MINISTERIET FOR
BØRN, UNDERVISNING
OG LIGESTILLING
STYRELSEN
FOR IT OG LÆRING

De nationale tests måleegenskaber

September 2016

De nationale tests måleegenskaber

BAGGRUND

De nationale test blev indført i 2010 for at forbedre evalueringskulturen i folkeskolen. Hensigten var bl.a. at give lærerne et bedre indblik i elevernes faglige niveauer gennem deres skoletid – også set i forhold til det faglige niveau blandt resten af landets elever. Tidligere var det folkeskolens afgangsprøve, der var den primære kilde til viden om elevernes faglige niveauer set i forhold til resten af landet.

Testene udgør et blandt flere værktøjer, som kan bidrage til, at læreren får overblik og kan vurdere elevernes udbytte af undervisningen. Da det ikke er alle områder af fagene, der kan eller skal testes med de nationale test, kan testresultaterne ikke stå alene i evalueringen af eleverne. Resultaterne kan også bidrage til skole-hjem-samarbejdet.

Der er ti obligatoriske test á 45 minutters varighed i løbet af elevernes skoletid. Disse er fordelt på seks forskellige fag og seks forskellige klassetrin. Fire af de ti test er i dansk, læsning på fire forskellige klassetrin og to af testene er i matematik. Alle test består af tre profilområder, som afgrænser de områder af faget, som eleverne testes i.

HVAD ER AFGØRENDE FOR TESTENES MÅLEEGENSKABER?

Hvor god en test er til at vurdere elevernes faglige niveau i et område af faget afhænger blandt andet af den tid, der er afsat til at afvikle testen. Jo længere tid eleverne testes, jo flere opgaver – og dermed bedre grundlag – er der til at bedømme elevens faglige niveau ud fra. Omvendt kan særligt de yngre elever blive trætte og ukoncentrerede, hvis testene varer for længe. De nationale test varer som udgangspunkt 45 minutter.

De nationale test bygger på en adaptiv algoritme, som løbende tilpasser opgavernes sværhedsgrader til den enkelte elevs niveau. Det betyder, at eleven starter med en middelsvær opgave, og hvis eleven svarer korrekt, er den næste opgave lidt sværere. Hvis eleven svarer forkert, er den næste opgave lidt lettere. Det fortsætter, indtil elevens faglige niveau er bestemt med en vis sikkerhed. Metoden optimerer testenes måleegenskaber og gør det muligt at opnå en vurdering af elevens faglige niveau med størst mulig sikkerhed inden for rammerne af en typisk lektion på 45 minutter.

Fakta – sådan bliver opgaverne til

Opgaverne til de nationale test bliver udviklet af faglige opgavekommissioner, der er nedsat inden for hvert fag. Her udvikler fagfolk opgaver, der har høj kvalitet og er tilpasset de områder af faget, som testes. Opgaverne bliver udviklet på baggrund af de Fælles Mål, der er fastsat inden for faget. I testene inddrages kun de områder af Fælles Mål, som kan testes inden for rammerne af it-baseret og selvrettende test.

De nationale test trækker på spørgsmål fra en stor opgavebank, men inden opgaverne finder vej til den, bliver de afprøvet på ca. 700 elever. Her gennemgår opgaverne en omfattende statistisk analyse, som både vurderer, om opgaverne måler på det, de skal, og som konsoliderer den enkelte opgaves sværhedsgrad. Når det er sket, kommer opgaverne ind i opgavebanken, som løbende bliver opdateret for at sikre, at der er tilstrækkelige opgaver på alle sværhedsgrader.

DEBAT OM NATIONALE TEST

Den debat, der har været om de nationale tests måleegenskaber, har hovedsageligt drejet sig om tre spørgsmål:

- Måler testene det, de skal?
- Hvor god er testen til at vurdere den enkelte elevs faglige niveau?
- Måler testene det samme, når de måler eleverne to gange i træk?

Nedenfor vil de tre spørgsmål blive kommenteret. Der er særligt lagt vægt på at kommentere og illustrere via fagene dansk læsning og matematik, da det er i de fag, at eleverne testes flest gange i løbet af skoletiden.

MÅLER TESTENE DET, DE SKAL?

