved de blev underkastet en prøve på, om de i deres udformning var psykometrisk (statistisk) holdbare. Afprøvningen afslørede de opgaver, der af den ene eller anden årsag havde problemer – psykometrisk set – og medførte, at ca. halvdelen af opgaverne ikke blev godkendt. En konsekvens heraf er, at de to profilområder *Afkodning* og *Tal og algebra* ikke har indgået i testene, da der ikke var tilstrækkelig med opgaver til at dække områderne.

Denne form for afprøvning afslører dog ikke, om opgaverne fagligt set er korrekte eller gode. Det kan kun fageksperter i faget og på klassetrinnet afgøre.

For alle opgaver gælder, at de dels har været underkastet en psykometrisk afprøvning, dels er udformet af opgavekonstruktører, der må formodes at besidde de rette kompetencer (uddannelse, fag og trin). Desuden har de været underkastet en faglig kvalitetssikring, der igen må formodes at være tilstrækkelig. Med det review, panelet har gennemført, har opgaverne hermed også været underkastet en faglig 'afprøvning'.

Udviklingen af de enkelte opgaver, f.eks. hvordan en opgave skal udformes, hvilke elementer den skal opbygges af osv, er defineret ud fra nogle rammer:

- opgaver skal være udformet med udgangspunkt i de udvalgte **trinmål** og de **faglige profilområder**.
- opgaver skal indeholde **items, der er veldefinerede** såvel i forhold til trinmål, som til de kundskaber og færdigheder, der konkret testes.
- fagets **kundskaber og færdigheder** skal i videst muligt omfang inddrages i testen.
- til hvert item skal der knyttes en **identifikation** bestående af en række elementer (f.eks. opgavenummer, klassetrin, profilområde, trinmål, kendeord).

Identifikationen er interessant, fordi den er udgangspunktet for, hvordan den adaptive mekanisme kan sortere og udvælge i de enkelte opgaver/items i opgavebanken. Desuden skal f.eks. en udvælgelse på baggrund af kendeord og rapporteringen heraf løbende være med til at sikre, at alle udvalgte trinmål er dækket rigeligt ind med opgaver.

For lærerne er der desuden værdifuld information at hente i identifikationsoplysningerne.

Panelet har gennemført et fagligt review baseret på et repræsentativt udsnit af opgavebanken - for herved entydigt at kunne konkludere, hvorvidt opgaverne fagligt set er:

- konstrueret på en måde, at de fremstår som valide, korrekte og genkendelige i deres form - under forudsætning af de givne rammer (profilområder, trinmål, items).
- opbygget på en hensigtsmæssig måde mht. brugen af illustrationer, overskrifter, instruktioner, formuleringer osv.

Opgavebankens måde at udvælge opgaver på til den enkelte elev er ud fra et fagligt synspunkt også interessant. Den er nemlig afgørende for, om den enkelte elev i praksis bliver tildelt en opgave, der svarer til vedkommendes niveau - og om det er den rigtige opgave, eleven præsenteres for. Panelet har vurderet den mekanisme i opgavebanken, der udvælger opgaverne. Vurderingen af algoritmen har primært haft et psykometrisk udgangspunkt, men også en række faglige forhold er blevet vurderet.:

- om den enkelte elevs test sammensættes af opgaver med udgangspunkt i et bredt udvalg af de udvalgte trinmål og dele heraf.
- om testen omfatter en fordeling af items vedrørende hvert profilområde, der giver basis for en velunderbygget tilbagemelding på profilområder.

I panelets gennemgang af elevforløb er der tegn på, at den adaptive algoritme ikke i alle tilfælde har udvalgt opgaver på en optimal måde. Dette kan skyldes forskellige forhold, der bør undersøges nærmere. Dog er der i de psykometriske analyser ovenfor redegjort for forhold omkring den adaptive algoritme, der kan være årsagsforklarende også i den faglige tilgang.

Panelets vurdering af de tre test fremgår af de følgende afsnit.

5.2. Generelle observationer

Panelet har observeret en række forhold, der gør sig gældende på tværs af fagene, og som kan være en følge af måden, opgaverne konstrueres på - eller måden opgaveinstruktørerne instrueres på.

5.2.1. Opgavevarianter

Panelet har fundet, at et proportionalt set større antal opgaver består af varianter over den samme opgave. Med varianter menes opgaver med samme problemstilling (billede, tegning, diagram) eller med samme tekst (enslydende opgaver), der anvendes i flere opgaver, eller opgaver, der overlapper helt - bortset fra distraktorerne (opgaveformuleringen, spørgsmålet og det rigtige svar er det samme).

Panelet kan samtidig konstatere, at den adaptive algoritme ikke tager højde for, om en elev tidligere har set en variant af samme opgave. Det er et problem, idet en elev, der tidligere er blevet præsenteret for en tekst, i forbindelse med en ny opgave naturligvis vil være bedre stillet end de elever, der ikke har mødt teksten før.

Den forventede størrelse af opgavebanken taget i betragtning, må det formodes, at opgavebanken i et vist omfang kan indeholde varianter over en opgave. Det er dog helt centralt, at den adaptive algoritme tager højde herfor. Samtidig er det afgørende, at antallet af varianter ikke udgør en så stor andel af opgavebanken, at det bliver afgørende for, hvorvidt der er opgaver nok til at fylde en time og samtidig ikke forøger risikoen for Teaching-to-the-test. Se næste afsnit.

Panelet anbefaler, at brugen af opgavevarianter i opgavebanken reduceres eller i det mindste, at udvælgelsesproceduren indrettes således, at en elev aldrig møder to eller flere varianter af samme grundopgave i et testforløb.

5.2.2. Teaching-to-the-test

Givet opgavebankens størrelse, og da en del af opgaverne ser ud til at blive brugt meget hyppigt, vurderer panelet, at der pt. er en betydelig risiko for teaching-to-the-test.

Panelet vurderer, at de lærere, der er til stede under testenes afvikling, forholdsvis let vil kunne huske en række af problemstillingerne og følgende år gøre ekstra meget ud af disse i deres undervisning. Et forhold, der forstærker denne antagelse er det forholdsvis høje antal af varianter.

Panelet anbefaler, at opgavebanken udbygges, således at risikoen for teaching-to-the-test mindskes. Antallet af opgaver skal være rigeligt, således at efterspørgslen af opgaver ikke må medføre en hyppig udnyttelse af enkelt opgaver.

5.2.3. Brug af flere items pr. opgave

Panelet finder anvendelsen af flere items i opgaverne problematisk. Ud fra et fagligt synspunkt betyder en opgave med mange items, at det varer tilsvarende længe, inden eleven kommer videre til et andet emne i faget.

Dertil kommer, at der for nogle opgavers vedkommende er flere items, der ikke har samme sværhedsgrad. Dette er specielt problematisk for de elever, der har svært ved at løse opgaverne. (Og for estimeringen af elevens dygtighed, ff. afsnit 4.1.3)

Endvidere anvendes der hyppigt opgavetyper med samme antal spørgsmål og svarmuligheder. Kan eleven eksempelvis besvare fire ud af fem items, vil det, at der resterer én svarmulighed, gøre, at eleven får et gratis svar. I internationale sammenhænge konstruerer man typisk disse 'multiple matching'- opgavetyper på en sådan måde, at eleven ved sit sidste svar har to eller tre valgmuligheder.

Nedenstående tabel viser fordelingen af opgaver i opgavebanken efter antallet af items i opgaven.

Tabel 5.1 Antal opgaver og antal items pr. opgave

	Opgaver	Antal items pr opgave									
Profilområde	i alt	1	2	3	4	5	6	7	8	9	10
Sprogforståelse	118	87	5	6	10	3	2				5
Tekstforståelse	100	59	6	7	9	13	3	1	1		1
Geometri	127	63	16	9	11	16	10	1		1	
Matematik i anvendelse	63	20	1	21	8	4	4	3	1	1	
Energi og energiomsætning	112	53	7	7	8	20	11	3	3		
Stoffer og materialer	150	66	4	8	30	28	10	4			
Anvendelse og perspektiver	90	36	3	5	20	24	1	1			
Opgaver i alt	760	384	42	63	96	108	41	13	5	2	6

5.2.4. Antal opgaver og spredning i opgavesværhed

Panelet kan konkludere, at der mangler mange opgaver i opgavebanken. Dette skyldes ikke alene det forhold, at der mangler to profilområder og dermed også alle opgaverne herfra. Samtlige de profilområder, der har indgået i maj/juni 2007, bør udvides med mht. antal opgaver per trinmål og opgaver per profilområde. Som det fremgår af nedenstående tabeller, er det langt fra alle trinmål, der kan siges at være repræsenteret tilfredsstillede i opgavebanken.

