Rådgivningsgruppen for evaluering af de nationale test

Anbefalinger, januar 2020

Indhold

1 Indledning

1.1 Baggrund

I 2013 evaluerede Rambøll de nationale test for Børne- og Undervisningsministeriet[1]. Evalueringen havde fokus på testenes indhold og effekt og gav viden om skolernes anvendelse af testene og om kvaliteten af testene i de enkelte fag.

Evalueringen blev gennemført tre år efter indførelsen af de nationale test i 2010. Den daværende undervisningsminister indstillede i sin redegørelse til Folketinget fra december 2013[2], at der skulle gennemføres en ny evaluering 3-5 år efter.

Børne- og Undervisningsministeriet igangsatte derfor i 2018 en ny evaluering af de nationale test med henblik på at skabe et solidt og kvalificeret grundlag for at tage stilling til den fremadrettede udvikling og brug af testene.

Samtidig blev der nedsat en rådgivningsgruppe af forskere, lærere, skolelever, skoleledere, forældrerepræsentanter og repræsentanter for kommunalforvaltningen. I efteråret 2018 afholdte gruppen tre møder med henblik på at udarbejde anbefalinger til evalueringstemaer. Gruppen modtog blandt andet input fra faglige oplæg og to dialogmøder med lærere, ledere og forvaltninger.

I november 2018 indstillede rådgivningsgruppen en række spørgsmål, som evalueringen burde besvare. Spørgsmålene samler sig inden for følgende temaer:

- Validitet, usikkerhed og reliabilitet.
- Testenes egenskaber i relation til undervisningen og med perspektiv til internationale erfaringer.
- Lærernes, skoleledernes og kommunalpolitikernes anvendelse af de nationale test.
- Eleverne i testsituationen og brugbarheden af testresultaterne.
- Forældrenes forståelse af de nationale test.

Med udgangspunkt i disse spørgsmål udarbejdede Børne- og Undervisningsministeriet en konkret opgavebeskrivelse for evalueringen. Børne- og Undervisningsministeriet besluttede at udføre evalueringen i to dele:

- En teknisk analyse af testenes statistiske egenskaber og den adaptive algoritmes funktion.
- En undersøgelse af brugen og betydningen af testene.

Styrelsen for IT og Læring (STIL) har udført første delopgave, mens Det Nationale Forsknings- og Analysecenter for Velfærd (VIVE) har udført anden del. VIVE har desuden været ansvarlig for et fagligt review af STIL's tekniske analyser.

1 https://dk.ramboll.com/~/media/Files/RM/Rapporter/Nationale%20test%202013.pdf
2 https://www.ft.dk/samling/20131/almdel/BUU/bilag/71/1312728.pdf

1.2 Rådgivningsgruppens kommissorium og arbejde

Rådgivningsgruppens arbejde har taget udgangspunkt i et kommissorium udarbejdet af Børne- og Undervisningsministeriet[3].

Kommissoriet giver gruppen følgende formål:

Rådgivningsgruppen skal rådgive ministeriet om opgaveløsningen med evalueringen, og gruppen skal have fokus på evalueringens formål, evalueringstematikker og bidrage til evalueringsspørgsmål samt efterfølgende bidrage til at kvalitetssikre evalueringens resultater. Gruppen skal på baggrund af evalueringen rådgive ministeriet om det videre arbejde med nationale test og evaluering i folkeskolen.

Rådgivningsgruppen har i efteråret 2018 udarbejdet de spørgsmål, der har været udgangspunkt for evalueringens opgavebeskrivelse.

På tre møder i januar 2020 har gruppens medlemmer drøftet VIVE's og STIL's afrapporteringer samt diskuteret anbefalinger til det fremadrettede arbejde med nationale test. Rådgivningsgruppens medlemmer har haft mulighed for at oversende deres kommentarer om evalueringen til VIVE. Gruppen har ikke som helhed taget stilling til evalueringens kvalitet eller godkendt evalueringen. Gruppen har lagt hovedvægten på at diskutere anbefalinger til fremtiden.

