

Evalueringen af de nationale test

Tværgående evalueringsrapport



Lasse Hønge Flarup

Evalueringen af de nationale test – Tværgående evalueringsrapport

© VIVE og forfatterne, 2020

e-ISBN: 978-87-7119-740-2

Arkivfoto: Lars Degnbol/VIVE

Projekt: 301403

VIVE – Viden til Velfærd

Det Nationale Forsknings- og Analysecenter for Velfærd

Herluf Trolles Gade 11, 1052 København K

www.vive.dk

VIVEs publikationer kan frit citeres med tydelig kildeangivelse.

Forord

Folketinget vedtog i marts 2006 indførelsen af de nationale test. Den første obligatoriske testrunde blev gennemført i foråret 2010. De nationale test var ét blandt flere elementer i et lovforslag fra december 2005 om fornyelse af folkeskolen for at forbedre det faglige niveau blandt eleverne gennem styrket, løbende evaluering i folkeskolen.

Børne- og Undervisningsministeriet igangsatte evalueringen af de nationale test med udgangspunkt i to overordnede undersøgelsesspørgsmål på baggrund af anbefalinger fra den nedsatte rådgivningsgruppe:

Evalueringens to undersøgelser

En analyse af den statistiske usikkerhed, reliabiliteten og øvrige måleegenskaber forbundet med de nationale test.

En undersøgelse af betydningen og brugen af de nationale test.

Rådgivningsgruppen har udarbejdet forslag til de samlede undersøgelsesspørgsmål, der er afgivet til Børne- og Undervisningsministeriet. Ministeriet har med en enkelt tilføjelse om ”fagenes formål” bedt Styrelsen for It og Læring (STIL) og VIVE – Det Nationale Forsknings- og Analysecenter for Velfærd om at udarbejde henholdsvis første og anden undersøgelse. VIVE har endvidere haft til opgave at sammenfatte en kvalitetssikring af STILs tekniske beregninger og analyser af de nationale test gennem et forskerreview.

Evalueringen bygger på en kombination af STILs dokumentation, registerdata, en systematisk litteratursøgning, interview, observationer, surveydata og workshops.

Evalueringen af de nationale test består af seks rapporter og en bilagsrapport: én tværgående evalueringsrapport og fem delrapporter, der omhandler hvert sit emne, samt en bilagsrapport til delrapport 5. Chefanalytiker Lasse Hønge Flarup er projektleder på evalueringen af de nationale test og har udarbejdet den tværgående evalueringsrapport med udgangspunkt i de fem delrapporter:

Delrapport 1: Review af evalueringen af de statistiske aspekter ved de nationale test
Af forsker Peter Rohde Skov og chefanalytiker Lasse Hønge Flarup

Delrapport 2: De nationale tests samvariation med karakterer
Af forsker Peter Rohde Skov og chefanalytiker Lasse Hønge Flarup

Delrapport 3: Kortlægning af sammenlignelige test
Af forsker Tine Louise Mundbjerg Eriksen, chefanalytiker Lasse Hønge Flarup og forsker Peter Rohde Skov

Delrapport 4: De nationale tests sammenhæng med fagenes formål
Af senioranalytiker Martin Foldager Hindsholm, analytiker Niels Westermann Brændgaard og chefanalytiker Lasse Hønge Flarup

Delrapport 5: Anvendelsen af de nationale test inkl. bilagsrapport

Af seniorforsker Bente Bjørnholt, chefanalytiker Lasse Hønge Flarup, senioranalytiker Martin Foldager Hindsholm, analytiker Niels Westermann Brændgaard, praktikant Christina Munkholm Andersen, forsker Niels Bjørn Grund Petersen og forsker Sidsel Vive Jensen.

Bibliotekar Anne Nørgaard-Pedersen har stået for litteratursøgningen. Studenterne Cecilia Juel Schlosser, Emilie Hestbæk Jacobsen, Amalie Damgaard Johansen, Clara Maria Pedersen, Helene Kni Rasmussen, Sara Lentz Jørgensen, Anders Winkler, Cecilie Bundgaard Lohse, Cianne Isabel, Emil Bakkensen Johansen, Helena Elisabeth Ravn, Julie Lund Hansen, Karl Magnus Møller, Matthias Røy Wagner, Sofie Jarlstrøm Clausen samt videnskabelig assistent Ronja Rosenberg Grøn har bidraget til dataindsamling og databehandling.

Rapporterne har været i eksternt review og er blevet kvalitetssikret af forskere og praktikere på feltet. Rapporterne er endvidere kvalitetssikret af forsknings- og analysechef Carsten Strømbæk Pedersen, udviklingsdirektør Mette Deding, forskningsdirektør Torben Tranæs og forsknings- og analysechef Mads Leth Jakobsen.

Vi takker for værdifulde kommentarer fra de tilknyttede medarbejdere fra Børne- og Undervisningsministeriet. Vi takker desuden de mange forvaltningschefer, forskere, opgavekommissionsmedlemmer, politikere, skoleledere, lærere, elever og forældre, der har deltaget i undersøgelsen gennem spørgeskemaer, interview, workshops. Uden deres deltagelse ville disse rapporter ikke have været mulige.

Carsten Strømbæk Pedersen

Forsknings- og analysechef for VIVE Børn og Uddannelse
2020

Indhold

1	Evalueringen af de nationale test	6
1.1	Formål	6
1.2	Læsevejledning	8
2	Sammenfatning af evalueringen af de nationale test	9
2.1	Tværgående perspektivering	9
2.2	Tværgående resultater	11
3	De nationale test	24
3.1	Baggrund	24
3.2	Testenes indhold	25
	Litteratur	30

1 Evalueringen af de nationale test

Børne- og Undervisningsministeriet igangsatte evalueringen af de nationale test med udgangspunkt i to overordnede undersøgelsesspørgsmål på baggrund af anbefalinger fra den nedsatte rådgivningsgruppe:

Evalueringens to undersøgelser

En analyse af den statistiske usikkerhed, reliabiliteten og øvrige måleegenskaber forbundet med de nationale test.

En undersøgelse af betydningen og brugen af de nationale test.

Rådgivningsgruppen har udarbejdet forslag til de samlede undersøgelsesspørgsmål, der er afgivet til Børne- og Undervisningsministeriet. Ministeriet har med en enkelt tilføjelse om ”fagernes formål” bedt STIL og VIVE om at udarbejde henholdsvis første og anden undersøgelse. VIVE har dog også til opgave at sammenfatte en kvalitetssikring af STILs tekniske beregninger og analyser af de nationale test gennem et forskerreview.

Denne rapport er den tværgående delrapport i VIVEs samlede evaluering af de nationale test. Evalueringen er både summativ og formativ og danner grundlag for en redegørelse til Folketinget, ligesom den danner grundlag for, at rådgivningsgruppen udarbejder anbefalinger. Det summative sigte har til formål at se på resultaterne af brugen af de nationale test. Det formative sigte anvendes med henblik på, at den viden, evalueringen bibringer, kan indgå i de valg, der træffes om den fremadrettede brug og udvikling af de nationale test. VIVEs evaluering af de nationale test belyser styrker såvel som svagheder i indholdet og brugen af de nationale test i folkeskolen.

1.1 Formål

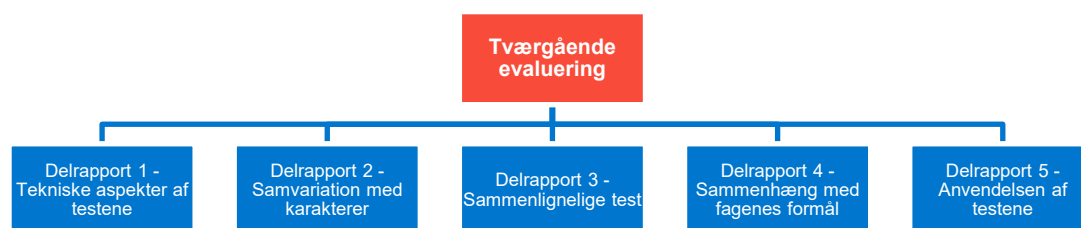
Denne tværgående rapport samler resultaterne fra evalueringen af de nationale test. Evalueringen har til formål at belyse styrker såvel som svagheder omkring indholdet og brugen af de nationale test i folkeskolen samt give et vidensgrundlag, der kan danne afsæt for det fremadrettede arbejde med udvikling og brug af de nationale test i folkeskolen. Evalueringen svarer konkret på følgende, overordnede undersøgelsesspørgsmål:

Undersøgelsesspørgsmål

Har de nationale tests indhold og udformning styrket skolernes evalueringskultur og derigennem elevernes faglige niveau?

Evalueringen af de nationale test kan groft sagt deles ind i to aspekter – der ser på henholdsvis indholdet og anvendelsen. Undersøgelsen af indholdet af de nationale test består af fire separate undersøgelser, der sætter fokus på forskellige aspekter af testenets egenskaber. Undersøgelsen af anvendelsen af de nationale test ser på, hvordan aktører på alle niveauer anvender

testenes resultater. VIVE har struktureret besvarelsen i følgende fem delrapporter samt en tværgående evaluering.



De fire første delrapporter omhandler primært egenskaber ved de nationale test og sammenlignelige test, mens delrapport 5 omhandler anvendelsen af testene til evaluering. Nedenstående figur illustrerer indholdet af de fem delrapporter.

Datagrundlag

Den tværgående evaluering baseres primært på de fem delrapporter samt materiale vedrørende baggrunden for udviklingen af de nationale test og materiale udarbejdet i forbindelse med de nationale test af Børne- og Undervisningsministeriet. Delrapporterne refererer desuden løbende til hinanden og inddrager relevant viden på området. De fem delrapporter baseres på en kombination af forskellige datakilder. Datakilderne uddybes i de enkelte delrapporter. Herunder beskrives de kort:

Delrapport 1 baseres på Styrelsen for It og Lærings evaluering af de statistiske aspekter af de nationale test samt fire forskere på områdets review af samme evaluering.

Delrapport 2 baseres primært på registerdata for de nationale test og karakterer i folkeskolens afgangsprøver.

Delrapport 3 baseres på en systematisk litteratur- og testsøgning i online databaser og hjemmesider.

Delrapport 4 baseres på registerdata over samtlige aktive opgaver i de nationale test samt kvalificerende workshops med medlemmer af de opgavekommissioner, der udvikler opgaverne.