Tabel 5.2 Oversigt over opgavernes fordeling på trinmål inden for Dansk/Læsning

Profilområde	Trinmål	Antal opgaver
Sprogforståelse	01:Sprogets poetiske funktion	16
Sprogforståelse	02:Gøre rede for betydningen af sproglige virkemidler og bruge dem	2
Sprogforståelse	03:Gøre rede for samspillet mellem sprog, tekst, genre, indhold og situation	16
Sprogforståelse	04:Forholde sig analytisk og vurderende til sprog, sprogbrug og sprogrigtighed	75
Sprogforståelse	05:Kende forskellige sætningstyper og sætningsled samt ordklasserne og deres funktion i sproget	7
Sprogforståelse	06:Bruge regler for sammensætnnger og afledninger	2
Tekstforståelse	01:Fortolke, vurdere og perspektivere tekster og andre udtryksformer ud fra såvel umiddelbar oplevelse som analytisk forståelse	8
Tekstforståelse	02:Kende forskellige genrer og deres blandingsformer inden for fiktion og ikke-fiktion	8
Tekstforståelse	03:Gøre rede for genre, komposition, fortælleforhold, fremstillingsform, tema og motiv, sprog og stil samt meningen i tekster	52
Tekstforståelse	04:Læse norske og svenske tekster	9
Tekstforståelse	05:Anvende viden om litteraturens foranderlighed gennem tiderne og om, at litteraturen afspejler den tid, den er blevet til i	5
Tekstforståelse	06:Benytte varierende læsemåder afpasset efter formålet - oversigtslæse, punktlæse og nærlæse	12
Tekstforståelse	07:Fastholde det væsentlige i en tekst	6
I alt		218

Tabel 5.3 Oversigt over opgavernes fordeling på trinmål inden for Matematik

Profilområde	Trinmål	Antal opgaver
Geometri	01:Benytte geometriske metoder og begreber i beskrivelse af fysiske objekter fra dagligdagen, herunder figurer og mønstre	15
Geometri	02:Undersøge og beskrive enkle figurer tegnet i planen	19
Geometri	03:Kende til grundlæggende geometriske begreber som vinkler og parallelitet	14
Geometri	04:Arbejde med fysiske modeller og enkle tegninger af disse	31
Geometri	05:Kende til forskellige kulturers metoder til at angive dybde i billeder	8
Geometri	06:Undersøge de enkelte tegnemetoders anvendelighed til beskrivelse af form og afstand	13
Geometri	07:Måle og beregne omkreds, areal og rumfang i konkrete situationer	4
Geometri	08:Tegne, undersøge og eksperimentere med geometriske figurer, bl.a. ved at benytte computer	23
Matematik i anvendelse	01:Vælge og benytte regningsarter i forskellige sammenhænge	16
Matematik i anvendelse	02:Anvende og forstå enkle informationer, som indeholder matematikfaglige udtryk	7
Matematik i anvendelse	03:Anvende faglige redskaber, herunder tal, grafisk afbildning og statistik, til løsningen af matematiske problemstillinger fra dagligliv, familieliv og det nære samfundsliv	14
Matematik i anvendelse	04:Arbejde med enkle procentberegninger, herunder ved rabatkøb	15
Matematik i anvendelse	05:Beskrive og tolke data og informationer i tabeller og diagrammer	11
I alt		190

Tabel 5.4 Oversigt over opgavernes fordeling på trinmål inden for Fysik/Kemi (del 1/2)

Profilområde	Trinmål	Antal opgaver
Energi og energiomsætning	01:Anvende enkle fysiske og kemiske begreber til at beskrive hverdagens fænomener som regnbuen, elektricitet i hjemmet og korrosion	41
Energi og energiomsætning	07:Beskrive og forklare energioverførsel som fotosyntese, ånding og elektrisk energioverførsel	27
Energi og energiomsætning	13:Kende til fordel og ulemper ved udnyttelsen af forskellige energiformer, herunder vedvarende energikilder	21
Energi og energiomsætning	14:Give eksempler på, at der ved fremstilling af energi ofte produceres stoffer og varme, der påvirker miljøet	8
Energi og energiomsætning	15:Beskrive og forklare energioverførsel ved udvalgte eksempler fra teknikken, som transport og brændselsceller	10
Energi og energiomsætning	18:Kende eksempler på produktionsprocesser og delprocesser, heraf som gæring og katalyse	5
Stoffer og materialer	01:Anvende enkle fysiske og kemiske begreber til at beskrive hverdagens fænomener som regnbuen, elektricitet i hjemmet og korrosion	10
Stoffer og materialer	02:Kende til enkle modeller, herunder forestillingen om, at alt stof er opbygget af partikler	7
Stoffer og materialer	03:Beskrive nogle grundstoffer og kemiske forbind-elser samt enkle træk i det periodiske system	15
Stoffer og materialer	04:Kende nogle generelle egenskaber ved hverdagens stoffer og materialer, som tilstandsformer, ledningsevne og surhedsgrad	20
Stoffer og materialer	06:Kende jordens og månens bevægelser og nogle af de virkninger, der kan iagttages på jorden som årstider, tidevand og formørkelser	8
Stoffer og materialer	07:Beskrive og forklare energioverførsel som fotosyntese, ånding og elektrisk energioverførsel	9
Stoffer og materialer	08:Kende udvalgte stoffers kredsløb i naturen som kulstof, nitrogen og vand	8
Stoffer og materialer	09:Kende til nogle forestillinger om stofopbygning, herunder det periodiske system	12
Stoffer og materialer	11:Kende nutidens forestilling om solsystemets opbygning	9
Stoffer og materialer	13:Kende til fordel og ulemper ved udnyttelsen af forskellige energiformer, herunder vedvarende energikilder	11
Stoffer og materialer	14:Give eksempler på, at der ved fremstilling af energi ofte produceres stoffer og varme, der påvirker miljøet	14
Stoffer og materialer	15:Beskrive og forklare energioverførsel ved udvalgte eksempler fra teknikken, som transport og brændselsceller	5
Stoffer og materialer	16:Beskrive udvalgte produkters og materialers vej fra fremstilling til bortskaffelse	9
Stoffer og materialer	18:Kende eksempler på produktionsprocesser og delprocesser, heraf som gæring og katalyse	13

Tabel 5.5 Oversigt over opgavernes fordeling på trinmål inden for Fysik/Kemi (del 2/2)

Profilområde	Trinmål	Antal opgaver
Anvendelse og perspektiver	01:Anvende enkle fysiske og kemiske begreber til at beskrive hverdagens fænomener som regnbuen, elektricitet i hjemmet og korrosion	26
Anvendelse og perspektiver	08:Kende udvalgte stoffers kredsløb i naturen som kulstof, nitrogen og vand	10
Anvendelse og perspektiver	10:Kende nogle tidligere kulturers forestilling om universets opbygning	7
Anvendelse og perspektiver	12:Beskriver forhold, hvor udviklingen af teknologi er tæt forbundet med fysisk og kemisk viden	7
Anvendelse og perspektiver	13:Kende til fordel og ulemper ved udnyttelsen af forskellige energiformer, herunder vedvarende energikilder	12
Anvendelse og perspektiver	14:Give eksempler på, at der ved fremstilling af energi ofte produceres stoffer og varme, der påvirker miljøet	3
Anvendelse og perspektiver	18:Kende eksempler på produktionsprocesser og delprocesser, heraf som gæring og katalyse	20
Anvendelse og perspektiver	19:kende til eksempler på ekektronisk styring i hverdagen	2
Anvendelse og perspektiver	20:Anvende it-baserede redskaber til dataopsamling og præsentation som temperaturregistrering og adgangskontrol	2
Anvendelse og perspektiver	21:Formulere spørgsmål og indsamle relevante data	1
I alt		352

Desuden kan panelet konkludere, at der mangler opgaver inden for det helt lette niveau (1) og det helt svære niveau (5). Dette er uddybet nærmere i afsnit 4.2.8.