3 https://www.uvm.dk/-/media/filer/uvm/adm/pdf18/jul/180821-kommissorium-for-raadgivningsgruppen-for-evaluering-af-de-nationale-test.pdf?la=dat

1.3 Rådgivningsgruppens sammensætning

Det fremgår af kommissoriet, at rådgivningsgruppen er sammensat med henblik på at repræsentere forskellige holdninger, kompetencer og erfaringer. Medlemmerne er:

Professor Peter Dahler-Larsen (formand), Københavns Universitet

Professor Jeppe Bundsgaard, Aarhus Universitet

Professor emeritus Peter Allerup, Aarhus Universitet

Professor emeritus Svend Kreiner, Københavns Universitet

Professor Simon Calmar Andersen, Aarhus Universitet

Professor Lotte Bøgh-Andersen, Aarhus Universitet

Lektor Lena Lindenskov, Aarhus Universitet

Professor Rolf Vegar Olsen, Oslo Universitet

Udviklingschef Niels Rasmus Ziska, Hogrefe Psykologisk Forlag

Forældrerepræsentant René Bomholt – erstattet af Rasmus Edelberg, Skole og Forældre

Skoleelev Sonja Agerbæk-Pedersen – erstattet af Leanor Dall/ Niklas Vistesen, Danske Skoleelever

Skoleleder Thomas Dandanell Nielsen, Nørre Alslev Skole, Guldborgsund Kommune

Lærer Stine Hertz, Skovshoved Skole, Gentofte Kommune

Lærer Signe Tofft, Dyssegaardsskolen, Gentofte Kommune

Direktør Jakob Ryttersgaard, Skoleforvaltningen, Aalborg Kommune

Lærer og kommunal konsulent Rikke Kjærup, Ballerup Kommune – udtrådt af gruppen i januar 2020.

Kirsten Balle, centerchef for Udvikling og Pædagogik i Lyngby-Taarbæk Kommune – udtrådt af gruppen i januar 2020.

Undervejs i arbejdet har rådgivningsgruppens undergået følgende ændringer: Rasmus Edelberg har afløst René Bomholt. Leanor Dall og Niklas Vistesen har i fællesskab afløst Sonja Agerbæk-Pedesen. Rikke Kjærup og Kirsten Balle er trådt ud af gruppen pga. ændrede ansættelsesforhold.

2 Anbefalinger

Kommissoriet giver gruppen denne opgave:

At modtage evalueringens resultater og bidrage med anbefalinger til ministeren i den efterfølgende opfølgning og beslutningsproces.

Som grundlag for rådgivningsgruppens drøftelse har medlemmerne modtaget et fortroligt udkast til både VIVE's og STIL's afrapportering. Det offentliggjorte evalueringsmateriale er i alt væsentligt identisk med det udkast, gruppen har modtaget.

Kommissoriet rummer ikke et pålæg til rådgivningsgruppen om at forholde sig til omkostninger og tidsforbrug knyttet til forskellige beslutninger om de nationale test. Rådgivningsgruppen har i bred forstand drøftet, at nogle anbefalinger kan koste tid og penge, ligesom der kan være besparelser i tid og penge knyttet til andre anbefalinger.

Ifølge kommissoriet er rådgivningsgruppen sammensat med henblik på at repræsentere forskellige positioner, kompetencer og erfaringer.

Dette har rådgivningsgruppen fortolket således, at indstillingen fra rådgivningsgruppen skal give plads til både enighed og meningsforskelle. Ligeledes har det været rådgivningsgruppens ønske at afspejle de argumenter for og imod en given anbefaling, som har været særligt afgørende for de enkelte medlemmers stillingtagen, for på den måde at give eksempler på de afvejninger, der ofte må præge en efterfølgende beslutning.

Følgende spilleregler har været gældende i rådgivningsgruppens arbejde: Hvert enkelt medlem har haft mulighed for at fremsætte forslag til anbefalinger. Alle modtagne forslag er blevet drøftet. Ethvert forslag til anbefalinger er taget med i indstillingen, medmindre det med forslagsstillerens accept er trukket tilbage eller skrevet sammen med et andet forslag.