Delrapport 5 baseres på spørgeskemadata fra lærere, skoleledere, kommunale forvaltninger, forskere, spørgeskemadata fra evalueringen af de nationale test i 2013, observationsdata fra elever og lærere, interviewdata fra elever, lærere, vejledere, skoleledere, skolebestyrelsesformænd, kommunalforvaltninger, kommunalpolitikere, folketingspolitikere, workshopdata fra medarbejdere i Børne- og Undervisningsministeriet, testdata om de nationale test.

Delrapport	Undersøgelsesspørgsmål	Indhold
1. Review af evalueringen af de statistiske aspekter ved de nationale test	• Har STIL på tilfredsstillende vis besvaret rådgivningsgruppens evalueringsspørgsmål om de nationale tests statistiske usikkerhed, reliabilitet og øvrige måleegenskaber?	Sammenfatter eksterne revieweres bedømmelse af STILs gennemgang af de tekniske aspekter af de nationale test. Undersøger testenes statistiske usikkerhed, validitet, reliabilitet og øvrige måleegenskaber.
2. De nationale tests samvariation med karakterer	• Hvad er samvariationen mellem elevers præstationer i testene og karakterer i 8. og 9. klasseprøverne?	Undersøger og giver svar på samvariationen mellem elevers resultat i de nationale test og samme elever i folkeskolens 8. og 9. klasseprøver. Undersøger, hvor valide testene er på gruppeniveau.
3. Kortlægning af sammenlignelige test	• Hvilke test findes, der i formål, indhold og omfang minder om de danske nationale test?	Kortlægger nationale og internationale test på baggrund af en række karakteristika. Undersøger andre tests karakteristika sammenlignet med de nationale test og giver et overordnet billede af testlandskabet til inspiration.
4. De nationale tests sammenhæng med fagernes formål	• I hvilket omfang er der sammenhæng mellem de nationale test og de centrale dele af faget og fagernes formål jf. Fælles Mål?	Undersøger sammenhængen mellem opgaverne i de nationale test og Fælles Mål for de fire obligatoriske testfag: dansk (læsning), matematik, engelsk og fysik/kemi.
5. Anvendelsen af de nationale test	• Hvordan opleves de nationale test som evalueringsredskab? • Hvordan bruges de nationale test i dialogen og opfølgningen på tværs af lokale politikere, forvaltning, skoleledere, lærere, elever og forældre?	Undersøger, hvorvidt og hvordan de nationale test anvendes som evalueringsredskab alene og i sammenhæng med andre datakilder og evalueringer på nationalt, kommunalt og skoleniveau. Fokus er særligt, hvorvidt og hvordan nationale test understøtter en evalueringskultur inden for og på tværs af niveauer.

1.2 Læsevejledning

Rapporten er inddelt i tre kapitler. Kapitel 1 beskriver overordnet designet af evalueringen af de nationale test. Kapitel 2 beskriver de tværgående perspektiver af evalueringen og sammenfatter resultaterne af de enkelte delrapporter. Kapitel 3 beskriver de nationale tests baggrund og indhold.

2 Sammenfatning af evalueringen af de nationale test

Evalueringen af de nationale test afreporteres i fem selvstændige, men forbundne delrapporter. Hver delrapport har selvstændige analyser og afdækker selvstændige undersøgelses-spørgsmål. Der er dog en række temaer, der bliver berørt i flere af delrapporterne.

Sammenfatningen i den tværgående rapport deles op i to underkapitler. Det første underkapitel 2.1 indeholder de tværgående og overordnede konklusioner, som evalueringen som en samlet analyse har fundet frem til. Det andet underkapitel 2.2 indeholder en sammenfatning af resultaterne fra hver delrapport. For en mere fyldig gennemgang af resultaterne og analyserne henvises til de konkrete delrapporter.

2.1 Tværgående perspektivering

Den tværgående perspektivering samler op på evalueringens fund og sætter dem ind i en fælles kontekst. Den tværgående perspektivering er udtryk for VIVEs samlede vurdering på baggrund af de indsamlede data og de gennemførte analyser.

De nationale test er en typisk test ...

De nationale test er én faglig test blandt mange forskellige faglige test og prøver i grundskolen.

Faglige test måler områder inden for et fag og ikke hele faget eller alle aspekter af et fags formål. Folkeskolens afgangsprøver dækker heller ikke alle aspekter af fagene. De nationale test måler tilsvarende kun de dele af faget, som der testes i. De er således ikke udtryk for elevers fulde kunnen inden for et fag, men et udtryk for deres kunnen inden for de områder, der testes i. Og det ved praktikerne godt. Resultaterne fra de nationale test anvendes derfor primært som én videnskilde blandt flere supplerende videnskilder, ligesom Børne- og Undervisningsministeriets vejledninger til de nationale test også beskriver anvendelsesmulighederne.

Alle tests resultater er forbundet med en grad af usikkerhed. Og denne usikkerhed er forbundet med testens evne til at måle præcist og til at måle det, den er designet til at måle. De nationale tests resultater er også forbundet med usikkerhed.

... og en atypisk test

De nationale test har nogle karakteristika, der gør dem specielle i sammenligning med andre test. Både i Danmark, og når man sammenligner med andre landes erfaringer.

De nationale test har et dobbelt formål, der er rettet mod både pædagogisk brug og brug som styringsredskab. Det gør testen speciel sammenlignet med flertallet af andre test i ind- og udland.

Det adaptive princip, hvor testen tilpasser opgavers sværhedsgrad til eleven, er også relativt sjældent. Der er ikke andre test i Danmark, der gør det samme, og meget få test i udlandet. Det er derfor ikke en testform, der på nuværende tidspunkt er meget erfaring med fra andre steder end de danske nationale test.

Validitetsdiskussionen om de nationale test fylder

Diskussionen vedrørende de nationale tests validitet og reliabilitet fylder og har fyldt meget både i medier og i praksis. Evalueringen indikerer, at diskussionen har givet anledning til tvivl hos praktikerne, hvilket har haft betydning for deres syn på testene og deres anvendelsespotentiale. Man står i en venteposition, hvor man er i tvivl om, hvorvidt man kan stole på resultaterne eller ej.

Evalueringen viser, at de nationale test er usikre, når det kommer til den enkelte elevs resultat. Usikkerhed på elevniveau er forventeligt blandt lignende test. Der findes dog meget lidt viden om, hvor usikre andre test er på elevniveau, da området er relativt uudforsket og vanskeligt at sammenligne på tværs af test.

Resultaterne kan anvendes til generalisering og styring, da de har høj eksternt validitet og til en vis grad også er internt valide, dvs. måler det, de er designet til at måle. Høj ekstern validitet betyder, at de kan generaliseres til et udtryk for elevernes faglige niveau. Evalueringen viser også, at der er mulighed for at forbedre både målesikkerheden og den interne validitet.

Det vil sige, at den tvivl, mange har om, hvorvidt man meningsfuldt kan anvende data på aggregeret niveau, når nu data er usikre for den enkelte elev, bør være afklaret. Det kan man godt, men selvfølgelig inden for de metodiske rammer, som er gældende for data af denne type, og niveauet man aggregerer til. Det skal bemærkes, at der ikke er noget, der tyder på, at de nationale test er mindre eller mere pålidelige end andre sammenlignelige test.

Testsituationen rummer dilemmaer

Selve testsituationen rummer dilemmaer. Nogle elever oplever testsituationen positivt, mange oplever den som neutral og få oplever den negativt. Særligt blandt de yngre elever er der udfordringer i forhold til længden af testen. Ligeledes spiller det adaptive princip ind i oplevelsen af testsituationen, hvor nogle, både lærere og elever, oplever det som ubehageligt, dels at længden på testen kan forlænges, dels at alle elever stilles spørgsmål, som er for svære at besvare.

Længden af testen hænger sammen med testens præcision. Jo længere en test, desto mere præcis test, da man vil kunne svare på flere opgaver. Så hvis man forkorter testens længde, så bliver testen mere upræcis.

Testens adaptive princip er i teorien med til at forkorte testens længde, da det gør det muligt hurtigere at finde elevens niveau. Så en afskaffelse af det adaptive princip vil alt andet lige kræve en længere test for at opnå et lige så præcist resultat.

På samme måde vises resultaterne fordelt på de tre profilområder inden for faget. Det giver et større detaljeringsniveau i forhold til at teste forskellige områder af et fag. Men samtidig gør opdelingen også, at resultatet for hvert profilområde er mere upræcist, end hvis man lagde profilområderne sammen, jf. Delrapport 1. En sammenlægning af profilområderne vil potentielt kunne forkorte testens længde.

Det dobbelte formål volder udfordringer

De nationale test er designet til både at være et pædagogisk redskab og et styringsredskab. Men det dobbelte formål volder udfordringer i forhold til anvendelsen af testen resultater.

Evalueringen viser klart, at man bør være meget påpasselig med at anvende en enkelt elevs resultat som enkeltstående udtryk for elevens faglige niveau. Resultatet er for usikkert til, at det kan stå alene, og den interne validitet kunne være bedre. Det vanskeliggør anvendelsen som et enkeltstående testresultat i det pædagogiske arbejde, om end VIVEs data viser, at

lærerne oftest oplever, at elevernes resultater stemmer overens med lærerens opfattelse af elevens faglige niveau.

Som pædagogisk redskab på klasseniveau er der bedre muligheder for at anvende data. Men der er uklarhed om, hvordan man omsætter den viden, som testene potentielt bidrager med, til pædagogisk praksis. Uklarheden kan både bygge på manglende viden, manglende tid, og at diskussionen om testenets validitet har fyldt så meget, som den har.

På styringsniveau og som ledelsesinformation er data dog pålidelige med høj ekstern validitet. Data bidrager særligt på kommunalt og nationalt niveau som et værdifuldt styringsredskab administrativt og i mindre grad politisk. Tilsvarende har data fra de nationale test høj værdi for den forskning, der anvender data. Den eksterne validitet er også med til at forhøje den informationsværdi, skoleledelserne kan have, for de ledere, der formår at forene deres styring med den pædagogiske praksis.

Hvad er det fremtidige behov?

Evalueringen viser klart, at der er behov for data, der kan bruges pædagogisk af lærerne i skolerne, og data, der kan bruges som styringsredskab på højere niveauer. Og VIVE vurderer, at hvis de nationale test afskaffes, så vil der være behov for at udvikle et eller flere nye redskaber til at dække disse behov, der kan indgå i samspil med andre eksisterende datakilder, så som trivselsmålinger og afgangsprøvekarakterer.