Panelet vurderer, at det er tvivlsomt, om alle trinmål er repræsenteret dækkende. Desuden vurderes de enkelte trinmål at være repræsenteret ved et antal opgaver, der varierer meget. Også de enkelte læringsmål under profilområderne vurderes at være ujævnt repræsenteret.

I panelets arbejde har det været påpeget, at der i udviklingsprocessen har været arbejdet med en detaljeret fordeling af opgaver og item per trinmål, I nedenstående tabel 5.5 (eksempel fra Matematik) ses antallet af opgaver udarbejdet til hvert profilområde inden afprøvningen. Antallet er aftalt ud fra en forlods vurdering af områdets vægt i matematikundervisningen. Antallet og fordelingen af godkendte opgaver afviger, som det fremgik ovenfor af Tabel 5.3, en del fra det planlagte.

*Tabel 5.5 Oversigt over opgavernes **planlagte** fordeling på trinmål inden for Matematik-Geometri*

Geometri	Antal grund-opgaver	Antal Opgaver i alt	Antal items i alt
I alt	57	221	567
G1 Benytte geometriske metoder og begreber i beskrivelse kter fra dagligdagen, herunder figurer og mønstre	5	19	63
G2 Undersøge og beskrive enkle figurer tegnet i planen	9	33	77
G3 Kende til grundlæggende geometriske begreber som llelitet	9	36	108
G4 Arbejde med fysiske modeller og enkle tegninger af disse	10	37	62
G5 Kende til forskellige kulturers metoder til at angive dybde	3	12	26
G6 Undersøge de enkelte tegnemetoders anvendelighed til form og afstand	6	24	69
G7 Måle og beregne omkreds, areal og rumfang i konkrete	7	28	44
G8 Tegne, undersøge og eksperimentere med geometriske ed at benytte computer	8	32	118

Panelet anbefaler, at opgavebanken udbygges, således at antallet af opgaver per trinmål og per profilområde forøges og således, at alle niveauer af opgavesværhed dækkes lige godt ind i forhold til efterspørgslen af opgaver.

5.2.5. Elevforløb

Panelet har vurderet en række elevforløb. De følger det ønskede forløb med et udgangspunkt i opgaver med middel sværhedsgrad og med valg af opgaver skiftevis fra de forskellige profilområder. Rigtighedsprocenter har også de værdier, som de bør have - ca. 50 %. Panelet har dog fundet adskillige forløb, der adskiller sig fra det ønskelige: Dels er der forløb, der er meget langvarige, og dels er der forløb, hvor elever gætter (klikker sig igennem), og derfor tømmer hele opgavebanken.

Det er ikke fundet beskrevet, hvordan opgaver udvælges, således at eleven kan undgå at møde tidligere besvarede opgaver i større omfang.

Panelet har vurderet elevforløb ud fra en psykometrisk vurdering med en række betragtninger om den adaptive algoritme. Dette følger af afsnit 4.2.6 til 4.2.8.

5.2.6. Distraktorer

Panelet har vurderet brugen af distraktorer (forkerte svarmuligheder) i opgaverne.. Panelet finder brugen af distraktorer problematisk på tre områder:

- Der er eksempler på opgaver, hvor flere distraktorer ikke er entydigt forkerte, og det betyder, at der er flere rigtige svarmuligheder i forhold til et item – modsat hensigten.
- Ligeledes er der eksempler på, at distraktorer genbruges i forskellige opgaver. Det er et problem, såfremt eleven præsenteres for opgaver, hvor nogle af distraktorerne går igen. Sandsynligheden for, at eleven kan slutte sig til det rigtige svar ved at udelukke distraktorer øges dermed.
- Endelig er der eksempler på opgaver, der gør brug af distraktorer, der ikke altid forekommer plausible.

Panelet anbefaler, at opgavebanken gennemgås med henblik på at få gjort brugen af distraktorer stringent og konsekvent.

5.2.7. Overskrifter

Brugen af overskrifter synes ikke konsekvent på tværs af fagene. Såfremt opgaverne skal gøre brug af overskrifter, bør overskriftens funktion overvejes, og den bør i så fald anvendes konsekvent, så eleven ikke oplever et skift.

Panelet anbefaler at definere brugen af overskrifter mere præcist og at implementere brugen konsekvent på tværs af opgaver.

5.2.8. Illustrationer, sprog og formuleringer

I panelets gennemgang af opgaver er der fundet opgaver, der indeholder sproglige fejl, uklare formuleringer – eller som bærer præg af manglende konsekvens i forhold til tegnsætning. Desuden er der eksempler på, at illustrationer ikke er præcise nok i forhold til den måde de skal aflæses/anvendes på.

Panelet anbefaler, at opgaverne generelt underkastes korrekturlæsning, og at upræcise illustrationer erstattes.

5.2.9. Scrollning

Panelet vurderer, at der ganske ofte er brug for at scrolle, når eleven skal danne sig et overblik over opgaven. Det kan være problematisk for både yngre og ældre elever. Dels tager det tid at orientere sig og få overblik over opgaven, og den tid går fra elevens opgaveløsning. Dels kan det virke forvirrende for de yngre elever.

Panelet anbefaler, at opgaverne konstrueres så scrollning helt kan undgås.

5.2.10. Metadata

Panelet har i sin gennemgang af de metadata, der dels skal være knyttet til den enkelte opgave - dels til det enkelte item, fundet, at der mangler metadata på opgaveniveau, og at der helt mangler metadata på itemniveau.

Særligt er brugen af kendeord fraværende, og der, hvor kendeord er anvendt, er det ofte gjort usystematisk og inkonsekvent. Opgavemål er også et eksempel på metadata, der er udfyldt usystematisk og inkonsekvent.

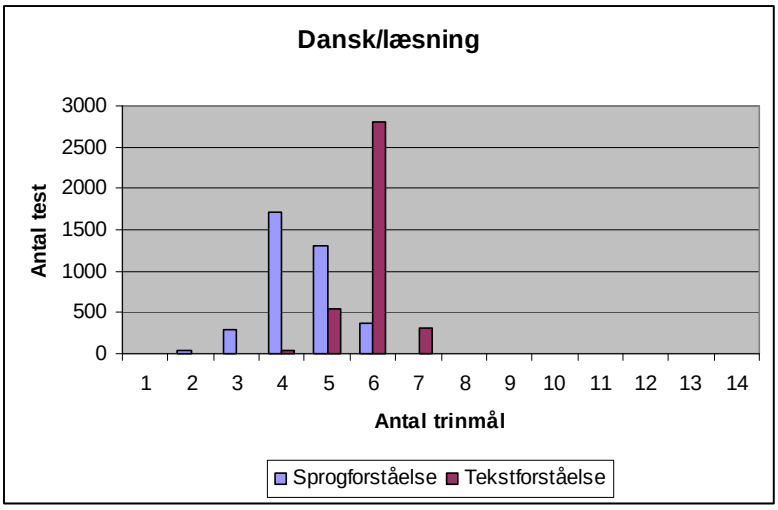
Panelet anbefaler, at samtlige metadata bør gennemgås, og de steder, hvor metadata er anvendt usystematisk og inkonsekvent (f.eks. kendeord), bør opbygningen begynde helt forfra.

5.2.11. Trinmål i de enkelte testforløb

For at kunne vurdere eleven ud fra alle de af fagets aspekter, der er tænkt ind i de nationale test, skal eleven præsenteres for et bredt udsnit af opgaver fra samtlige valgte trinmål.

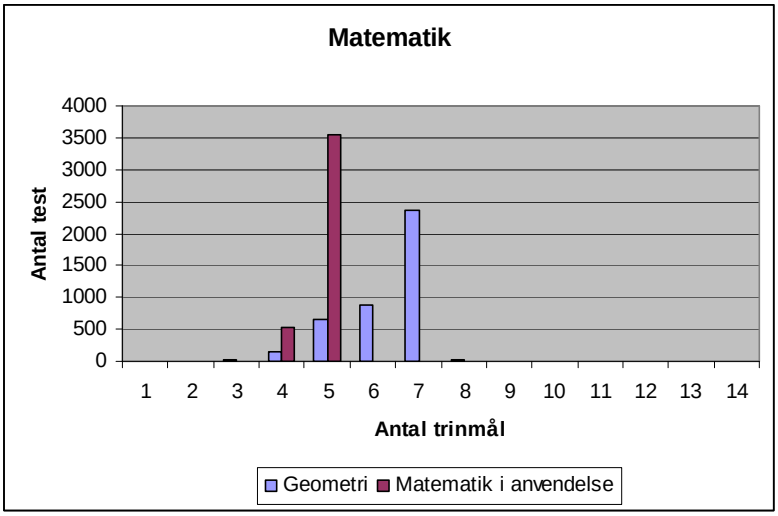
Opgørelsen er baseret på de faktiske ca. 4.000 forløb pr. fag i udtrækket. En stor andel af disse test har varet mere end 45 minutter, og har formentlig omfattet langt flere opgaver end tiltænkt.