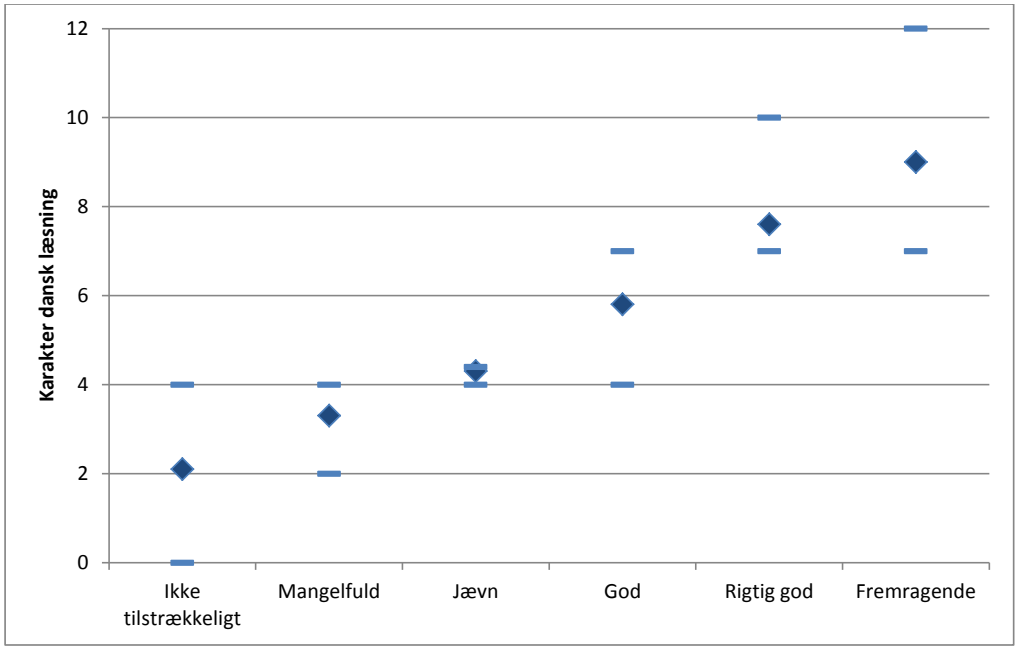
Et af de kritikpunkter, der har været af de nationale test, har gået på, at testene måler for snævert i forhold til de færdighedsområder, det er meningen, at de skal måle på. Der testes alene i færdigheder, som det er muligt at afprøve i en it-baseret og selvrettende test. Derfor bør testresultaterne aldrig stå alene i evalueringen af elevernes undervisningsudbytte.

Hver test tester i tre faglige områder, de såkaldte profilområder. For eksempel består testen i dansk læsning af en test i sprogforståelse, en test i afkodning og en test i tekstforståelse. Det er altså kun dele af faget, eleven bliver testet i, og det gør sig også gældende for de øvrige fag.

For at få en indikation af om testene samlet set ser ud til at måle det samme som andre tilsvarende test og prøver, kan man se på sammenhængen mellem elevernes testresultat i de nationale test og deres efterfølgende præstation i de relevante dele af folkeskolens prøver i 9. klasse.

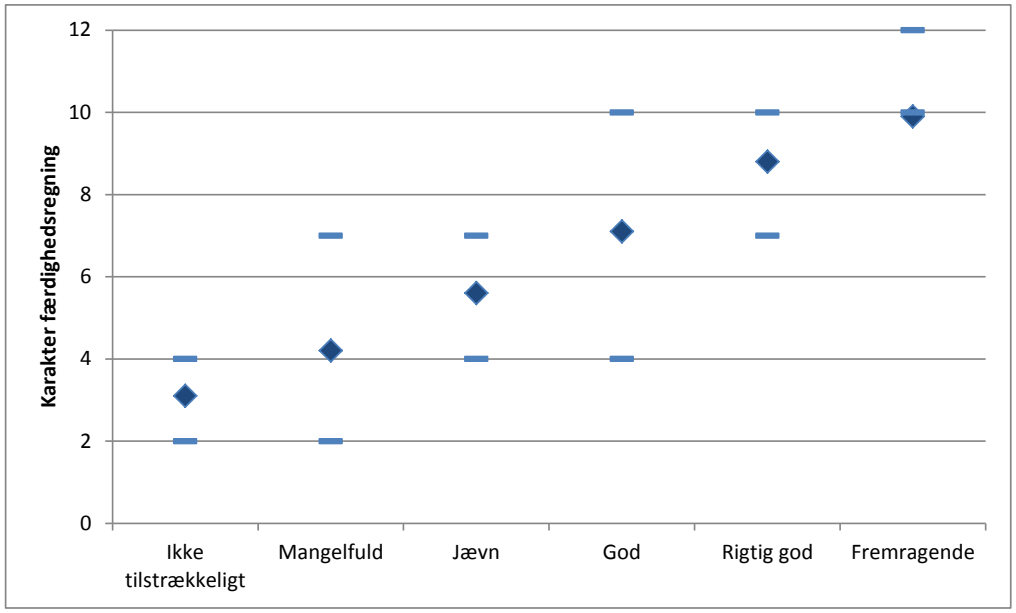
Den øvelse er lavet i figur 1 og figur 2 for den elevårgang, der tog 9.klasseprøver i foråret 2015. Figuren viser, at elevernes tidligere testresultater i matematik og læsning i 6. og 8. klasse hænger tæt sammen med deres efterfølgende karakterer i hhv. færdighedsregning og læsning i folkeskolens prøver i 9. klasse. For begge fag er der en statistisk signifikant sammenhæng.