Hvis man ikke afskaffer de nationale test, er der behov for at arbejde med reliabiliteten og den interne validitet samt med at gøre det nemmere for lærere og skoleledere at arbejde konstruktivt med testene – eksempelvis gennem bedre vejledninger og mere handlingsorienteret oversættelse af data til pædagogisk anvendelse – ligesom der bør arbejdes med fortællingen om, hvad de nationale test egentlig kan og skal måle, og hvad de ikke kan og skal måle.

2.2 Tværgående resultater

Dette kapitel samler resultaterne fra de fem delrapporter. Først behandles emnerne vedrørende de nationale tests reliabilitet, interne validitet og eksterne validitet (Delrapport 1+2). Derefter behandles de nationale tests sammenhæng med fagenes Fælles Mål (Delrapport 4). Så behandles anvendelsen af de nationale test, først på skole og kommunalt niveau, og dernæst på nationalt niveau (Delrapport 5). Til sidst behandles kortlægningen af sammenlignelige test (Delrapport 3).

2.2.1 De nationale tests præcision, validitet og sammenhæng med fagene

Styrelsen for It og Læring (STIL) har gennemført en evaluering af de statistiske aspekter af de nationale test

I forbindelse med evalueringen af de nationale test, er det blevet udarbejdet en evaluering af de statistiske aspekter af de nationale test. Børne- og Undervisningsministeriet har besluttet, at STIL skal gennemføre denne evaluering.

STILs evaluering består af en validering af den tekniske beregning bag de nationale test, dvs. spørgsmål om, hvorvidt de nationale test regner rigtigt, om opgavernes sværhedsgrader stadig er korrekte og stadig passer til Rasch-modellen¹, og om det er muligt at forbedre den

¹ Læs mere om Rasch-modellen i afsnit 3.2.1

adaptive algoritme med henblik på at reducere den statistiske usikkerhed. Derudover undersøger STIL, hvorvidt målesikkerheden af elevernes færdigheder kan forbedres ved at kombinere resultater fra forskellige profilområder. Dette gøres ved at undersøge, om profilområderne måler forskellige aspekter af den samme bagvedliggende færdighed og dermed, om testresultaterne fra profilområderne kan slås sammen og dermed forbedre sikkerheden i testene.

VIVE har til opgave at reviewe evalueringen gennem nedsættelse og facilitering af en uafhængig gruppe af danske, såvel som nordiske forskere, med særlig viden om test af elever. Forskerne vurderer styrker og svagheder ved resultaterne af STILs dokumentation og analyser af de nationale tests usikkerhed, reliabilitet og øvrige måleegenskaber. Reviewerne bemærker, at STIL har gjort et stort arbejde med at dokumentere de statistiske aspekter af de nationale test, så som den statistiske sikkerhed og reliabilitet. Læs mere om de statistiske aspekter af de nationale test i Delrapport 1.

STILs evaluering af de statistiske aspekter af de nationale test er omfattende, og der er behov for uddybende forklaringer og argumentation

Reviewerne påpeger, at der er en række områder, hvor der er behov for yderligere forklaringer eller argumentation for valgene, truffet i forbindelse med både selve opbygningen af de nationale test og STILs evaluering af de tekniske aspekter. Der er ligeledes en række kritikpunkter forbundet med opbygningen af de nationale test samt konkrete forslag til forbedringer.

Opgaverne vælges på den rigtige måde, og elevdygtighederne og usikkerhederne beregnes korrekt

STIL dokumenterer, at opgaverne i de nationale test vælges på den rigtige måde og at elevdygtighederne og usikkerhederne om elevernes resultater beregnes korrekt. Det vil sige, at STIL har udelukket, at eventuelle fejl eller usikkerheder i de nationale test skyldes tekniske programmeringsfejl i beregningerne.

Målingerne er usikre på elevniveau

STIL dokumenterer, at sikkerhedsintervallerne for elevernes præstationer er brede, og at reliabiliteten er lav for nogle af testene. Det betyder ifølge reviewerne, at målesikkerheden er relativt usikker på elevniveau. Usikkerheden har særligt betydning for lærernes anvendelse af den enkelte elevs resultat, som derfor bør foretages med forbehold og ikke uden supplerende viden. Der er dog ikke noget, der tyder på, at de nationale test er ekstraordinært usikre eller mere usikre på elevniveau end andre tilsvarende test. Det bemærkes dog, at der generelt er begrænset viden om usikkerheden blandt alternative test. STIL dokumenterer, at usikkerheden er størst for de dygtigste elever. Reliabiliteten refererer til, om testen er stabil og vil give de samme resultater, hvis man gentager målingen.

93 % af alle obligatoriske testforløb i skoleåret 2017/2018 blev afsluttet med en statistisk usikkerhed under 0,55 SEM (Standard Error of Measurement), hvilket er den anvendte skæringsværdi i de nationale test. Reviewerne kritiserer STIL for ikke tilstrækkeligt at have argumenteret for, at skæringsværdien bør være 0,55 SEM, ligesom STIL ikke reflekterer over, hvad SEM bør være, når der er tale om en pædagogisk test som de nationale test. Det betyder, at det er vanskeligt at forholde sig til, om den valgte værdi er den korrekte eller ej.

Reliabiliteten er højest for dansk (læsning), matematik og engelsk, mens den for fysik/kemi ligger lavere. Man bør overveje, om den nuværende konvertering af resultaterne til en percentil-

skala (dvs. til den normbaserede skala) er formålstjenstlig, da det leder til paradoksale resultater, hvor resultaterne er mest sikre i hver sin ende af skalaen, men usikre i midten, selvom usikkerheden i de rå resultatscorer er størst for de dygtigste elever.

Ved at forlænge den tid, en test tager, vil det være muligt for eleverne at besvare flere opgaver, hvilket vil være med til at reducere den statistiske usikkerhed. Læs mere om den statistiske usikkerhed i Delrapport 1. En forlængelse af testtiden vil dog potentielt have konsekvenser for elevernes oplevelse af testsituationen, som i forvejen opleves som lang, særligt i de små klasser. For mere om oplevelsen af testsituationen læs Delrapport 5.

STIL foreslår selv at øge antallet af polytome opgaver samt at justere algoritmen i testsystemet, så opgaver med størst mulig informationsværdi vælges. "Polytome opgaver" er opgaver, hvor der er flere delspørgsmål, der tilsammen kan udtrykke om eleven har svaret rigtigt på hele opgaven eller kun dele – i modsætning til dikotome opgaver med eksempelvis ja/nej-svar. Det er dog ikke, ifølge reviewerne, entydigt, at brugen af flere polytome opgaver vil forbedre de nationale tests præcision.

Antallet af svære opgaver bør øges for at forbedre præcisionen

Der er for få svære opgaver i opgavebanken til de nationale test. STIL dokumenterer, hvor mange opgaver der er i opgavebanken, hvordan opgaver afprøves, og besvarelsene fra opgaveafprøvningsstatistik analyseres. STIL dokumenterer, at der er mangel på svære opgaver til de dygtigste elever i flere af profilområderne. Dette betyder, at eleverne ikke får den rette information om, hvor dygtige de er i de enkelte fag, da testen er upræcis. Med flere svære opgaver er det muligt at skelne mellem dygtige og meget dygtige elever, hvilket også vil medføre større sikkerhed i testene om elevdygtigheden, generelt. Der er enighed blandt reviewerne om, at antallet af svære opgaver bør øges, da det vil forbedre de nationale tests præcision.

Metoder til bestemmelse af sværhedsgrader bør undersøges nærmere

STIL finder endvidere, at der er forskel på opgavernes estimerede sværhedsgrad, når disse beregnes på baggrund af de adaptive testforløb (obligatoriske test), og når de beregnes i lineære afprøvningsforløb (opgaveafprøvning). Reviewerne efterspørger, at metoderne til bestemmelse af opgavernes sværhedsgrader bør undersøges nærmere, da der er stor forskel på opgavernes sværhedsgrad, afhængig af, om de er fra lineære eller adaptive test (som de nationale test er baseret på). Læs mere om sværhedsgraderne i Delrapport 1.

Samling af profilområderne vil øge præcisionen i målingerne

Det vil forbedre testenens målesikkerhed, hvis de nuværende tre profilområder, der findes for hver af de nationale test, bliver samlet til én skala. STIL vurderer, at elevernes resultater fra tre profilområder kan samles til ét samlet resultat med en større statistisk sikkerhed, end hvad der er tilfældet i dag.

Analyserne i VIVEs Delrapport 2 viser i forlængelse heraf, at et samlet mål for resultatet af en national test har større samvariation med karaktererne i folkeskolens 9. klasseprøver i tilsvarende fag, end de tre mål, der knytter sig til de tre profilområder enkeltvis. Det vil sige, at resultater fra de enkelte profilområder har lavere præcision med hensyn til at forudsige elevernes præstationer i 9. klasse end et samlet mål for hver national test. Et samlet mål vil derfor være mere præcist med hensyn til at afdække elevernes faglige niveau.

Det er dog afhængigt af, at det undersøges og testes, om profilområderne kan sammensættes til ét samlet mål for elevernes dygtighed inden for det enkelte fag. Reviewerne udtrykker, at der

mangler et teoretisk argument for samling af profilområderne til én skala. Læs mere om mulighederne for at øge præcisionen af resultatet ved at samle de tre profilområder i Delrapport 2 og 1. Delrapport 4 uddyber endvidere forskellen i, hvordan de enkelte test dækker fagenes mål og bredde.

Der er sammenhæng mellem resultater i de nationale test og afgangsprøverne...

Tidligere undersøgelser har vist sammenhænge mellem de nationale test og senere karakterer i 8. klasses standpunktskarakterer og folkeskolens 9. klasseprøver. STILs beregninger, jf. Delrapport 1, og nye undersøgelser gennemført af VIVE, jf. Delrapport 2, viser ligeledes, at der er samvariation mellem elevernes resultater i de nationale test og i folkeskolens afgangsprøver. De fundne korrelationer og sammenhænge i Delrapport 2 er på niveau med niveauet fra andre analyser af standardiserede test og karakterer. Det vil sige, at de nationale test har et forventeligt niveau. Læs mere om samvariationen mellem de nationale test og andre testresultater i Delrapport 2.