Figur 5.1 Fordeling af test på antallet af trinmål der indgår - Dansk/Læsning



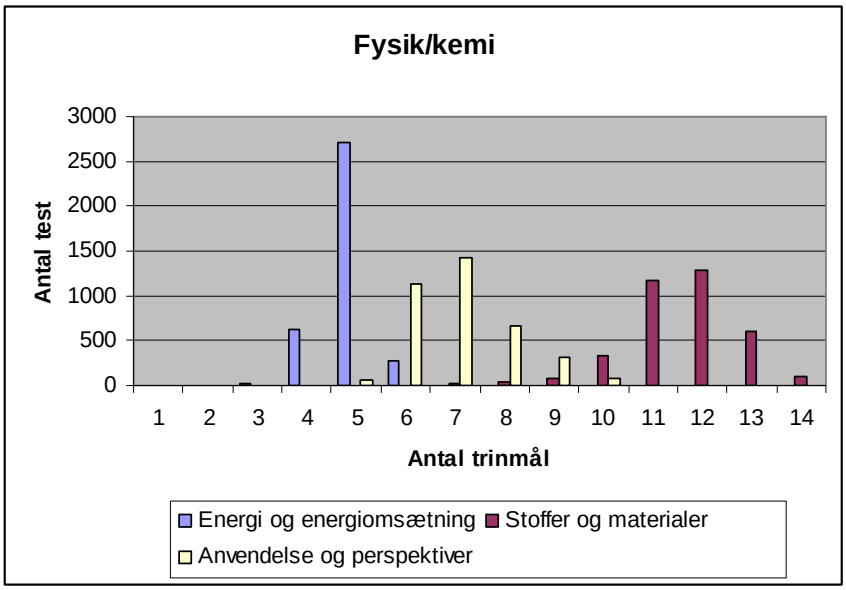
Det antal trinmål, der indgår i opgavebanken for Dansk/Læsning er hhv. seks for Sprogforståelse og syv for Tekstforståelse. I forhold til Sprogforståelse indgår højst fire ud af seks trinmål i halvdelen af testene. For Tekstforståelse indgår mindst seks ud af syv i langt de fleste test.

Figur 5.2 Fordeling af test på antallet af trinmål der indgår - Matematik



Det antal trinmål, der indgår i opgavebanken, er hhv. otte for Geometri og fem for Matematik i anvendelse. I Matematik i anvendelse indgår stort set alle trinmål i langt de fleste test. For Geometri gælder det, at mere end halvdelen af testene omfatter syv trinmål.

Figur 5.3 Fordeling af test på antallet af trinmål der indgår - Fysik/Kemi



Det antal trinmål, der indgår i opgavebanken for Fysik/Kemi, er hhv. seks for Energi og energiomsætning, 14 for Stoffer og materialer, samt 10 for Anvendelse og perspektiver. Det ses, at for Energi og energiomsætning indgår mindst fem ud af seks trinmål i langt de fleste test. Mht. Anvendelse og perspektiver er det således, at mere end halvdelen af testene omfatter syv eller færre af de 10 trinmål. For Stoffer og materialer indgår 12 eller flere af de 14 trinmål i mere end halvdelen af testene.

5.3. Vurdering af Fysik/Kemi

Panelets overordnede konklusion på det faglige review indenfor faget Fysik/Kemi er, at maj/juni testen på en rimelig måde har afdækket elevernes færdigheder.

Fagligt set fungerer de valgte profilområder godt, og næsten alle trinmål dækkes på en rimelig måde. Der er mange gode opgaver, men samtidig også mange opgaver, der bør rettes til. Og antallet af opgaver bør udvides markant. Panelet finder, at testen bør suppleres med opgaver, der tester forståelse og anvendelse af viden på baggrund af generelle faglige principper. Der bør også i højere grad bruges opgaver, som giver en faglig baggrund, som eleven kan argumentere fagligt ud fra.

Konklusionen på det faglige review er således, at opgavebanken kan udvides kvantitativt, videreudvikles kvalitativt og derved kan komme til at fremstå som et af flere anvendelige evalueringsværktøjer i fremtiden.

Der er dog fundet nogle problemer i testen/opgaverne. Disse gennemgås i det følgende.

5.3.1. Sammenfaldende profilområder

En af de første problemstillinger, som panelet ønsker at fremhæve, er, at to af fagets profilområder[6] overvejende tester den samme færdighed, nemlig paratviden. Dette understøttes af den psykometriske validering af profilområderne, jf. afsnit 4.2.4.

Problemet med profilområderne kan føres tilbage til den måde, faget traditionelt har været inddelt på. Profilområderne er kendetegnende fagområder og er den opdeling, som lærerne vil kunne genkende og finde anvendelig – også i relation til tilbagemeldingen og den fremadrettede planlægning. Når opgaverne inden for profilområderne tilsyneladende tester de samme typer af færdigheder, skyldes det hovedsageligt, at der reelt er tale om to fag (Fysik og Kemi), der skal testes med én test, men også, at opgaverne ikke er konstrueret på en måde, der gør, at andre færdigheder hos eleven testes. Opgavekonstruktionen har med andre ord taget udgangspunkt i den samme måde at teste på - inden for de to sammenfaldende profilområder.

Panelet anbefaler, at fagets tre profilområder bibeholdes, da de fagligt vurderet er de rette områder at anvende. Panelet anbefaler dog samtidig, at man ændrer måden at konstruere opgaver på, således at en langt større andel af opgaverne tester forståelse og anvendelse af viden på baggrund af generelle faglige principper. Herved bør de to sammenfaldende profilområder også blive forskellige. Ideelt set bør der ikke være opgaver, der tester paratviden – men gerne opgaver, der tester paratforståelse.

5.3.2. Opgavevariation

I forlængelse af ovenstående problematik omkring test af paratviden, finder panelet, at opgavevariationen ikke er helt som forventet. Ud over, at for mange opgaver tester elevernes paratviden - det sker ofte inden for de trinmål, der er formuleret åbent - er der også forholdsvis for mange opgaver inden for energiomsætninger. Modsat mangler der helt opgaver, der tester elevens viden om naturvidenskabelige arbejdsprocesser og evnen til at argumentere fagligt ud fra nogle givne oplysninger.

[6] Profilområde 7. Energi og energiomsætning samt profilområde 8. Stoffer og materialer

I Fysik/Kemi er der mange ens varianter af opgaver. Dette problem er omtalt under de tværgående problemstillinger, men nævnes igen her, da varianterne i Fysik/Kemi udgør en stor andel af den samlede opgavemængde. Hvis antallet af de ovennævnte opgavetyper reduceres, vil antallet af varianter være markant.

Panelet anbefaler, at der konstrueres langt flere opgaver inden for Fysik/Kemi. Dette vil nedsætte risikoen for teach-to-the-test og reducere problemet med varianter. Konstruktionen bør tage udgangspunkt i et ønske om at opnå en større variation af faglige problemstillinger. Der bør defineres et koncept for, hvordan man kan teste elevens viden om naturvidenskabelige arbejdsprocesser og evnen til at argumentere fagligt på basis af givne oplysninger.

5.3.3. Illustrationer og sprog

I det forgående afsnit 5.2 om tværgående problemstillinger er problematikken omkring upræcise illustrationer nævnt, men den optræder også hér. Dog kan der i forhold til Fysik/Kemi tilføjes, at der anvendes for mange og for komplekse illustrationer f.eks. af elektriske kredsløb, hvor tegningerne skal aflæses for flere sæt af parametre. Opgaver med denne type illustrationer vil kunne gøres mere enkle og overskuelige ved brug af mere simple illustrationer. Komplekse illustrationer må ikke i sig selv blive afgørende for, om eleven kan løse en problemstilling, der kan illustreres enkelt på anden vis.