Figur 1: De nationale test i dansk læsning 8. klasse og folkeskolens prøve i 9. klasse



Anm.: Gennemsnitskarakter (firkant) samt 25 pct. og 75 pct. percentiler (vandret streg)

Figur 2: De nationale test i matematik 6. klasse og folkeskolens prøve i 9. klasse



Anm.: Gennemsnitskarakter (firkant) samt 25 pct. og 75 pct. percentiler (vandret streg)

Elever, der opnår et testresultat i de nationale test i dansk læsning 8. klasse på niveauet 'God', får med stor sandsynlighed karakteren 4 eller 7[1] i folkeskolens prøve året efter, mens elever, der opnår et testresultat i dansk læsning 8. klasse på niveauet 'Rigtig god', med stor sandsynlighed får karakteren 7 eller 10 i folkeskolens prøve året efter.

I en rapport fra konsulentfirmaet DAMVAD i 2014[2] påvises det i øvrigt, at der er en sammenhæng mellem de resultater eleverne opnår i de nationale test og i den internationale PISA-undersøgelse, jf. boks 1. Dette gælder både for testene i dansk og matematik.

Boks 1. Uddrag af Damvad-rapport om PISA og de nationale test (s. 5):

"Der er en tydelig sammenhæng mellem resultaterne fra de nationale test og resultaterne fra PISA-undersøgelserne. Sammenhængen kan observeres på tværs af profilområder i både læsning og matematik, men er ikke nødvendigvis jævnt fordelt."

"Den tydelige sammenhæng mellem resultaterne fra de nationale test og PISA betyder samtidig, at de to test uafhængigt af hinanden når til relativt enslydende vurderinger af elevers faglige niveauer. Det er en bekræftelse af, at de nationale test siger noget relevant om elevernes faglige niveau i de områder, hvori de testes."

HVOR GOD ER TESTEN TIL AT VURDERE DEN ENKELTE ELEVS FAGLIGE NIVEAU?

En anden kritik er gået på, om testene har for høj en statistisk usikkerhed i forhold til at vurdere elevernes faglige niveau i de områder af faget, som testes.

I de nationale test er det muligt af få angivet den statistiske usikkerhed på elevens testresultat[3]. Dette er ikke en mulighed i mange andre test og prøver.

Elevernes resultater i testene afrapporteres via forskellige skalaer. På den *kriteriebaserede skala*, der bl.a. kan anvendes i forældrebrevene, er der seks niveauer rangerende fra "ikke tilstrækkelig" til "fremragende".

Det faglige niveau, elevens testresultat er beregnet til, er det mest sandsynlige på baggrund elevens testresultat, men det kan ikke afvises med en mindre sandsynlighed, at elevens testresultat ligger lige over eller under. Nogle elevers faktiske niveau kan ligge i gråzonen mellem to niveauer, hvilket gør vurderingen af, om en elev fx skal vurderes "god" eller "rigtig god", mere usikker.

[1] Henholdsvis 25 pct. og 75 pct. percentiler

[2] PISA-relatering af de kriteriebaserede nationale test. DAMVAD 2014 (http://www.uvm.dk/-/media/UVM/Filer/Udd/Folke/PDF14/Okt/141008-Kriteriebaserede-test-delrapport-1.ashx)

[3] En fordel ved den måde, de nationale tests er bygget op på, er, at læreren undervejs i testafviklingen kan se en vurdering af den statistiske sikkerhed i vurderingen af elevernes faglige niveau på sin skærm. Det giver læreren mulighed for at lade testen vare længere end de normale 45 minutter, hvis læreren vurderer, at det er nødvendigt for at opnå en højere sikkerhed.