Resultatet indikerer, at de nationale test har en høj ekstern validitet, hvilket som sagt betyder, at de kan generaliseres til et udtryk for elevernes faglige niveau. Det vil sige, at resultatet indikerer, at de nationale test er gode til at anvende i analyser på gennemsnits- og gruppeniveau, fordi resultaterne af testene er en god stedfortræder for de enkelte elevers faglige niveau, selvom resultaterne i sagens natur ikke er identiske med det faglige niveau. Det giver de nationale test en informationsværdi, der kan anvendes på skole-, kommune- og nationalt niveau til at vurdere elevernes faglige niveau. Det gælder også, selvom de – i lighed med andre faglige test og prøver – kun måler en del af det, der undervises i. Både de nationale test og afgangsprøverne er udtryk for niveauet i de ting, der testes i, og ikke andre dele, så som alsidig udvikling eller trivsel. Der er et overlap mellem det, de nationale test og afgangsprøverne måler, om end det ikke er et fuldstændigt overlap.

... og sammenhængen stiger med øget samtidighed

Analyserne viser ligeledes, at samvariationen – altså korrelationen mellem de to resultater – er stigende med øget samtidighed. Desto tættere tidsmæssigt på hinanden den nationale test og afgangsprøven er taget, jo stærkere sammenhæng mellem de to faglige resultater. Dette understøttes også af, at de samtidige sammenligninger mellem test taget i 8. klasse og standpunktskarakterer, er ligeså stærke eller stærkere end sammenhængene imellem test taget i 8. klasse og karakterer i 9. klasse. Dette resultat er, som forventet, givet, at eleven udvikler sig, som årene går, og at der forventes mindre udvikling, jo tættere testen er på prøven.

Faglige resultater hænger ved

Elever, der opnåede lave resultater i de nationale test, opnår i gennemsnit også lave karakterer i de samme fag ved folkeskolens 9. klasseprøver. Tilsvarende gælder, at elever, der klarede sig godt i de nationale test, også i gennemsnit får højere karakterer ved afgangsprøverne end de elever, der klarede sig mindre godt. Læs mere i Delrapport 2.

Der er stærke sammenhænge i dansk (læsning) og matematik og mindre stærke sammenhænge i fysik/ kemi

Nogle fag har en lavere sammenhæng mellem resultater i de nationale test og karakterer end andre. De stærkeste sammenhænge findes i fagene dansk (læsning) og matematik for de nationale test i 8. klasse og karakterer i folkeskolens 9. klasseprøver. Det skyldes dels faget, dels at der er kort tid mellem testen og prøven.

Særligt de nationale test i fysik/kemi samvarierer i mindre grad end øvrige fag med karakter i 9. klasse i samme fag. En medvirkende forklaring kan være, at fysik/kemi er et treårigt fag fra

7.-9. klasse, hvor læreren planlægger undervisningen af pensum på tværs af alle årene, men hvor den nationale test finder sted i 8. klasse. Det vil sige, at de nationale test potentielt indeholder emner, som eleverne ikke har gennemgået på testtidspunktet. Læs mere om fysik/kemi i Delrapport 2, hvor samvariationen analyseres, og Delrapport 4, hvor sammenhængen mellem opgaverne i testen og fagets mål analyseres.

2.2.2 Sammenhængen med fagenes formål

De nationale tests tekniske kobling til fagenes Fælles Mål

Opgaverne til de nationale test udvikles af opgavekommissioner bestående af praktikere på bestilling af Styrelsen for Undervisning og Kvalitet (STUK). Opgaverne kobles til et bestemt fag inden for bestemte profilområder og sværhedsgrader. Opgaverne kobles desuden i det administrative system til Fælles Mål. Fælles Mål er organiseret i tre niveauer; i) kompetenceområder, ii) færdigheds- og vidensområder og iii) færdigheds- og vidensmål. Læs mere om udviklingen af opgaver i de nationale test i Delrapport 4.

Der er stor variation i bredden af testenes sammenhæng med Fælles Mål

Der er store og centrale dele af fagene, der ikke dækkes af de nationale test. Omfanget varierer på tværs af fag. Det er dog meningen, at testene ikke skal teste hele fag, men kun dele af faget, hvilket også er tilfældet for andre test og prøver.

Mens nogle af testene – i dansk (læsning) og engelsk – dækker de Fælles Mål relativt snævert og går i dybden med udvalgte områder, dækker de øvrige obligatoriske test – i matematik og fysik/kemi – større dele af fagenes Fælles Mål og er således mindre fokuserede.

Testene i matematik, engelsk og fysik/kemi har stor variation i dækningen af færdigheds- og vidensområderne og færdigheds- og vidensmålene i Fælles Mål. En del af forklaringen kan være, at testformatet i de nationale test (fx at de er it-baserede og multiple choice) ikke egner sig til at teste kompetencer og kun i nogen grad færdigheder. Disse dele af fagene dækkes derfor enten slet ikke eller i lav grad af testene. Det drejer sig eksempelvis om områder som 'Kommunikation' og 'Modellering'.

Der er omstændigheder, der gør, at den fundne sammenhæng undervurderes. For det første er der opgaver, som ikke i systemet er kategoriseret inden for Fælles Mål, men som hører til faget alligevel. For det andet har opgavekommissionerne, der udvikler opgaverne, kun mulighed for at koble en opgave til ét kompetenceområde, ét færdigheds- og vidensområde, ét færdigheds- og vidensmål i Fælles Mål, selvom nogle opgaver kan tilknyttes flere områder og mål, da løsningen af opgaven kræver viden fra flere af fagets områder. Læs mere om sammenhængen mellem indholdet i de nationale test og de fire obligatoriske testfags mål i Delrapport 4.

De nationale test i dansk (læsning) er en læsetest, ikke en dansktest

De nationale test i dansk adskiller sig fra de andre nationale test ved ikke at teste et fag men alene ét fagområde. Dansktesten tester således i Fælles Mål-termer udelukkende kompetenceområdet læsning – og konkret halvdelen af dette kompetenceområdes seks færdigheds- og vidensområder. De tre områder fra Fælles Mål, der dækkes, svarer 1-1 til testens tre profilområder.

Sammenhængen mellem indholdet af testen i dansk (læsning) og fagets samlede formål er derfor begrænset. Testen tester alene læsning og ikke de andre centrale dele af danskfaget.

Dette er dog et bevidst valg truffet i forbindelse med indførelsen af de nationale test. Testens sammenhæng med kompetenceområdet 'læsning' er derimod stærk. Der identificeres et potentiale i at inddrage læsehastighed i testningen af afkodning, som burde være teknisk mulig og umiddelbart vil kvalificere testen. Læs mere om sammenhængen mellem indholdet i de nationale test og faget dansk i Delrapport 4.

De nationale test i matematik rammer bredt, men er udfordret på test af færdigheder

De nationale test i matematik dækker ikke kompetenceområdet 'Matematiske kompetencer'. Det er et bevidst fravalg, da det er vurderet for svært at teste inden for de nationale tests format. Matematiktestene dækker de tre øvrige kompetenceområder, der svarer til testen's profilområder. Inden for kompetenceområderne er der dog stor variation i, i hvilken grad både færdigheds- og vidensområder samt færdigheds- og vidensmål dækkes. Dette skyldes eksempelvis, at nogle områder og mål ganske enkelt er nemmere at teste end andre. Dette gælder særligt områder og mål, som involverer færdigheder – fx det at undersøge, beskrive eller tegne.

Sammenlignet med særligt testene i dansk (læsning) og engelsk dækker matematiktestene dele af matematikfaget ganske bredt. En stor andel af fagets Fælles Mål er i en eller anden grad berørt, men der er stor forskel på, i hvilken grad områder af faget er dækket, og flere centrale færdigheder testes ikke som følge af testens format. Læs mere om sammenhængen mellem indholdet i de nationale test og faget matematik i Delrapport 4.

De nationale test i engelsk varierer på de to klassetrin

De nationale test i engelsk er knyttet til kompetenceområderne 'Skriftlig kommunikation' og 'Mundtlig kommunikation'. Kompetenceområdet 'Kultur og samfund' dækkes således ikke i testen målrettet 7. klassetrin, og dækkes kun i meget ringe grad af testen målrettet 4. klassetrin.

Inden for de dækkede kompetenceområder er der stor variation i, i hvilken grad færdigheds- og vidensområderne er dækket. I testen målrettet 7. klasse er der tale om, at færdigheds- og vidensområderne enten er dækket af mange opgaver eller ingen opgaver. Således er kun tre områder dækket med mere end én opgave. I testen målrettet 4. klasse er opgaverne lidt mere spredt. Variationen skyldes særligt testens format. Læs mere om sammenhængen mellem indholdet i de nationale test og faget engelsk i Delrapport 4.

De nationale test i fysik/kemi er præget af stor bredde, men skævhed i opgavernes fordeling

Den nationale test i fysik/kemi dækker tre af fire af fagets kompetenceområder i Fælles Mål. Inden for kompetenceområderne er der knyttet spørgsmål til hver af færdigheds- og vidensområderne. Der er altså tale om en meget bred test. Antallet af opgaver tilknyttet hvert færdigheds- og vidensområde varierer dog meget. Det kan delvist forklares af, at der ikke er nogen klar kobling mellem testens profilområder og stukturen i Fælles Mål, ligesom der ikke systematisk arbejdes med at dække alle dele af fagets Fælles Mål i udarbejdelsen af opgaver. Som i de øvrige fags tilfælde, er der områder af faget fysik/kemi, som er særligt vanskelige at teste i nationale test. Det gælder blandt andet for kompetenceområdet 'Kommunikation'. Læs mere om sammenhængen mellem indholdet i de nationale test og fysik/kemi i Delrapport 4.

Der er særligt for fysik/kemi et mismatch mellem, hvad testen tester og bredden af elevernes faglige kunnen på tidspunktet for testafviklingen. Testen tester nemlig de samlede Fælles Mål for hele udskolingen (7.-9. klassetrin), mens testen gennemføres på 8. klassetrin. Lærerne bestemmer selv, i hvilken rækkefølge de underviser i områderne i Fælles Mål, hvilket betyder, at eleverne risikerer at få testopgaver, der relaterer sig til områder af faget, som eleverne endnu

ikke er undervist i. En problematik, der også fremhæves i Delrapport 5, ligesom det kommer til udtryk i den relativt svage samvariation mellem testresultaterne i de nationale test og afgangsprøverne jf. Delrapport 2.

2.2.3 Anvendelse af de nationale test på skoler og i kommuner

Anvendelsen af de nationale test på skole-, kommune- og nationalt niveau behandles i Delrapport 5. De nationale test har til formål at fungere både som et pædagogisk redskab og et styringsredskab, jf. kapitel 3 i denne rapport. Delrapport 5 afdækker endvidere, hvorvidt nationale test anvendes i overensstemmelse med de to formål, mens Delrapport 3 kortlægger, om sammenlignelige nationale og internationale test ligeledes har to formål.