Tillige har panelet fundet, at der er en uklar og i hvert fald varieret praksis mht. brug af fagsprog/hverdagssprog. Denne fremgangsmåde er formentlig anvendt for at gøre opgaverne mere tilgængelige, men det skaber reelt en fare for det modsatte: At eleven bliver forvirret over det upræcise ordvalg og opfatter opgaven som svært tilgængelig.

Panelet anbefaler, at undervisningens indhold perspektiveres mere bredt gennem brug af varierede illustrationer, samt at antallet af billedbaserede svaritems og af parametre, der skal aflæses på den enkelte illustration, begrænses.
Desuden anbefaler panelet, at opgaverne skrives i et præcist (niveau-relevant) fagsprog - ikke i hverdagssprog.

5.3.4. Distraktorer

Også mht. distraktorer er der i forhold til Fysik/Kemi nogle særlige konstruktionsmæssige udfordringer ved de reviewede opgaver. Panelet har fundet, at en del af de svarmuligheder, der gives i opgaveformuleringen, og som er tænkt som forkerte svar, rent faktisk er mulige svar på opgaven. De mulige svar er blot ikke noget, eleven er blevet undervist i. En vidende fagmand, eller en meget vidende/dygtig elev på klassetrinnet, besidder denne viden, hvorfor fagmanden eller den meget vidende/meget dygtige elev vil have en større sandsynlighed for at få fejl i disse opgaver.

Panelet anbefaler, at opgaver med denne type distraktorer enten ændres eller fjernes helt.

5.3.5. Brug af begreber/terminologi

Den danske folkeskole anvender forskellige lærebøger i faget Fysik/Kemi. Det giver en udfordring, når der fokuseres på begrebsterminologi, idet nogle lærebøger sværger til den græsk/latinske terminologi, mens andre lærebøger sværger til den danske terminologi. Det vil derfor være således, at eleverne på et klassetrin må formodes enten at have et indgående kendskab til begreber i den ene eller den anden terminologi. Netop derfor ser panelet det som problematisk, at terminologien ikke anvendes konsekvent i udformningen af opgaver. Enkelte opgaver går sågar direkte på at omsætte fagbegreber mellem græsk/latinske ord og tilsvarende danske. Sådanne opgaver bør ikke forekomme. Ligeledes har panelet fundet, at navngivningen af kemiske forbindelser er inkonsekvent.

Panelet anbefaler på denne baggrund, at testen tilpasses med en konsistent brug af terminologi, således at begge sæt konsekvent anvendes med det ene (og altid det samme) i parentes. Panelet anbefaler ligeledes en mere konsekvent navngivning af kemiske forbindelser.

5.4. Vurdering af Dansk/Læsning

Testen i Dansk/Læsning adskiller sig fagligt fra to øvrige fag, fordi testene skal teste færdigheder bredt i faget. I Dansk er fra Undervisningsministeriet side på forhånd truffet et valg om at fokusere snævert på en særlig del af faget, nemlig læsning. Mere herom nedenfor.

De tre profilområder i Dansk/Læsning er: *Tekstforståelse*, *Afkodning* og *Sprogforståelse*. Panelet har, på trods af at området ikke indgik i maj/juni testen, forholdt sig til profilområdet *Afkodning*, for således at kunne vurdere faget som en samlet helhed.

Danskfaget er i meget høj grad et kulturbærende dannelsesfag. De it-baserede test har fra projektets start haft fokus på at udvikle opgaver i en lille del af faget, nemlig læsning, som der normalt ikke er selvstændig fokus på. På den baggrund er det forståeligt, at overordentlig mange af de stillede opgaver kan karakteriseres som opgaver, der måler elevens indsigt i dansk sprog, litteratur og kultur snarere end som opgaver i læsning.

Det må dog fastholdes, at opgaverne skal teste læsning – dvs. teste færdigheder i at forstå tekster for bl.a. at *tilegne* sig viden. Når fagområdet hedder Dansk/Læsning betyder det, at det er den del af læsning, der indgår i danskfaget.

Det betyder, at mange af de udformede opgaver ikke kan betragtes som læseopgaver, da de tester *kendskab til dansk kultur*, og ikke *tilegnelse* af oplevelse og viden fra tekst.

Det er særdeles veldokumenteret, at effektiv tekstforståelse bygger på to centrale komponenter: ordafkodning og sprogforståelse. Disse to faktorer forklarer hver især en stor selvstændig del af variationen i tekstforståelse både blandt begynderlæsere og ældre elever. Tilsammen giver de tre definerede profilområder et fyldestgørende billede af den enkelte elevs færdigheder inden for læsning. Mål for elevens hastighed bør rapporteres særskilt for de enkelte profilområder, da de tre profilområder afdækker helt forskellige færdigheder.

Panelets gennemgang af opgavebanken konkluderer, at hovedparten af opgaverne i Dansk/Læsning ikke er brugbare i deres nuværende udformning. Teksterne fra nogle opgaver vil kunne anvendes som udgangspunkt for udarbejdelse af nye opgaver.

Panelet finder også, at det ikke er veldefineret, hvilke færdigheder i læseforståelse testen skal afdække, og anbefaler derfor at skærpe kendeordene i retning af komponenter i *læsning*, snarere end i andre danskfaglige retninger.

I de nye opgaver bør der også arbejdes med en synkronisering af formuleringerne, således at samme opgave stilles på samme måde.

Panelet foreslår på ovenstående baggrund en forenkling af arbejdet med opgavekonstruktionen. Der bør tages udgangspunkt i en begrænsning af opgavetyperne, sådan at sammenhængen mellem opgaver og testmål kan blive klar. Antallet af testede trinmål og emner må herefter afhænge af, hvad der kan lade sig gøre at udforme gode opgaver til i praksis.

Panelet anbefaler for hver af de tre profilområder, at testene gør brug af to opgavetyper indenfor hvert område. Anbefalingen uddybes i det følgende.

Panelet anbefaler desuden, med udgangspunkt i, at de foreslåede opgavetyper til profilområderne anvendes på samtlige klassetrin, at de it-baserede test i læsning fungerer som en lang test med fælles opgavebank. Dette vil også til dels løse nogle af problemerne med manglende lette/svære opgaver på de enkelte klassetrin.

5.4.1. Afkodning

Testen i *Afkodning* skal afdække elevernes færdighed i at læse ord af varierende sværhedsgrad. *Afkodning* er årsagsrelateret til ringe læseforståelse og er derfor et vigtigt undervisningselement i læseundervisningen på alle klassetrin, der er omfattet af de it-baserede test.

Det vil være hensigtsmæssigt at udforme en lang afkodningstest, der kan bruges fra 2. til 8. klasse. Ordene vil spænde fra de letteste, først tilegnede - som *is* og *mor* - til ord, der vil kunne udfordre de allerdygtigste i 8. klasse f.eks. chihuahua og rokokokokotte.

Panelet anbefaler, at der i profilområdet udelukkende tages udgangspunkt i to opgaveformater, ordidentifikation af billedvalg (vælg/find det ord der passer til billedet) og ordkædeopgaver. Disse opgaveformater er begge velafprøvede og har vist høj validitet og reliabilitet, hvor de er afprøvet.

5.4.2. Sprogforståelse

Profilområdet *Sprogforståelse* drejer sig om den sprogforståelse, der er central i tekstlæsning. Formålet med dette profilområde er at hjælpe lærerne til at vurdere, om eventuelle læsevanskeligheder skyldes begrænsninger i sprogforståelsen (f.eks. som følge af, at dansk er et andetsprog). Udfordringen ved at stille opgaver i sprogforståelse er således, at kravene til læsefærdigheden skal være minimale.

Ellers er der jo tale om læseopgaver. Den optimale måde at teste sprogforståelsen på er at lade eleven høre ord og tekster og svare ved f.eks. at vælge et billede. Såfremt dette ikke er en mulighed, anbefaler panelet brug af to opgaveformater: Opgaver i ords betydning og opgaver i udtryk/faste vendingers betydning.

Snarere end at teste kendskab til eksisterende ordsprog (dansk kulturarv) kunne det være opgaver i at finde betydningen af *nye* udtryk og vendinger med ord i overførte betydninger.