Beregninger, jf. tabel 1, viser, at ni ud af ti testresultater (91 pct.) med statistisk sikkerhed[4] vurderes rigtigt i det beregnede faglige niveau eller i enten niveauet lige under eller niveauet lige over. De resterende 9 procent af elevernes testresultater har en større usikkerhed, der betyder, at elevens faktiske niveau ikke kan afvises at ligge i både niveauet lige under og i niveauet lige over det målte niveau. Størstedelen af disse elever er elever, som vurderes til en jævn præstation.

Tabel 1: Den statistiske sikkerhed på elevens testresultat på den kriteriebaserede skala

Elevens testresultat ligger med stor sandsynlighed i ...			
... det beregnede faglige niveau	... det beregnede faglige niveau eller niveauet lige under	... det beregnede faglige niveau eller niveauet lige over	... det beregnede faglige niveau eller niveauet lige over eller lige under
28 pct.	34 pct.	29 pct.	9 pct.

Anm: Enkelte testresultater (<0,02 pct.) er mere usikkert bestemt

Den *normbaserede percentilskala* er en værdi fra 1 til 100, som læreren kan bruge til sin egen bearbejdning af elevernes resultater. Det er således ikke en værdi, der oplyses i forældrebrevene. Lærere kan bruge værdien til at få en mere detaljeret vurdering af eleverne i forhold til den mindre finmaskede kriteriebaserede skala.

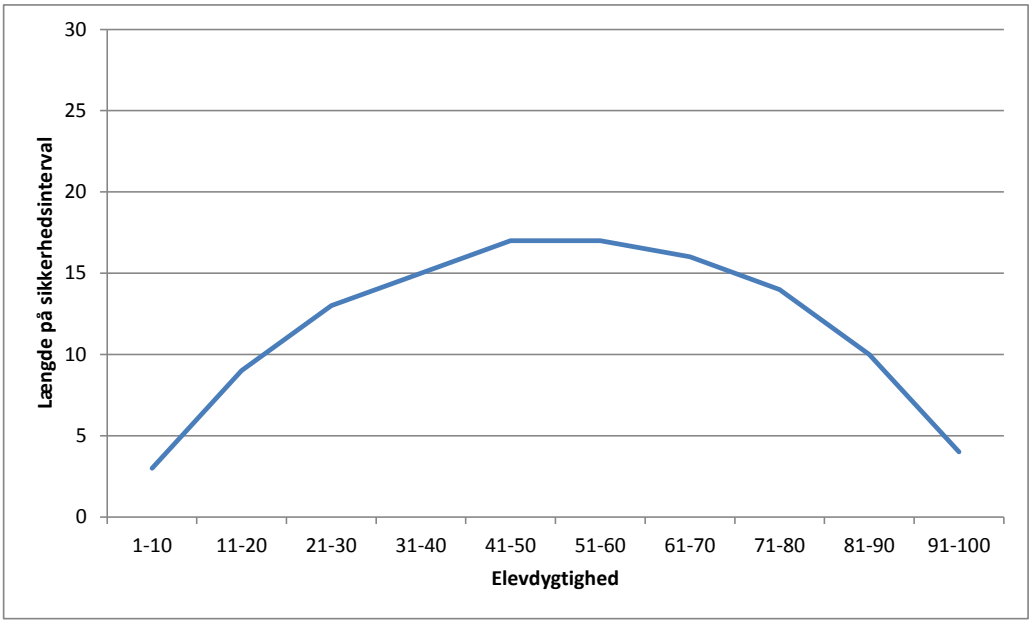
Usikkerheden er mere synlig på den normbaserede percentilskala. I gennemsnit er usikkerheden på ca. ± 12[5] point. Det vil sige, at det ikke kan afvises, at en elev, der scorer 75 point, reelt kan have en score, der ligger mellem 63 og 87 point. Det er vigtig at understrege, at elevens beregnede score er den mest sandsynlige værdi, men der er en vis sandsynlighed for, at den reelle score afviger fra denne.

Som figur 3 viser, er usikkerheden, omregnet til percentilskalaen, størst for de elever, der scorer middel, mens den er mindre for elever med høje eller lave scorer.