Der gennemføres flere frivillige nationale test for at følge elevernes progression

Registerdata viser, at omfanget af obligatoriske nationale test har været nogenlunde stabilt over en periode på seks skoleår. I samme periode er omfanget af gennemførte frivillige nationale test steget ganske betydeligt. Det hænger blandt andet sammen med, at kommuner og skoler ønsker at kunne følge elevernes progression, lige som der blandt forvaltninger, skoleledere og lærere er et ønske om at forberede eleverne til de obligatoriske test og gøre dem mere trygge ved testsituationen. Samtidig er der også kommet flere mulige, frivillige test, hvilket er med til at øge antallet af gennemførte frivillige test.

Testsituationen er ofte udramatisk, men de små elever oplever i særlig grad udfordringer med koncentrationen

Observationer af testgennemførelser og elevinterview viser, at eleverne oplever testsituationen meget forskelligt. Hovedparten af de interviewede elever er dog hverken særligt negative eller positive i beskrivelse af testen. I testsituation kommer frustrationer og ubehag typisk kun synligt til udtryk hos en enkelt eller få elever pr. testafvikling, mens der også er elever, der sætter pris på testen som en afveksling fra den almindelige undervisning. Der er ikke belæg i analysen for at sige, at de nationale test skaber hverken mere eller mindre ubehag eller glæde hos eleverne end andre test.

Eleverne oplever, at testen tager lang tid, og at det er svært at koncentrere sig. Særligt, når der er tale om forlængelser ud over de 45 minutter, som der er afsat til testene. I de mindste klasser observeres tegn på manglende koncentration allerede efter 10-15 minutter, hvilket dog ofte forbedres efter en pause.

Lærerne er generelt skeptiske over for nationale test som et pædagogisk redskab, men deres vurderinger har flere relevante nuancer

Lærerne er generelt skeptiske over for nationale test som et pædagogisk redskab. Mange lærere i spørgeskemaundersøgelsen er overordnet utilfredse med testens faglige indhold og kvalitet, og kun en mindre gruppe lærere oplever, at testen bidrager positivt til den pædagogiske praksis. Lærerne vurderer således generelt ikke, at de nationale test giver dem bedre indsigt i hverken enkeltelevers eller klassers faglige niveau på de områder, som eleverne bliver testet i. Lærerne stiller desuden spørgsmål ved, om resultaterne reelt udtrykker elevernes kompetencer inden for fagenes formål.

På den anden side vurderer lærerne imidlertid også, at elevernes resultater i de nationale test typisk stemmer overens med deres forventninger, og i de kvalitative interview fremstår lærernes vurderinger af de nationale test mere blandede. Flere lærere viser her på én og samme tid både en begejstring for mulighederne med de nationale test og en skepsis over for resultaterne.

Desuden er der lærere, der vurderer, at det er synd for børnene at teste dem, mens en anden gruppe lærere betragter test som en nødvendighed for deres pædagogiske praksis. Det spiller desuden en rolle for lærerne, at det pædagogiske formål med de nationale test er uklart, og de vurderer, at de nationale test primært er tænkt som et styringsredskab.

Endelig er nogle lærere kritiske over for, at nationale test er standardiserede og ikke kan tilpasses til den enkelte klasse, mens andre lærere vurderer, at standardiseringen og muligheden for at sammenligne med andre klasser giver en form for evidens og mulighed for at bekræfte deres egne vurderinger af eleverne mere bredt.

Begrænset systematik i lærernes pædagogiske anvendelse af nationale test

Der er i mindre grad fastsat klare retningslinjer for den pædagogiske anvendelse af nationale test. Lærerne er generelt i tvivl om, hvorvidt og hvordan de skal bruge nationale test i deres pædagogiske praksis. Ofte er det op til den enkelte lærer.

Lærerne bruger i begrænset omfang nationale test som grundlag for deres undervisning. Det skyldes ifølge lærerne, at nationale test ikke afspejler det, lærerne underviser i, at det kan være vanskeligt at handle på baggrund af nationale test, og at det er meget ressourcekrævende at bruge nationale test som afsæt for en faglig udvikling. Læs mere om sammenhængen mellem testene og fagenes mål i Delrapport 4.

Lærerne anvender i højere grad nationale test i dansk og matematik end i andre fag.

Lærerne anvender i højere grad nationale test i dansk og matematik sammenlignet med særligt fysik og i nogen grad engelsk. Det kan hænge sammen med, at det særligt er i de fag, at skoleledelsen følger op på elevernes resultater, og det er ofte i dansk og matematik, at der findes faglige vejledere. Det er imidlertid dansklærerne, som vurderer de nationale test i dansk som mindst anvendelige sammenlignet med nationale test i andre fag. VIVE vurderer, at deres skepsis kan hænge sammen med, at de nationale test i dansk ikke tester hele faget men alene delelementer. Læs evt. delrapport 4 for en uddybning.

Der gives begrænset mundtlig feedback til eleverne

Kun lige over halvdelen af lærerne giver deres elever individuel mundtlig feedback efter en national test. Når lærerne giver feedback, sker der med størst fokus på de elever, der har klaret sig dårligst i testen, og primært med det formål at opmuntre dem. Den individuelle feedback involverer oftest en drøftelse af elevernes oplevelse af testsituationen. Cirka halvdelen af lærerne, der giver eleverne individuel feedback, anvender feedbacksituationen til at pege på handlemuligheder for eleven.

De nationale test betragtes som et vigtigt styringsredskab på kommunalt niveau og til en vis grad på skoleniveau

Særligt forvaltningschefer og til en vis grad skoleledere oplever, at de nationale test er et vigtigt styringsredskab, som muliggør en dialog og opfølgning inden for og på tværs af kommuner og skoler. De kommunale forvaltningschefer vurderer, at nationale test styrker deres mulighed for at følge med i skolernes udvikling, og er et kvalificeret udgangspunkt for dialog med skolerne. I forlængelse heraf påpeger forvaltningscheferne i interview, at de ville erstatte nationale test med andre målinger af elevernes faglige niveau, hvis nationale test afskaffes.

Forvaltningschefernes opfattelse og anvendelse af de nationale tests er blevet styrket siden 2013. De fleste skoler indgår da også mindst én gang årligt i en dialog med forvaltningen om

resultaterne af de nationale test. Forud for dialogen har både skole og forvaltning typisk identificeret områder, hvor skolen kan forbedre sig. Dialogerne er oftest fremadskuende og fører til fremadrettede konkrete aftaler. Desuden opfattes dialogerne overvejende som tillidsbaserede.

Forvaltningschefer og til dels kommunalpolitikere finder desuden, at det er vigtigt med nationale målinger, som giver mulighed for at vurdere, om kommunens elever er særligt udfordrede i forhold til elever andre steder i landet.

Skoleledernes vurderinger af nationale test som styringsredskab er mere blandende. Stort set lige andele skoleledere er henholdsvis positive og negative over for nationale test som en kilde til ledelsesinformation om elevernes faglige niveau. Skolelederne er mest positive over for muligheden for at følge elevernes progression.

Selvom de nationale test generelt anvendes i dialogerne til at skabe overblik over, hvordan eleverne klarer sig, vurderer både skoleledere, forvaltningschefer og lokal politikere, at de nationale test ikke kan stå alene, hvorfor der inddrages en lang række andre datakilder i dialogerne på tværs af skoler, forvaltninger og politikere.

Systematiske procedurer understøtter den styringsmæssige anvendelse af nationale test på kommunalt niveau, mens det er mere svingende på skoleniveau

Mens der på kommunalt niveau er forholdsvis faste procedurer for opfølgning på nationale test i dialogen mellem forvaltning og skole, er det mere svingende, om der på skolerne er systematiske retningslinjer for opfølgning på de nationale test. De kommunale forvaltninger bruger typisk de nationale test aktivt og følger systematisk op på både tilfredsstillende og utilfredsstillende resultater fra skolerne. Det sker via systematiske "læringsamtaler" med skolerne, der typisk gennemføres en til to gange om året. Mens der på nogle skoler er tilsvarende systematiske procedurer for intern opfølgning på de nationale test, så som faste møder, er det ikke tilfældet på andre skoler. Samtidig kan der på én skole være forskel på systematikken i opfølgningen på tværs af fag.

De nationale test udgør en mindre del af skoleledernes styring og ledelse

Generelt tegner Delrapport 5 et meget blandet billede af, hvorvidt og hvordan skolelederne bruger nationale test. Over halvdelen af skolelederne bruger nationale test til at holde øje med det samlede faglige niveau på skolen og oplever, at nationale test øger deres kendskab til eleverne og styrker deres samarbejde med lærerne. Der er imidlertid også en betydelig andel skoleledere, der ikke bruger de nationale test eller bruger dem i mindre grad.

Nationale test synes primært at understøtte en tillidsbaseret dialog mellem lærere og skoleledere, men oplevelsen heraf varierer. Den nationale test synes i mindre grad at give anledning til konkrete indsatser og beslutninger på skoleniveau.

Nationale test synes i det hele taget at udgøre en mindre del af skoleledernes styrings- og ledelsesgrundlag, selv om der er relativt stor forskel på skoleledernes opfølgning på tværs af skoler. I overvejende grad bruger skolelederne nationale test til at vurdere skolens samlede, faglige progression og i mindre grad til at vurdere den enkelte elev eller klasse.

De kvalitative interview tyder på, at lærerne er mere positive over for de nationale test på de skoler, hvor skolelederne formår at koble den ledelsesmæssige dialog med lærerne til lærernes pædagogiske praksis og dermed anvende de nationale test som både et pædagogisk og et styringsmæssigt redskab.

Nationale test udgør en lille del af skolernes evalueringspraksis

Skoleledere og lærere er generelt enige om, at evaluering er et vigtigt redskab i arbejdet med at styrke den faglige kvalitet på skolerne, og de oplever, at der findes evalueringskompetencer på skolerne. Der synes imidlertid at være forskel på skolerne, hvad angår skoleledernes kendskab til og systematik i opfølgningen på test- og evalueringsresultater. Nationale test opleves generelt ikke at understøtte skolernes evalueringskultur. Undersøgelsen peger på, at det hænger sammen med, at lærere og skoleledere er skeptiske over for validiteten af nationale test. Læs mere om validiteten af de nationale test i Delrapport 1 og 2.