5.4.3. Tekstforståelse

Inden for profilområdet *Tekstforståelse* er det vigtigt, at opgaverne ikke kan løses, uden at eleven har læst teksten. Det er ligeledes vigtigt, at opgaverne er prøver i at få oplysninger og indsigt ud af tekster. Endvidere er det centralt, at der bliver udformet spørgsmål, der afdækker forskellige niveauer af elevernes tekstforståelse, og der bør være eksempler på spørgsmål, hvor eleverne skal kombinere flere oplysninger fra teksten for at kunne svare korrekt. Opgaver i *Tekstforståelse* skal derimod ikke være prøver i forhåndskendskab til danskkulturel dannelse.

Til profilområdet er det at stille flere spørgsmål i samme opgave problematisk, idet spørgsmålene typisk er af temmelig forskellig sværhedsgrad. Derved udsættes dårlige læsere for alt for vanskelige del-spørgsmål. En opgaverevision skal sikre, at sådanne opgaver kræver tekstforståelse, ikke stavefærdighed, kulturel indsigt, eller blot at kunne finde det ord, der ikke giver syntaksbrud.

Til profilområdet *Tekstforståelse* anbefaler panelet to opgavetyper: opgaver med tekst og et valg mellem flere svarmuligheder og cloze-test med valg mellem flere alternativer.

5.5. Vurdering af Matematik

5.5.1. Profilområder

Profilområderne har en meget tydelig tilknytning til hovedområder og trinmål i faghæftet Fælles Mål for Matematik. Profilområderne er valgt, fordi de er velkendte for matematiklærerne og udgøres af områderne *Tal og Algebra*, *Geometri* og *Matematik i anvendelse*.

Opdelingen i profilområder er således fagligt velfungerende, og de udvalgte trinmål repræsenterer faget rimelig bredt. Opgaverne vurderes for størstedelens ved-

kommende at være fagligt i orden, dog kan mange af dem rent sprogligt forenkles en del.

Panelet anbefaler at fastholde opdelingen i de tre profilområder.

5.5.2. Validitet

Matematiktesten skønnes begrænset valid på grund af tre forhold. Det ene forhold er, at der mangler opgaver i banken helt ned på trinmålsniveau. Det andet forhold er, at profilområdet *Tal og algebra* mangler. Det tredje er, at der i forhold til visse opgaver i review'et er skabt usikkerhed om, hvad der egentlig testes i opgaverne, bl.a. pga. forskellige opgaveskabeloner, sproglige instruktioner, tekstmængde mm.

Panelet anbefaler, at de opgaver, hvor der er usikkerhed om, hvad der egentlig testes, udsættes for en nærmere analyse. Dertil anbefaler panelet, at opgavebanken udvides kraftigt mht. antal opgaver per trinmål.

5.5.3. Tilbagemelding på test

Tilbagemeldingen på testen - sådan som lærerne får den - er blevet vurderet af panelet. Konklusionen er, at den i sin nuværende form ikke er helt så anvendelig som ønsket. Der mangler dels et profilområde og dermed ca. 1/3 af faget. Desuden mangler der opgaver inden for nogle af trinmålene.
Som tilbagemeldingen fremstår pt., får læreren en for overordnet tilbagemelding til, at den kan bruges til den fremadrettede planlægning og løbende evaluering. Hvis dét er meldingen til læreren, skal vedkommende selv ud på opgaveniveau for at blive inspireret.

Panelet anbefaler, at tilbagemeldingen udsættes for et fagligt review, og at man på den baggrund beslutter om tilbagemeldingen skal udsættes for designændringer.

5.5.4. Opgaveskabeloner

Gennemgangen af Matematik har rettet søgelyset mod en række eksempler, der kunne tyde på, at de tilgængelige opgaveskabeloner har været styrende for en opgaves indhold og dens antal af item. Dette skønnes ikke at være hensigtsmæssigt set ud fra et fagligt synspunkt.

Det anbefales, at den videre udvikling af faget Matematik i høj grad bør præcisere hvilke opgaveskabeloner, der er velegnede til udvikling mod de enkelte trinmål.

6. Anbefalinger

6.1. Forskningstilknytning

6.1.1. Tættere samarbejde med forskere

Panelet anbefaler, at der ved udviklingen af opgaver i højere grad inddrages forskere og eksperter, både i de enkelte fag og i psykometriske spørgsmål. I panelets gennemgang af opgaver har det vist sig, at central viden om, hvordan elementer i fagene kan testes, ikke er nået hele vejen igennem opgaveudviklingen. Dette kommer efter panelets opfattelse bl.a. til udtryk i de manglende profilområder, hvor de mange kasserede opgaver tyder på et gennemgående problem, og i Dansk/Læsning, hvor langt størstedelen af opgaverne skal ændres.

Forskere og eksperter bør inddrages allerede fra forberedelsen af opgavekonstruktionen og løbende inddrages i godkendelsen af opgaverne.

6.1.2. Åben dokumentation af de nationale test

Dokumentationen for de psykometriske egenskaber ved testene bør lægges åbent frem, således at eventuelle kritikere kan kontrollere, om testene fungerer som beskrevet. Den åbne dokumentation er en nødvendig forudsætning for, at testene kan fremstå som forskningsbaserede.

6.2. Psykometrisk opbygning

6.2.1. Fastlægge itembank-værdier på rigtig tid og population

Det har vist sig, at opgaver, der var godkendt under afprøvningen, ikke havde samme sværhed (eller var inhomogene) under testene i maj-juni.

Panelet anbefaler derfor, at sværhedsgrader for opgaverne i opgavebanken estimeres ud fra det samlede materiale fra maj-juni for at sikre, at den fremtidige estimering af dygtigheden ikke skævvrides af parameterværdier, der ikke er gyldige på tidspunktet for de obligatoriske test eller er estimeret på en ikke-repræsentativ population.

Fremover bør det tilstræbes, at afprøvningen af opgaver foregår i den samme del af skoleåret som de test, som opgaverne skal indgå i.

Dette kan muligvis også medføre, at opgaver, der ikke var homogene under afprøvningen, viser sig at være det i den rigtige testperiode (se anbefaling om fornyet afprøvning af afviste opgaver.)

6.2.2. Fastlæggelse af præstationsscores

Der anvendes flere forskellige repræsentationer af elevdygtigheder og opgavesværheder i testene på en uigennemskuelig og inkonsistent måde. Dette bør ændres.

Projektet må bl.a. tage stilling til, om der ønskes præstationsscorer, der kan sammenlignes over tid, eller præstationsscorer, der refererer specifikt til et bestemt prøvetidspunkt.

Det er imidlertid helt afgørende, at der anvendes samme transformation for elevdygtigheder som for opgavesværheder.

6.2.3. Sikkerhed på bedømmelser

I den adaptive algoritme, der har været anvendt i maj-juni, fortsætter testen indtil enten eleven eller læreren afbryder den, eller til der ikke er flere opgaver. Den nye algoritme, som er implementeret i oktober, opererer med en ”sikker” bedømmelse af testen, når den statistiske usikkerhed for bedømmelsen af elevens resultat er mindre end eller lig en bestemt grænseværdi.

Den i øjeblikket valgte værdi indebærer tilsyneladende en stor risiko for fejlklassifikationer på dygtighedsskalaen med fem trin. I værste fald kan det være tilfældigt, om en elev indplaceres på trin to eller trin fire på dygtigheds-skalaen (1-2-3-4-5-skalaen).

Det anbefales, at der foretages analyser af, hvor stor risiko der er for at placere eleverne forkert med forskellige grænseværdier, før man tager stilling til, hvilken grænseværdi, der skal accepteres.

Dette kan naturligvis få indflydelse på, hvor mange opgaver/items, en elev skal besvare, inden estimatet af dygtigheden er tilstrækkeligt sikkert, og dermed på testenes varighed og/eller antallet og udformningen af opgaverne.

6.3. Opgavebank

6.3.1. Yderligere kontrol af opgavebanken

Den afprøvning af opgavebanken, der blev gennemført forud for maj-juni, bør suppleres med yderligere kontroller, som også bør gennemføres i forbindelse med den fortsatte udvikling og afprøvning af opgaver. Nedenfor er nævnt eksempler på sådanne mere vidtgående kontroller. Der er dog også andre relevante kontroller, der bør gennemføres.