Det enkelte medlem har haft mulighed for at støtte anbefalingen eller ikke støtte anbefalingen. Det har ligeledes været muligt ikke at tage stilling til den enkelte anbefaling. Begrundelser herfor kan eksempelvis være, at man ikke har følt sig fagligt rustet på det specifikke område, som anbefalingen omhandler, at man kan lade anbefalingen passere med de argumenter for og imod, der er opstillet, eller at man har fundet, at stillingtagen til anbefalingen i sidste instans bør være en politisk afgørelse. I lyset af medlemmernes stillingtagen og af rådgivningsgruppens drøftelser er anbefalingerne blevet justeret løbende, og der er givet mulighed for ny stillingtagen.

En anbefaling er afslutningsvis placeret under overskriften "Rådgivningsgruppen anbefaler", hvis ingen medlemmer har ønsket ikke at tilslutte sig.

En anbefaling er afslutningsvis placeret under overskriften "Der er bred tilslutning i rådgivningsgruppen til at anbefale", hvis mindst et medlem, men højst syv medlemmer af gruppen har ønsket ikke at tilslutte sig.

En anbefaling er afslutningsvis placeret under overskriften "Et mindretal i rådgivningsgruppen anbefaler", hvis otte eller flere medlemmer af gruppen har ønsket ikke at tilslutte sig.

Som konsekvens af disse spilleregler gøres der opmærksom på:

- Argumenter for og imod en anbefaling udgør eksempler på, hvad der har præget rådgivningsgruppens diskussioner, men skal ikke forveksles med en fuldstændig analyse.
- Argumenter for og imod en anbefaling udgør eksempler på, hvordan individer i gruppen har udtalt sig om deres vurderinger af evalueringsresultater og deres synspunkter om de nationale test. At argumenterne er gengivet i indstillingen betyder ikke, at Rådgivningsgruppen som helhed har godkendt hvert enkelt af de gengivne argumenter.

Rådgivningsgruppens medlemmer har haft mulighed for ved navns nævnelse i fodnoter at give udtryk for, at de støtter hhv. ikke støtter den enkelte anbefaling. Ingen medlemmer har ønsket at benytte sig af denne mulighed.

Rækkefølgen og nummereringen af de opstillede anbefalinger er ikke udtryk for en prioritering af deres relative vigtighed.

I tilgift til de opdelte anbefalinger har rådgivningsgruppens medlemmer haft mulighed for under eget forfatterskab at skrive såkaldte positionspapirer. Et positionspapir giver for eksempel mulighed for en længere sammenhængende fremstilling af begrundelser for en given stillingtagen til de forskellige anbefalinger. Der er fremkommet fem positionspapirer (se bilag).

2.1 Rådgivningsgruppen anbefaler

Generelle principper

ANBEFALING 1

Det anbefales, at folkeskolens formålsparagraf og fagenes formål i bred forstand bringes i erindring, når der tages beslutninger om nationale test, når testene i givet fald udformes, og når man fortolker og anvender testresultater. Af samme grund er det et fælles ansvar at udvise varsomhed i betoningen af de nationale test i forhold til skolens samlede styring og udvikling.

Argumenter for at støtte

I. De nationale test dækker en begrænset del af de værdier og målsætninger, som kommer til udtryk i folkeskolens formålsparagraf og i fagenes formål. Den konkrete udformning af test (hvad angår varighed, teknik mm.) kan sætte yderligere grænser for, hvad der måles i sådanne test.

ANBEFALING 2

Det anbefales, at arbejdet med de nationale test følger disse principper.

- Test skal underbygge en udfoldet og nuanceret evalueringskultur.
- Test skal måle forhold, der fremmer realiseringen af folkeskolens formålsparagraf.
- Beslutningstagere skal være opmærksomme på, hvilke problemer testen skaber.
- Test skal være transparente i forhold til, hvad de måler, hvordan de måler, og hvordan det er intenderet at de bruges.