Kompetencerne til at tolke og omsætte resultaterne af nationale test varierer

Generelt vurderer skoleledere, at både de selv og lærerne har de nødvendige kompetencer til at tolke og anvende resultaterne fra nationale test. Der synes imidlertid at være forskel på skolerne, hvad angår skoleledernes kendskab til mulighederne for systematik i opfølgning på de nationale testresultater og øvrige evalueringsresultater.

Dertil kommer, at nogle skoleledere vurderer, at de ikke har de nødvendige fagfaglige kompetencer til at kunne understøtte den pædagogiske anvendelse af de nationale test og dermed underbygge, at testene får en faglig relevans for lærerne. Der er få lærere, der bruger de vinningsmuligheder, der findes for elevernes resultater i nationale test. Det skyldes ifølge lærerne, at det er tidskrævende og kræver særlige kompetencer, som ikke alle lærere oplever, at de har. På flere skoler spiller skolens faglige vejledere derfor en vigtig rolle i forhold til at tolke resultaterne og identificere konkrete indsatser.

Afklaring om testenenes validitet har stor betydning

Der synes at være en udfordring på skolerne i forhold til at tolke resultaterne af nationale test. En betydelig andel (mellem 41 og 59 %) lærere svarer "ved ikke" i spørgeskemaundersøgelsen på flere spørgsmål vedrørende fortolkning af resultaterne i de nationale test og deres målsikkerhed. Det indikerer manglende viden hos lærerne om de statistiske aspekter af de nationale test.

På tværs af lokalpolitikere, forvaltningschefer, skoleledere og lærere er der enighed om, at det er vigtigt at afklare, hvorvidt de nationale test er valide, hvis de skal give mening og anvendes både som styrings- og pædagogisk redskab. De vurderer, at kritikken af de nationale tests validitet udfordrer anvendelsen af nationale test som et effektivt evalueringsredskab, og at kritikken i nogle tilfælde kan udgøre en stopklods for den fortsatte anvendelse af de nationale test, både som pædagogisk og styringsmæssigt redskab. Læs mere om oplevelsen af testenenes validitet i Delrapport 5, og om testenenes validitet og reliabiliteten i Delrapport 1 og 2.

Nationale test indgår i skole-hjem-samarbejdet men sammen med andre typer af data

Næsten alle lærere orienterer forældrene skriftligt om resultaterne af deres børns resultater i de nationale test. Lærerne oplever, at forældrene kan have svært ved at forstå de skriftlige orienteringer, mens forældrene ikke vurderer, at dette er et problem i interview. Det kan dog hænge sammen med, at de interviewede forældre alle er bestyrelsesformænd og forholdsvis ressourcestærke. Lærerne oplever da også, at det særligt er de ressourcestærke forældre, der har interesse i de nationale test. En af kommentarerne fra reviewerne i Delrapport 1 var, at man bør overveje den form, man formidler elevernes resultater til forældrene i forhold til de skalaer, som resultaterne præsenteres på.

Lidt over halvdelen af lærerne drøfter også resultaterne med forældrene mundtligt, typisk i forbindelse med skole-hjem-samtaler. De interviewede forældre er overvejende glade for den information, de får fra de nationale test. Det er imidlertid karakteristisk, at nationale test udgør en lille del af datagrundlaget for skole-hjem-samarbejdet.

Tilsvarende gør sig gældende i bestyrelsesarbejdet, hvor bestyrelserne bliver orienteret om elevernes resultater i nationale test, men i mindre grad handler og diskuterer resultaterne. I det omfang det sker, bygger dialogen og initiativerne også på andre datakilder.

2.2.4 Anvendelse af nationale test på nationalt niveau

På nationalt niveau anvendes de nationale test i høj grad administrativt

Den administrative anvendelse af de nationale test er ganske betydelig i forhold til at understøtte styring og ledelse på tværs af niveauer. Data fra de nationale test anvendes til at kvalificere Børne- og Undervisningsministeriets vejledningsindsatser over for skoler og kommuner og som et udvælgelseskriterie for tilsyn med folkeskolens faglige kvalitet. Her giver de nationale test som faglig indikator mulighed for at følge elevkohorters faglige progression over en årække og identificere faglige udsving allerede i indskolingen og på mellemtrinnet.

Derudover bruges nationale test indirekte som politisk beslutningsgrundlag som en del af ministeriets statusredegørelser til Folketinget om folkeskolens generelle udvikling, og de indgår som en del af det vidensgrundlag, der videresendes til ministeren og forligskredsen, og danner afsæt for bl.a. policy-udvikling, følgeforskningen til folkeskolereformen samt rekvirerede analyser og forskning.

Politisk anvendes de nationale test i væsentlig grad indirekte

Den politiske værdi af de nationale test vurderes meget forskelligt fra folketingsmedlem til folketingsmedlem. Variationen spænder fra ingen værdi til stor værdi.

Den direkte politiske anvendelse på nationalt niveau af data fra de nationale test synes meget begrænset. Data forelægges kun forligskredsen bag folkeskolereformen gennem den årlige statusredegørelse for folkeskolens udvikling. Udviklingen i elevernes resultater præsenteres for politikerne på et aggregeret og ikke detaljeret niveau, og ifølge de interviewede politikere er det derfor vanskeligt at træffe beslutninger alene på baggrund af resultaterne.

Den indirekte politiske anvendelse af de nationale test gennem fx forskning er langt større. Særligt er Folketingsmedlemmerne optagede af følgeforskningen til folkeskolereformen, som blandt andet baserer sig på data fra de nationale test. En stor del af den forskningsviden, som Folketinget har på skoleområdet, indeholder data fra nationale test.

Data fra de nationale test er værdifulde i forskning

Forskere, der anvender data fra de nationale test, oplever, at data fra de nationale test har stor forskningsmæssig anvendelighed og er værdifulde for deres forskning. Langt hovedparten af forskerne er desuden helt eller overvejende enige i, at adgang til data fra de nationale test er værdifuld for, at deres forskning kan bidrage til at forbedre praksis.

Data anvendes til samfundsvidenskabelig forskning og primært i undersøgelser af sammenhængen mellem indsatser og elevers faglige resultater eller undersøgelser, der bidrager med viden om, hvad der øger elevernes læring og faglige niveau. Data fra de nationale test i dansk (læsning) og matematik anvendes mest.

Forskerne ser overvejende ingen alternative datakilder af samme kvalitet, og mere end to ud af tre finder adgang til data fra de nationale test afgørende for, at de kan besvare deres forskningsspørgsmål.

2.2.5 Sammenlignelige test

VIVE har kortlagt 106 sammenlignelige test fra Danmark og internationalt med det formål at kortlægge, hvilke test findes der i formål, indhold og omfang minder om de danske nationale test. Det gøres for at skabe overblik over lignende test og dermed skabe mulighed for at blive inspireret til at indhente yderligere viden om konkrete test. Kortlægningen viser endvidere, hvordan testlandskabet ser ud, og hvilke typer af test der bliver anvendt. Testene er ikke nødvendigvis direkte alternativer til de danske nationale test, men kan bruges til at søge inspiration om form og erfaringer i. For mere om sammenlignelige test læs Delrapport 3.

Testene er oftest fra USA og tester oftest indskolingen

Ud af de 106 test er 56 fra USA, mens 14 test er danske. Desuden findes 13 test i resten af Norden, 13 test i Europa eksklusive Norden, 7 i de resterende undersøgte lande, samt 3 internationale test.

90 % af testene målretter sig indskolingen i USA og Europa. De kan dog, ligesom de danske nationale test, godt være udviklet til at dække flere klassetrin. I USA ses eksempelvis, at de fleste test er målrettet hele skolegangen.

Formålet er typisk pædagogisk

Testene er typisk tiltænkt som pædagogiske redskaber, men uden for Europa ses oftere test med styring som formål, dog i ca. 50 % af tilfældene i en kombination af et styringsredskab og et pædagogisk værktøj ligesom de danske nationale test.

De danske nationale tests målgruppe og modtager minder om andre test med lignende anvendelsesformål

De danske nationale test er karakteriseret ved at være tiltænkt som både et pædagogisk værktøj og et styringsredskab. De er målrettet årgange på tværs af indskolingen, mellemtrinnet og udskolingen. Dette er også tilfældet for de øvrige test, som har det dobbelte formål.

Tilsvarende målrettes resultaterne af testen typisk både hjemmet, læreren og myndighed, præcis som det også forekommer i de danske nationale test.

Varigheden af de danske nationale test er kortere sammenlignet med andre test med lignende anvendelsesformål

Er en test kategoriseret som et pædagogisk værktøj, er den typisk målrettet indskolingen eller hele skolegangen, og testen vil maksimalt tage 45 minutter.

Anvendes testen derimod alene som et styringsredskab og ikke et pædagogisk værktøj, er den målrettet mellemtrinnet og udskolingen eller hele skolegangen. Her vil testen typisk tage mere end 45 minutter.

Er en test en kombination af begge anvendelsesformål er den typisk designet til hele skolegangen, og den vil have en tendens til at vare mere end 45 minutter. De danske nationale test varer typisk en lektion, dvs. 45 minutter, hvilket er kortere sammenlignet med øvrige test med samme anvendelsesformål.

Adaptive test er ikke udpræget i Norden

Brugen af adaptive test er meget begrænset, og generelt findes der ingen adaptive test i Norden ud over de danske nationale test. Denne testform er altså usædvanlig, og man må forvente mindre eller ingen erfaring med denne type test i landene, herunder Danmark.

Test er oftest manuelt bedømte

De danske test minder mest om test i USA, hvor cirka halvdelen af testene bedømmes manuelt, cirka en tredjedel er selvscorende i testsystemet, mens resten bedømmes både manuelt og automatisk i testsystemet. Sammenlignet med resten af Norden er selvscorende bedømmelse betydeligt mere udbredt i Danmark.

Information om tests reliabilitet er svær at indhente og sammenligne

Blandt et udsnit på 11 af de kortlagte test er der søgt efter information om reliabilitet. Det er dels sparsomt med information både inden for og på tværs af testene, dels er der ikke konsistens i valget af reliabilitetsmål på tværs af test. Derudover er de enkelte mål meget kontekstafhængigt. Værdierne af Standard Error of Measurement (SEM) afhænger af den underliggende model, der anvendes til at score en given test. Det er med til at vanskeliggøre sammenligning af SEM på tværs af test. Ligeledes vil en test-retest kunne foretages på mange forskellige samples, som i større eller mindre omfang vil have betydning for korrelationen. Det vil sige, at det ikke uden en væsentligt dybere analyse er muligt at sammenligne de danske nationale tests reliabilitet med de fundne tests reliabilitet. Ud fra den information, der er indhentet på de 11 test, er der ikke noget, der tyder på, at de danske nationale test har en markant dårligere eller bedre reliabilitet end andre test.