Responsmønstre for elever bør vurderes, så det bliver klart, hvor stor en andel af eleverne, der i maj/juni testen havde responsmønstre, der afveg fra det forventede efter Rasch-modellen. Dvs. f.eks. med responsmønstre med uforholdsmæssigt mange fejl på lette opgaver sammenlignet med antallet af rigtige svar på de sværeste opgaver.

Den **psykometriske effekt** af forskellige opgaveskabeloner for samme type opgave bør undersøges. Det er klart fra panelets statistiske undersøgelse, at opgaveskabelonen i sig selv øver indflydelse på elevens chance for at svare korrekt på den stillede opgave.

6.3.2. Fornyet afprøvning af afviste opgaver

Der bør gennemføres en indholdsanalyse af de afviste opgaver, hvor det undersøges, om der blandt de afviste opgaver er nogle, der er afvist på grund af statistiske fejl, eller om de er afvist, fordi de måler noget andet end det, de er konstrueret til.

En omhyggelig indholdsanalyse af de afviste opgaver vil formodentlig kunne føre til, at man kan skelne mellem opgaver med konstruktionsfejl, opgaver, der måler noget forkert, og opgaver, hvor der ikke kan påvises nogen problemer. Fejlene i den første type opgaver skal naturligvis rettes, og de korrigerede opgaver afprøves på ny. Den sidste gruppe opgaver bør inkluderes uændrede i en ny afprøvningsrunde for at se, om de stadig afvises. Hvis det ikke er tilfældet, må det konkluderes, at der enten var tale om en statistisk type I fejl i den første afprøvning, eller at problemet skyldtes, at opgaven var fejlplaceret i den første del af skoleåret. Disse opgaver inkluderes i item-banken. Hvis opgaven derimod stadig afvises, må konklusionen være, at opgaven elimineres endeligt.

6.3.3. Inhomogene opgaver kasseres

Opgaver, der på baggrund af panelets analyser har vist sig inhomogene ved deres anvendelse i maj-juni testen, bør udgå af opgavebanken, indtil de evt. er godkendt ved en fornyet afprøvning.

6.3.4. Uddrage information af de afviste opgaver

Den omhyggelige indholdsanalyse af de afviste opgaver kan muligvis tillige føre til identifikation af opgaver, der netop er afvist, fordi de måler færdigheder, som adskiller sig fra de færdigheder, som de godkendte opgaver måler. Heri kan ligge inspiration til at definere profilområderne, så de er mindre indbyrdes afhængige, end det har vist sig at være tilfældet i den foreliggende testrunde.

6.4. Forløb

6.4.1. Bedre fordeling af brug af opgaver

Panelet har påvist brugsmønstre for opgaver, hvor nogle opgaver anvendes uforholdsmæsigt meget, mens andre anvendes uforholdsmæssigt lidt.

Det bør undersøges, hvad årsagen til disse anvendelsesmønstre er. På baggrund heraf bør der ændres i opgavebankens sammensætning og/eller i metoderne til udvælgelse af opgaver.

6.4.2. Sikre afvikling af test på 45 minutter

Erfaringerne fra maj-juni viser, at meget få forløb er afsluttet inden for 45 minutter. Panelet har ikke haft mulighed for at vurdere, om dette udelukkende skyldes en manglende stopklods i afviklingen af test, eller om det skyldes problemer med at få tilstrækkeligt sikre bedømmelser.

Panelet anbefaler, at problemet analyseres yderligere, således at de fornødne ændringer i relation til en tidsbegrænset afvikling kan blive indført. Panelet peger på, at det af hensyn til lærerens udbytte af testene også bør sikres, at en rimelig stor del af de udvalgte trinmål og emner inden for profilområdet indgår i hver elevs test.

6.4.3. Tilbagemelding om usandsynlige elevforløb

Det anbefales, at man fremover identificerer elever med påfaldende responsmønstre og indbygger det i tilbagemeldingen til læreren, således at læreren får en mulighed for at vurdere mulige årsager til det påfaldende mønster og tage højde herfor i sin anvendelse af testresultaterne og i forbindelse med kommende test

6.5. Opgaver

6.5.1. Revurder brugen af opgaver med flere items

Som udgangspunkt kan opgaver udformes som opgaver med flere items (polytome opgaver). De enkelte items skal, af hensyn til adaptiviteten, dog have omtrent samme sværhedsgrad.

Anvendelsen af denne type opgaver viser sig dog at være problematisk. Dels er mange polytome opgaver blevet afvist under afprøvningen, og dels har mange af de godkendte opgaver problemer med afhængighed mellem items og/eller items af varierende sværhedsgrad. Det er problematisk, for ikke at sige umuligt, at udlede en faglig fortolkning af besvarelsen af denne type opgaver, som vil være til nytte for læreren.

Panelet vurderer samtidig, at opgaver med mange items ikke er særligt velegnede set fra elevens synspunkt, eller særligt hensigtsmæssige i forhold til ønsket om at teste så mange trinmål/emner som muligt under testen.

Psykometrisk set er det tillige et problem, hvis der er lokale afhængigheder mellem items i en opgave, ligesom det er at foretrække, hvis de enkelte items hver især er homogene i forhold til opgavebanken.

På den anden side kan polytome opgaver give et større bidrag til afklaring af elevens dygtighed end et tilsvarende antal items i dikotome opgaver, forudsat at de har en passende sværhed for eleven. Anvendes opgaverne på en sådan måde, at de enten er for lette eller for svære for eleven, kan de derimod være mindre effektive end dikotome opgaver i denne henseende. Anvendelsen af dikotome opgaver er derfor stærkt afhængige af en velfungerende adaptiv procedure.

Da panelet vurderer, at det er vanskeligt at konstruere en serie items til en bestemt, ensartet sværhedsgrad, som samtidigt ikke er lokalt afhængige, forventer

panelet, at der også i kommende afprøvninger vil skulle kasseres relativt mange polytome opgaver.

Panelet anbefaler derfor, at der foretages omhyggelige analyser af den måde, de polytome opgaver fungerer på i forbindelse med de adaptive test, ud fra såvel et fagligt som et psykometrisk synspunkt, inden det besluttes i hvilket omfang polytome opgaver skal konstrueres og indgå i opgavebanken.

Panelet anbefaler, at polytome opgaver underkastes særlige analyser af lokal afhængighed, inden de accepteres i opgavebanken.

6.5.2. Anvend kun varianter uden genkendelsesvirkning,

En meget stor del af opgavebanken består af opgaver, der er varianter af den samme grundopgave. Panelet er opmærksomt på den fordel, det kan have for opgavekonstruktionen, men peger på, at det samtidig kan virke monotont for eleverne og må begrænse de emner, som opgavebanken beskæftiger sig med.

Panelet anbefaler derfor, at brugen af opgavevarianter i opgavebanken reduceres, eller i det mindste at udvælgelsesproceduren indrettes således, at en elev aldrig møder to eller flere varianter af samme grundopgave i et testforløb.

Panelet anbefaler dog under alle omstændigheder, at det sikres, at en elev ikke møder flere opgaver, hvor løsningen påvirkes af, at eleven har set en anden variant af opgaven, dvs. f.eks. opgaver:

- hvor eleven har fordel af at have set opgaven før, af at have sat sig ind i teksten eller i figuren.
- hvor svaret er det samme, og kun distraktorerne forskellige
- hvor opgaven er den samme, men iklædningen forskellig (forskellig skabelon, samme opgave)

6.5.3. Aldrig scrollning

Panelet vurderer, at der ganske ofte er brug for scrollning, når eleven skal danne sig et overblik over opgaven. Det kan være problematisk for både yngre og ældre elever. Dels tager det tid at orientere sig og få overblik over opgaven, og den tid går fra elevens opgaveløsning. Dels kan det virke forvirrede for de yngre elever.

Panelet anbefaler, at opgaverne konstrueres så scrollning helt kan undgås.

6.6. Tværgående faglige anbefalinger

6.6.1. Begræns antallet af opgaveskabeloner

Panelet er opmærksomt på, at det har været en målsætning at give testene variation gennem udnyttelse af it-mediet ved forskellige udformninger opgaverne (opgaveskabeloner).

Panelet vurderer imidlertid, at anvendelsen af forskellige opgaveskabeloner virker distraherende på den måde, at eleven skal bruge tid og opmærksomhed på at omstille sig til en ny type opgave. Endvidere har panelet påvist, at forskelle i opgaveskabeloner introducerer psykometriske skævheder i testene.