ANBEFALING 3

Det anbefales, at børne- og undervisningsministeren med inspiration fra rådgivningsgruppens anbefalinger beskriver en samlet strategi for nationale test i samarbejde med de relevante parter i og omkring skolen.

Argumenter for at støtte

I. Vigtige anbefalinger risikerer at falde mellem stolene, hvis deres indbyrdes sammenhæng og kvalitative relevans for en samlet løsning ikke identificeres og fastholdes i en strategisk plan.

II. Skolens parter bør vide, hvilken samlet løsning de kan forvente, hvis rådgivningsgruppens anbefalinger lægges til grund for en ny løsning.

Principper for tilrettelæggelse af test og hvem der omfattes af test

ANBEFALING 4

Det anbefales, at de nye nationale test prioriterer at bibringe god information til understøttelse af kollektive praksisser.

Argumenter for at støtte

I. Test baseret på det adaptive princip fremmer et individperspektiv i testene, og kan bidrage til at skabe urealistiske forventninger om, at en enkelt test kan give god og

sikker information om enkeltelever. Evalueringen viser tydeligt, at dette ikke er tilfældet med de nuværende adaptive prøver.

II. Med en lineær og overvejende fælles test, vil man bedre kunne formidle, at resultaterne primært giver brugbar information til belysning af situationen i klasser, skoler og på højere niveauer. I klasseværelset bør lærerne bruge informationen til at evaluere, hvad man er lykkedes med eller ikke er lykkedes med i undervisningen.

ANBEFALING 5

Det anbefales, at indholdet af testen har en tydelig sammenhæng til Børne- og Undervisningsministeriets "Fælles Mål". Indholdet af testen skal passe til fagets fagformål, kompetencemål, underliggende færdigheds- og vidensområder, samt de vejledende færdigheds- og vidensmål.

Argumenter for at støtte

I. Argumentet for tydelig sammenhæng mellem test og Fælles Mål er, at testen tester noget, der kan fremme elevens faglige kompetencer i det enkelte fag. Ved at indholdet af testen har sammenhæng til den daglige undervisning, vil testens indhold opleves genkendeligt og derved kunne benyttes som et hjælpsomt læringsredskab.

ANBEFALING 6

Det anbefales, at nationale test skal indgå i og underbygge en udfoldet og nuanceret evalueringskultur. Heri indgår evalueringsinstrumenter og idékataloger til fagpersoners samarbejde om udvikling af undervisning. Der skal i højere grad ske løbende professionsudvikling/efteruddannelse samt forskning og udviklingsarbejde om test- og evalueringskultur.

Argumenter for at støtte

I. Evalueringsinstrumenter som fx tests, *rubrics* (beskrivelser af kriterier, der kendetegner gode besvarelser), selv- og kammeratskabsvurdering, dialogværktøjer mv. kvalificerer både kendskabet til elevernes kompetencer og det faglige sprog i undervisningen.

II. Der er en tendens til, at det enkeltstående tal, der er resultatet af test, betragtes som det vigtigste eller eneste relevante udtryk for elevernes kompetencer og kvaliteten af undervisningen. Dette fører til, at alle de andre vigtige aspekter overses eller nedvurderes. Derfor skal der gøres en stor indsats for at testen ikke står alene.

III. Nationale test tester ofte afgrænsede og mere tekniske aspekter af fagene (fordi det er lettere at teste). Derfor er der brug for aktivt at fremme evaluering af mere avancerede kompetencer (at kunne skrive hensigtsmæssigt til en given målgruppe, at kunne omsætte virkelighedsnære problemstillinger til en matematisk model, at kommunikere med indlevelse og forståelse i en tværkulturel situation, at kunne danne hypoteser der forklarer et fysisk fænomen osv.).

IV. Der er en tendens til, at ikke-fagpersoner har en snæver opfattelse af mål og indhold i undervisning ("eleverne skal lære at stave og regne") og derfor vurderer succes ud fra for snævre kriterier. En udfoldet evalueringskultur giver lærere mulighed for at dele en mere nuanceret opfattelse af elevernes kompetencer med forældre og andre eksterne personer.