3 De nationale test

Dette kapitel beskriver kortfattet baggrunden for og indholdet af de nationale test. Formålet er at give læseren tilstrækkelig viden om de fundamentale aspekter af de nationale test til at kunne læse de analytiske kapitler, der vedrører VIVEs evaluering af de nationale test.

Undervisningsministeriet har selv ad flere omgange beskrevet dette, og kapitlet vil i væsentlig grad bestå af en gengivelse af disse beskrivelser (Børne- og Undervisningsministeriet, 2019a). Først beskrives baggrunden og det erklærede formål med de nationale test, og efterfølgende beskrives de tekniske aspekter af testene.

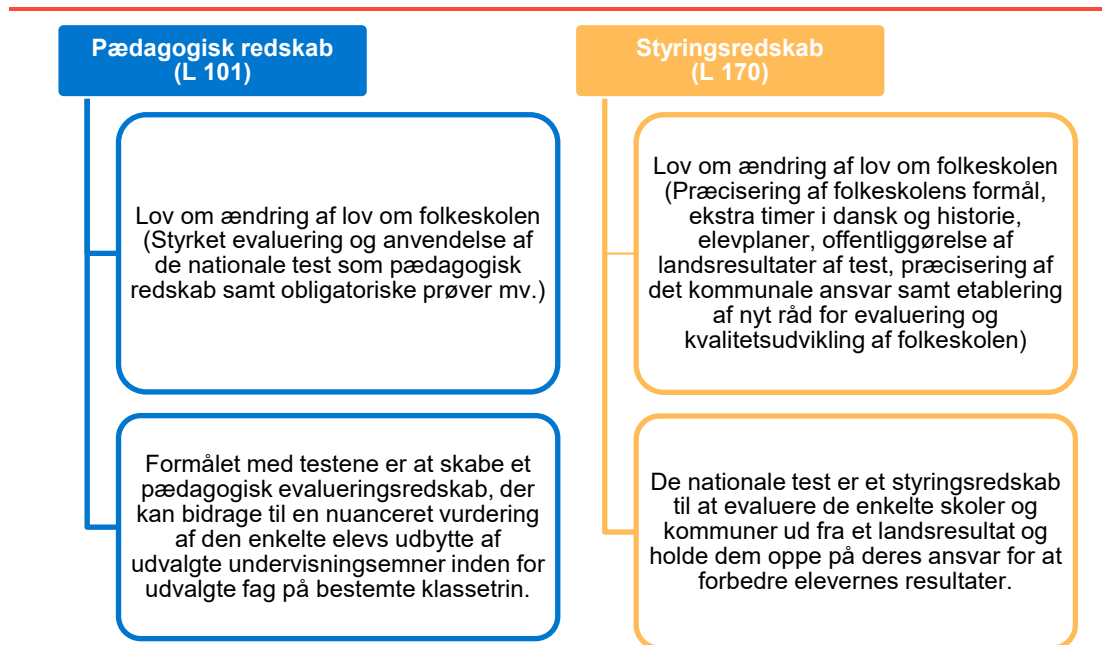
3.1 Baggrund

Indførelsen af de nationale test blev begrundet med henvisning til undersøgelser, der viste, at det faglige niveau blandt eleverne i den danske folkeskole var utilstrækkelig (EVA, 2004a; 2004b). Tilsvarende havde en OECD-rapport om grundskolen fra 2004 (OECD, 2004) påpeget, at den danske evalueringskultur var utilfredsstillende, og at der derfor var behov for at styrke arbejdet med at følge elevernes faglige resultater systematisk.

Testene er skabt som både et pædagogisk redskab og et styringsredskab

Formålet med de nationale test er todelt. De er skabt til at være såvel et pædagogisk redskab som et styringsredskab.

Figur 3.1 Testenes dobbelte formål



Kilde: Undervisningsministeriet (2005; 2006)

Formålet med indførelsen var, at man gennem øget brug af viden om elevernes faglige niveau kunne handle med rettidig omhu fra såvel lærerne, skolelederne, de kommunale forvaltninger og på nationalt niveau (Børne- og Undervisningsministeriet, 2019a). Målet var, at man på den

måde bedre kunne sætte ind i tide i forhold såvel til den enkelte elevs behov og tilrettelægge undervisningen for et bedre fagligt udbytte som klassens eller skolens behov. I 2006 vedtog Folketinget indførelsen af de nationale test med udgangspunkt i "Lov om ændring af lov om folkeskolen (styrket evaluering og anvendelse af de nationale test som pædagogisk redskab samt obligatoriske prøver mv.)". De nationale test var ét blandt flere elementer i et lovforslag fra december 2005 om fornyelse af folkeskolen for at forbedre det faglige niveau blandt eleverne gennem styrket, løbende evaluering i folkeskolen.

Den første obligatoriske testrunde blev gennemført i foråret 2010. Resultaterne fra de nationale test anvendes blandt andet til at følge den faglige udvikling på landsplan. Der er senest med folkeskolereformen i 2014 formuleret tre nationale mål for udvikling af folkeskolen, der skal danne afsæt for alle initiativer i folkeskolen og fungere som målestok for opfølgningen på, hvorvidt kommuners og skolars tiltag til udvikling af folkeskolen er lykkedes.

Nationale mål for folkeskolens udvikling

- Folkeskolen skal udfordre alle elever, så de bliver så dygtige, de kan.
- Folkeskolen skal mindske betydningen af social baggrund i forhold til faglige resultater.
- Tilliden til og trivslen i folkeskolen skal styrkes blandt andet gennem respekt for professionel viden og praksis.

De nationale test spiller en central rolle i forhold til at vurdere, hvorvidt de to første af de tre mål realiseres. De tre mål er operationaliseret i fire måltal, som danner afsæt for opfølgning på, hvorvidt målene indfris. I den sammenhæng spiller de nationale test en afgørende rolle, idet de indgår som datagrundlag for vurdering af, hvorvidt kommuner og skoler realiserer de tre første af måltallene.

Nationale måltal for folkeskolens udvikling

1. Mindst 80 % af eleverne skal være gode til at læse og regne i nationale test.
2. Andelen af de allerdygtigste elever i dansk og matematik skal stige år for år.
3. Andelen af elever med dårlige resultater i nationale test for læsning og matematik skal reduceres år for år.
4. Elevers trivsel skal øges.

3.2 Testenes indhold

Ti obligatoriske test og op til 32 frivillige

Der er ti obligatoriske nationale test i folkeskolen fordelt på fire fag. De fire fag er dansk (læsning)², matematik, engelsk og fysik/kemi. Alle elever, der undervises efter folkeskoleloven, skal gennemføre de obligatoriske test i udvalgte fag og på bestemte klassetrin. Elever kan i særlige tilfælde fritages. I de obligatoriske testfag er det muligt at tage samme test frivilligt før og efter den obligatoriske testrunde.

² I faget dansk testes alene i kompetenceområdet læsning.

Derudover er der i tre fag test, der kan tages på frivillig basis. De tre fag er geografi, biologi og dansk som andetsprog. Alle grundskoler (folkeskoler, specialskoler, privatskoler, friskoler og efterskoler mv.) kan gennemføre de nationale test på frivillig basis.

Testene kan gennemføres i to perioder: En i efteråret og en i foråret. I efteråret kan de frivillige test gennemføres, mens der både kan gennemføres frivillige og obligatoriske test i perioden fra d. 1. marts til og med den 30. april. Nedenstående oversigt viser testene, samt hvorvidt de er obligatoriske eller alene kan tages som frivillige test.

Tabel 3.1 Oversigt over obligatoriske og frivillige nationale test

Fag og klassetrin	1.	2.	3.	4.	5.	6.	7.	8.	9.
Dansk (læsning)									
Matematik									
Engelsk									
Fysik/kemi									
Biologi									
Geografi									
Dansk som andetsprog									

Frivillige test målrettet klassetrinet over eller under
 Obligatoriske test målrettet klassetrinet
 Frivillige test målrettet klassetrinet

3.2.1 Testenes form

De nationale test har følgende otte grundlæggende karakteristika:

Karakteristika	Uddybende beskrivelse
Hver test består af tre faglige profilområder	De nationale test tester et afgrænset område af fagene, og i hvert fag testes eleverne inden for tre faglige profilområder. Profilområderne er ens på tværs af klassetrin på nær engelsk, hvor der er forskel i 4. (lytning) og 7. (sprog og sprogbrug) klassetrin. Hver opgave i testen er tilknyttet ét af de tre profilområder.
De er it-baserede	Testene gennemføres på computer eller tablet online.
De er baseret på Rasch-modellen	Rasch-modellen er en statistisk model, hvor elevens dygtighed og opgavens sværhedsgrad måles på én og samme skala. Elevens dygtighed er defineret som sværhedsgraden på den opgave, hvor eleven har lige så stor sandsynlighed for at svare rigtigt og forkert. Alle opgaver i testen lever op til modellens krav. I testen er det ikke antallet af korrekte svar, der er af betydning for, hvor dygtig testen vurderer eleven til at være. Det er derimod sværhedsgraden af de opgaver, som eleven besvarer, der har betydning for, hvor dygtig testen vurderer eleven til at være.
De er adaptive	Det adaptive princip betyder, at testen tilpasser opgavernes sværhedsgrad til eleven i et forsøg på at tilpasse sig elevens faglige niveau. Testen individualiseres således, så eleverne ikke modtager de samme opgaver. Når der eksempelvis svares forkert på en opgave, vil den næste opgave have en lavere sværhedsgrad, og omvendt hvis man svarer rigtigt. Det betyder, at de svageste elever også får opgaver, som de kan svare på, og at de dygtigste elever også får opgaver, de ikke kan svare på. Tilpasningen sker på baggrund af en bagvedliggende algoritme, der trækker opgaver, hvis sværhedsgrad er baseret på elevens estimerede dygtighed.