Panelet anbefaler derfor, at der kun bruges de opgaveskabeloner, der er fagligt nødvendige, og kun samme opgaveskabelon til samme type opgave.

Panelet har ikke hermed vurderet, om de foreliggende skabeloner generelt er tilstrækkelige og optimale til de emner, som bør testes.

6.6.2. Antallet af opgaver skal mange-dobles

Opgavebanken bør i meget høj grad udbygges. Så vidt panelet kan vurdere, skal omfanget af opgaver som minimum tre-dobles for at leve op til de oprindelige forudsætninger.

Panelet har ikke grundlag for en konkret anbefaling vedrørende opgavebankens størrelse, men vurderer dog, at en udvidelse af opgavebanken ud over det hidtil forudsatte kan være nødvendig for at

- sikre mod meget hyppig udnyttelse af bestemte opgaver
- udvide antallet af testede emner (og reducere risikoen for teach-to-the-test)
- sikre tilstrækkeligt med opgaver med passende sværhed til en sikker vurdering af dygtigheden (og samtidigt sikre, at eleven ikke møder samme opgave ved fornyet testning)
- fastholde antal items ved begrænsning af antallet af opgaver med flere items

Som led i suppleringen skal det sikres, at der er tilstrækkeligt med opgaver til de mindst - og de mest - dygtige elever, herunder til gentagne test, men i øvrigt bør opgaverne fordeles relativt til den forventede udnyttelse.

6.6.3. Mere stringent opgavekonstruktion

Den konkrete gennemgang af opgaverne har vist, at opgaverne ikke følger samme retningslinier.

Panelet anbefaler, at opgavebanken gennemgås med henblik på at få gjort brugen af distraktorer stringent og konsekvent.

Panelet anbefaler, at brugen af overskrifter defineres mere præcist og implementeres konsekvent på tværs af opgaver.

Panelet anbefaler, at opgaverne generelt korrekturlæses.

Panelet anbefaler, at upræcise illustrationer erstattes.

6.6.4. Systematisk anvendelse af metadata

Panelet har observeret, at metadata for opgaverne er delvist fraværende og i mange tilfælde ikke retvisende.

Panelet anbefaler, at samtlige metadata gennemgås, og at man, de steder hvor metadata er anvendt usystematisk og inkonsekvent, begynder forfra med opbygningen.

6.7. Specielt om Dansk/Læsning

6.7.1. Fasthold profilområder

Panelet anbefaler, at de tre profilområder fastholdes, da de er fagligt velbegrundede. For hvert af de tre profilområder anbefales, at testene gør brug af kun to opgavetyper. Anbefalingen uddybes i det følgende.

Panelet anbefaler desuden, med udgangspunkt i, at de foreslåede opgavetyper til profilområderne anvendes på samtlige klassetrin, at de it-baserede test i læsning fungerer som en lang test med fælles opgavebank. Dette vil også til dels løse nogle af problemerne med manglende lette/svære opgaver på de enkelte klassetrin.

6.7.2. To opgavetyper per profilområde

Til profilområdet *Afkodning* vil det være hensigtsmæssigt at udforme en lang afkodningstest, som kan bruges fra 2. til 8. klasse. Ordene vil skulle spænde fra de letteste, først tilegnede som *is* og *mor* til ord, der vil kunne udfordre de allerdygtigste i 8. klasse, f.eks. chihuahua og rokokokokotte.

Panelet anbefaler, at der i profilområdet *Afkodning* udelukkende tages udgangspunkt i to opgaveformater: ”find det ord der passer til billedet” og ordkædeopgaver. Disse opgaveformater er begge velafprøvede og har vist høj validitet og reliabilitet, hvor de er afprøvet.

Til profilområdet *Tekstforståelse* anbefaler panelet to opgavetyper: opgaver med tekst og ét valg mellem flere svarmuligheder og cloze-test med valg mellem flere alternativer.

Den optimale måde at teste *Sprogforståelse* på er at lade eleven høre ord og tekster og svare f.eks. ved at vælge et billede. Såfremt dette ikke er en mulighed, anbefaler panelet brug af to opgaveformater: Opgaver i ords betydning og opgaver i udtryk/faste vendingers betydning.

Snarere end at teste kendskab til eksisterende ordsprog (dansk kulturarv), kunne der være opgaver i at finde betydningen af *nye* udtryk og vendinger med ord i overførte betydninger.

Hvis de foreslåede opgavetyper til profilområderne anvendes på samtlige klassetrin, vil de it-baserede tests i læsning for de 3 profilområder kunne fungere som en lang test med fælles opgavebank. Dette vil også til dels løse nogle af problemerne med manglende lette/svære opgaver på de enkelte klassetrin.

6.8. Specielt om Matematik

6.8.1. Fasthold profilområder

Opdelingen i profilområderne *Tal og Algebra*, *Geometri* og *Matematik i anvendelse* er fagligt velfungerende, velkendte for matematiklærerne, og de udvalgte trinmål repræsenterer faget rimelig bredt. Opgaverne vurderes for størstedelens vedkommende fagligt at være i orden, dog kan mange af opgavernes sprogligt forenkles en del.

6.8.2. Analyser validitet

Panelet har fundet, at det i forhold til en række opgaver i Matematik er usikkert, hvad de tester, bl.a. pga. forskellige opgaveskabeloner, sproglige instruktioner, tekstmængde m.m. Panelet anbefaler derfor, at sådanne opgaver udsættes for en nærmere analyse.

6.8.3. Revurder af opgaveskabeloner

Det anbefales, at den videre udvikling af faget Matematik i højere grad præciserer hvilke opgaveskabeloner, der er velegnede til udvikling mod de enkelte trinmål inden for faget.

6.9. Specielt om Fysik/Kemi

6.9.1. Fasthold profilområder, ændre opgavekonstruktionen

Panelet anbefaler, at profilområderne *Energi og energiomsætning* samt *Stoffer og materialer* bibeholdes samtidig med profilområdet *Anvendelse og perspektiver*, da de fagligt vurderet er de rette profilområder at anvende.

Panelet anbefaler dog samtidig, at man ændrer måden at konstruere opgaver på, således at en langt større del af opgaverne tester forståelse og anvendelse af viden på baggrund af generelle faglige principper. Herved bør de to sammenfaldende profilområder også blive forskellige. Ideelt set bør der ikke være opgaver, der tester paratviden – men gerne opgaver, der tester paratforståelse.

6.9.2. Inddrag flere emner

Panelet anbefaler, at der konstrueres langt flere opgaver inden for Fysik/Kemi for at opnå en større variation af faglige problemstillinger. Der bør udvikles et koncept for, hvordan man kan teste elevens viden om naturvidenskabelige arbejdsprocesser og evne til at argumentere fagligt ud fra visse givne oplysninger.

Gennem disse tiltag reduceres risikoen for, at en forudsigelig opgavebank fører til ’teaching to the test’.

6.9.3. Anvend kun entydige distraktorer

Panelet har fundet, at en del af de svarmuligheder, der gives i nogle opgaver, og som er tiltænkt at give et forkert svar, rent faktisk kan være mulige rigtige svar på

opgaven. De mulige svar er blot ikke noget, eleven er blevet undervist i. En vidende fagmand eller den meget dygtige elev besidder denne viden, hvorfor den meget dygtige elev vil have en større sandsynlighed for at få fejl i disse opgaver.

Panelet anbefaler at opgaver med denne type distraktorer enten ændres eller fjernes helt.

6.9.4. Opstramning af illustrationer og sprog

Panelet anbefaler, at undervisningens indhold perspektiveres mere bredt gennem brug af varierede illustrationer, herunder illustrationer, der er mindre ordinære, f.eks. ved at illustrere energi med mad i stedet for med olietønder.

Panelet anbefaler, at antallet af billedbaserede svaritems begrænses, samt at antallet af parametre, der skal aflæses på den enkelte illustration, begrænses

Endvidere anbefaler panelet, at testen tilpasses med en konsistent brug af terminologi (herunder brug af danske kontra latinske/græske begreber), således at begge sæt konsekvent anvendes med det ene (og altid det samme) i parentes. Ligeledes anbefaler panelet en mere konsekvent navngivning af kemiske forbindelser.