[4] Der er her anvendt et sikkerhedsinterval på ± 1*SEM svarende til et 67 pct. sikkerhedsinterval til vurdering af **usikkerheden på individniveau**. Til vurdering af usikkerheden på et gennemsnit anvendes ofte et sikkerhedsinterval på ± 2*SEM svarende til et 95 pct. sikkerhedsinterval

[5] Der er tale om en lille tilnærmelse, da sikkerhedsintervallerne på percentilskalaen ikke er helt symmetriske

Figur 3: Den statistiske sikkerhed på elevens testresultat på percentilskalaen



MÅLER TESTENE DET SAMME, NÅR DE MÅLER ELEVERNE TO GANGE I TRÆK?

Endelig har der været sat spørgsmålstegn ved, om de nationale test måler ensartet, når eleverne gennemfører den samme test to gange med kort mellemrum.

De nationale test er som udgangspunkt udviklet som et redskab til de obligatoriske målinger på bestemte klassetrin. Muligheden for at gennemføre frivillige nationale tests i efterårssemesteret har dog gennem de seneste år været stigende. I alt gennemførte knap 320.000 elever i efteråret 2015 de frivillige nationale test. Ca. 35.000 af dem gennemførte to på hinanden følgende frivillige tests i samme fag.

Når man gennemfører to på hinanden følgende tests med få ugers mellemrum[6] er der mange faktorer, der kan spille ind i forhold til, om man kan sammenligne de to testresultater. Lærerens instruktioner og formålet med de to hurtige testafviklinger, elevens motivation og koncentration samt stabiliteten af lokalt it-udstyr er nogle af de forhold, der kan påvirke et testforløb.

I tabel 2 er modellen bag de nationale test afprøvet via computersimuleringer for at vurdere selve modellens målepræcision uafhængigt af elevernes motivation m.v., der måtte have betydning ved at afvikle to test med kort tids mellemrum. Konkret er testafviklingerne simuleret med to gentagne elevforløb for 5.000 elever.

[6] I gennemsnit var der 20 dage mellem

Udtrykt på percentilskalaen er forskellen i den beregnede elevdygtighed mellem de to simuleringer i gennemsnit lig nul med et interkvartilt[7] range på [-8; +8].

I alle profilområder er der desuden en statistisk signifikant positiv sammenhæng mellem elevdygtigheden bestemt ved de to simulerede testforløb[8]. Med undtagelse af de nationale test i sprogforståelse (profilområde 1) i dansk læsning 2. klasse ligger alle korrelationerne[9] i intervallet 0,82 - 0,93.

Tabel 2 Korrelationen mellem elevdygtigheden ved to simulerede testforløb

Test	Profilområde 1	Profilområde 2	Profilområde 3
Dansk læsning 2. klasse	0,78	0,93	0,91
Dansk læsning 4. klasse	0,82	0,89	0,90
Dansk læsning 6. klasse	0,82	0,86	0,87
Dansk læsning 8. klasse	0,84	0,87	0,88
Matematik 3. klasse	0,90	0,86	0,82
Matematik 6. klasse	0,89	0,86	0,89

Ser man på de faktiske resultater fra de ca. 35.000 elever, der gennemførte to på hinanden følgende frivillige tests i efteråret 2015, er der ligeledes en positiv statistisk signifikant sammenhæng. Denne sammenhæng er dog en anelse svagere end i de simulerede elevforløb.

Forskellen i korrelationerne baseret på observerede og simulerede elevforløb viser, at elevadfærden har en vis indflydelse på muligheden for at opnå det samme testresultat ved at gentage den samme test. Hvis man som skole afvikler de frivillige test med få ugers mellemrum, bør man således være særligt opmærksom på at tolke resultaterne varsomt og ud fra de forhold, som testene er afviklet under.

[7] 25 pct. og 75 pct. percentiler

[8] Korrelationskoefficienten er et udtryk for sammenhængen mellem to målinger og ligger i intervallet fra -1 til +1. Guideline til vurdering af korrelations koefficienter: '0,0-0,2'=meget svag; '0,2-0,4'=svag; '0,4-0,6'=moderat; '0,6-0,8'=stærk; '0,8-1,0'=meget stærk. (Evans, J.D. 1996: Straightforward statistics for the behavioral sciences)

[9] Korrelationerne er beregnet på baggrund af testresultater på logit skalaen