Karakteristika	Uddybende beskrivelse
De er selvscorende	Programmet bag de nationale test beregner resultatet for den enkelte elev. Det er således ikke lærerne, der bedømmer testbesvarelsenerne.
Der gives en tilbagemelding pr. profilområde samt en samlet vurdering	Eleven får en samlet vurdering af præstationen. Eleven kan også få en tilbagemelding pr. profilområde. Elevens lærer har mulighed for at se, hvordan eleven klarer sig overordnet pr. profilområde samt på de enkelte opgaver og herunder tidsforbruget.
En test tager én lektion at gennemføre (45 min.).	Testen tager minimum 45 min. De fleste elever gennemfører på denne tid. I de tilfælde, at programmet ikke har fundet en tilfredsstillende vurdering i løbet af 45 minutter, har læreren mulighed for at forlænge den enkelte elevs test med 15 min ad gangen. Der er også mulighed for at afslutte eller udsætte testen.
Resultaterne er ikke offentlige	Resultaterne af de nationale test må ikke offentliggøres. Kommunerne må offentliggøre, om kommunen eller de enkelte skoler lever op til de nationale resultatmål, og om kommunen eller de enkelte skoler har forbedret sig i forhold til sidste skoleår.

3.2.2 Test- og prøvesystemet

Testene er digitalt-selvrettende og gennemføres online. Testene bookes og gennemføres i test- og prøvesystemet på testogprøver.dk. Skolerne kan gennemføre test samme dag, som de booker. I samme system genereres og tilgås testresultaterne fra testene.

Undervisningsministeriet har udarbejdet vejledninger om de nationale test til forældre, lærere i alle fag, skoleledere, kommuner og lærere, som underviser elever med særlige behov og dansk som andet sprog, samt vejledninger i resultatvisningerne af de nationale test (Undervisningsministeriet, 2017; 2018a; 2018b; 2018c; 2018d; 2018e; 2018f).

Adgangen til testopgaver og resultater varierer med aktøren

Lærere, skoleledere og kommuner har adgang til testresultater på forskellige niveauer (hhv. elevniveau, klasseniveau og skoleniveau). Resultaterne af testene er fortrolige og må alene anvendes internt, når der foreligger en saglig grund. Børne- og Undervisningsministeriets vejledninger til de enkelte aktører uddyber, hvilke aktører der må udveksle testresultater med hvem. Anonymiserede testresultater må heller ikke offentliggøres.

Læreren skal som minimum give eleven og forældrene en skriftlig tilbagemelding om testresultatet af de obligatoriske test. Det er ikke fastlagt, hvornår tilbagemeldingen skal foregå, og det er således ikke fastlagt, at den skriftlige tilbagemelding skal se umiddelbart efter testgennemførelsen. Børne- og Undervisningsministeriet har udarbejdet to tilbagemeldingsskabeloner til fri afbenyttelse. Det er dog frivilligt, om man vil bruge disse eller udarbejde egne skriftlige tilbagemeldinger.

Fortroligheden gælder også testopgaver. Det er dermed ikke lovligt at offentliggøre testopgaver, hverken mundtligt eller skriftligt. Dette gælder også i tilbagemeldingen til elever og forældre. Det betyder, at man eksempelvis som lærer ikke må vise en opgave fra testen på klassen efter testgennemførelse. Det skyldes, at testopgaverne bruges igen. Hvis testopgaverne bliver kendt, for eksempel fordi de står i lærebøger, ændres deres sværhedsgrad (de bliver lettere), og de får mindre værdi i testen.

3.2.3 Tilbagemeldingsformer

Testresultaterne kan vises på to forskellige måder: normbaseret og kriteriebaseret. De to tilbagemeldingsformer kan ikke umiddelbart sammenlignes, da de giver forskellig information om eleverne (Børne- og Undervisningsministeriet, 2018c). Lærerne kan i dansk og matematik

vælge, om de anvender den normbaserede eller den kriteriebaserede i formidlingen til forældrene. I de øvrige fag skal den normbaserede anvendes.

Den normbaserede tilbagemelding

Den normbaserede tilbagemelding viser, hvordan en elev og en klasse har klaret sig i testen sammenlignet med første gang, der blev gennemført nationale test på landsplan. Det var i 2010 for de fleste fag og profilområder. Elevernes resultater vises på en skala fra 1-100. Hvis en elev for eksempel har fået resultatet 75 på denne skala, er elevens resultat bedre end eller lige så godt som 75 % af elevernes resultater i samme fag og test fra 2010. Den normbaserede tilbagemelding kan anvendes i alle fag.

Den normbaserede skala viser ikke, hvor eleverne er i forhold til et ønsket fagligt niveau (Børne- og Undervisningsministeriet, 2018c). 1-100 skalaen vises kun til lærerne og eventuelt skolelederen. Det normbaserede resultat på 100-skalaen omsættes til fem kategorier, i formidlingen til elever og forældre (Børne- og Undervisningsministeriet, 2019b).

- En del over gennemsnittet (91-100)
- Over gennemsnittet (66-90)
- Gennemsnittet (36-65)
- Under gennemsnittet (11-35)
- En del under gennemsnittet (1-10).

Den kriteriebaserede tilbagemelding

Den kriteriebaserede tilbagemelding er alene for dansk (læsning) og matematik. Resultatet vises som et udtryk for elevernes faglige niveau i de dele af fagene, som testes. Skalaen er udarbejdet i samarbejde med ministeriets opgavekommissioner. De har udvalgt repræsentative opgaver og vurderet, hvor mange og hvilke af disse opgaver en elev bør kunne besvare rigtigt på hvert trin på skalaen.

Det kriteriebaserede resultatet anvendes i formidling til lærerne, hvor skalaen består af seks kategorier:

1. Fremragende præstation
2. Rigtig god præstation
3. God præstation
4. Jævn præstation
5. Mangelfuld præstation
6. Ikke tilstrækkelig præstation.

3.2.4 Udvikling af testopgaver

Kontor for Prøver, Eksamen og Test i Styrelsen for Undervisning og Kvalitet (STUK) har det overordnede ansvar for de nationale test. STUK beskikker eksterne fagpersoner til deltagelse i opgavekommissioner, som er det producerende organ i udviklingen og revisionen af opgaver til de nationale test. Hvert fag har en opgavekommission, som består af minimum to personer med en formand. Formanden for opgavekommissionen har det overordnede ansvar for kvalitetssikringen og processen. Deltagerne i opgavekommissionerne producerer testopgaver inden for fagenes tilknyttede profilområder og vælger inden for, hvilke færdigheds- og videnom-

råder de falder. De vælger ligeledes, hvilken sværhedsgrad og opgavetype de producerer opgaver inden for. Når medlemmerne er færdige med at udvikle et antal testopgaver, har andre tilknyttede personer, der skal kvalitetssikre, efterfølgende ansvar for at gennemgå hver enkelt opgave for at sikre, at de lever op til kvalitetsfaktorerne.

Læringskonsulenterne i STUK har det overordnede ansvar for den faglige godkendelse af opgaverne. Når de er godkendt, bliver de sendt til Kontor for Prøver, Eksamen og Test i STUK. Derefter afprøver Styrelsen for It og Læring (STIL) opgaven og foretager en analyse, bygget på Rasch-modellen. Alle opgaverne er afprøvet på ca. 700 elever. Rasch-analysen er en statistisk analyse, der oprindeligt er udviklet til læseprøver og intelligenstag. I Rasch-modellen bliver elevens dygtighed og opgavens sværhedsgrad målt på én og samme skala. Elevens resultat (dygtighed) er defineret som sværhedsgraden på den opgave, hvor eleven har lige stor sandsynlighed for at svare rigtigt og forkert. Selve testen handler altså om at finde ud af, hvor svære opgaverne skal være, for at eleven har 50 % sandsynlighed for at svare rigtigt. Lidt forenklet kan man sige, at jo sværere en opgave er, desto færre elever kan svare på den – og omvendt. Man finder opgavernes sværhedsgrad ved at afprøve opgaverne empirisk. Rasch-modellens styrke er, at den testmetodisk sikrer test, hvor resultatet er et godt udtryk for elevernes samlede resultat, mens en ulempe er, at opgaverne tester et forholdsvist snævert fagligt område (Børne- og Undervisningsministeriet, 2018e).

3.2.5 Tidligere evaluering

Den seneste evaluering af de nationale test blev gennemført i 2013 (Rambøll, 2013). Det blev i denne sammenhæng besluttet, at der skulle igangsættes en ny evaluering efter en femårig periode. Derfor igangsatte Undervisningsministeriet i efteråret 2018 en evaluering af de nationale test. VIVE blev i foråret 2019 kontaktet med henblik på udarbejdelse af evalueringen.

Litteratur

- Undervisningsministeriet (2017). *Vejledning til nye resultatvisninger i de nationale test – til lærere i alle fag*. København: Undervisningsministeriet.
- Undervisningsministeriet (2018a). *Vejledning om de nationale test - til kommuner*. København: Undervisningsministeriet.
- Undervisningsministeriet (2018b). *Vejledning om de nationale test - til lærere, som underviser elever med særlige behov*. København: Undervisningsministeriet.
- Undervisningsministeriet (2018c). *Vejledning om de nationale test - til skoleledere*. København: Undervisningsministeriet.
- Undervisningsministeriet (2018d). *Vejledning til de frivillige nationale test i dansk som andet-sprog*. København: Undervisningsministeriet.
- Undervisningsministeriet (2018e). *Nationale test – information til forældre*. København: Undervisningsministeriet.
- Undervisningsministeriet (2018f). *Vejledning om de nationale test - til lærere i alle fag*. København: Undervisningsministeriet.
- Børne- og Undervisningsministeriet (2019a). *Baggrundsnotat*. København: Undervisningsministeriet.
- Børne- og Undervisningsministeriet (2019b). *Standardbrev til formidling af resultater*. København: Undervisningsministeriet.
- EVA (2004a). *Undervisningsdifferentiering i folkeskolen*. København: Danmarks Evalueringsinstitut – EVA.
- EVA (2004b). *Løbende evaluering af elevernes udbytte af undervisningen i folkeskolen*. København: Danmarks Evalueringsinstitut – EVA.
- OECD (2004). *OECD-rapport om grundskolen i Danmark - 2004*. København: Undervisningsministeriet.
- Rambøll (2013). *Evaluering af de nationale test i folkeskolen*. København: Rambøll.
- Undervisningsministeriet (2005). *Lov om ændring af lov om folkeskolen L101*. København: Undervisningsministeriet.
- Undervisningsministeriet (2006). *Lov om ændring af lov om folkeskolen L170*. København: Undervisningsministeriet.



DET NATIONALE FORSKNINGS-
OG ANALYSECENTER FOR VELFÆRD