V. En udfoldet og nuanceret evalueringskultur giver lærere, elever, forældre og skoleledere (og højere ledelseslag) redskaber til at tale nuanceret om styrker og svagheder på et informeret grundlag, så elever kan forstå, hvad de skal blive bedre til, og så beslutningstagere har bedre grundlag for at vurdere kvaliteten af undervisningen.

Argumenter for ikke at støtte

I. Det er dyrt at skabe rammerne.

ANBEFALING 7

Det anbefales, at Børne- og Undervisningsministeriet ved fremtidige ændringer i relation til nationale test sætter fokus på at fremme det pædagogiske formål og sammenknytte det med udvikling af evalueringskultur. Herunder at give lærere forbedrede vejledninger i at tolke resultater og anvende dem sammen med andre evaluerings- og dokumentationsdata til at tilpasse undervisning og særlige indsatser til elevernes behov og potentialer.

Argumenter for at støtte

I. Det pædagogiske formål og en styrket evalueringskultur var oprindeligt den centrale begrundelse for indførelsen af de nationale test.

II. Kun ved at foretage disse ændringer kan de nationale test blive et nyttigt pædagogisk redskab.

Argumenter for ikke at støtte

I. Det er korrekt, at ønsket om styrket evalueringskultur var et hovedargument for indførelsen af nationale test. Tager man udviklingen i de øvrige elementer i evaluering i skolen i betragtning, såsom elevplaner og en lang række afrapporterings- og dokumentationsformer, så er ordet evalueringskultur ikke længere det positive plusord, der kan tilgodese ny opbakning til test. Tiden er kommet til ikke blot at styrke og kvalificere evalueringskulturen, men også at sætte grænser for den.

II. Det legitime ønske om at styrke sammenhængen mellem testresultater og pædagogisk brug heraf kan tilgodeses ved en overgang til lineære test samt ved flere andre af de fremsatte anbefalinger.

ANBEFALING 8

Det anbefales, at alle aspekter af nationale test skal være let og billigst muligt tilgængelige for dem, der har legitime interesser (herunder fx rådata til forskere). Dette under hensyn til databeskyttelsesforordningen og under hensyntagen til at visse dele af testen (fx ankeropgaver) kan have høje krav om fortrolighed.

ANBEFALING 9

Det anbefales at erstatte det adaptive princip i de nationale test med et lineært princip, dog således at testopgaver og testudformning forinden er genstand for en grundig afprøvning.

Argumenter for at støtte

I. Selv om det adaptive princip teoretisk kan føre til de mest præcise testresultater, er det adaptive princip i praksis meget krævende med hensyn til antallet af testopgaver, den løbende afprøvning af opgavernes sværhedsgrad, statistisk kompetence, samt formidling og fortolkning af testresultater.

II. Det adaptive princip medfører, at enhver elev uanset dygtighed oplever ikke at kunne løse cirka halvdelen af de stillede opgaver, men samtidig bryder det adaptive testsystem båndet mellem, hvor dygtig eleven oplever at være, og hvor dygtig eleven faktisk er. Med lineære test vil man mere effektivt synliggøre fælles faglige standarder.

Rådgivningsgruppen anbefaler

III. Skoleledere, lærere, forældre, elever og andre føler sig ofte ude af stand til at forstå resultaterne fra de nuværende nationale test, herunder de statistiske forudsætninger. Hvis man tager udgangspunkt i, at teknologien skal tilpasse sig brugerne, ikke omvendt, så taler det for, at man fremover prioriterer brugervenlighed langt højere. Det kan gøres gennem et lineært testprincip (i kombination med øget krav til gennemsigtighed i udviklingen af testopgaver).

IV. Selv om det adaptive princip teoretisk set kan give anledning til de mest præcise testresultater, så er flertallet af test på internationalt plan (gennemgået i VIVE's analyse) ikke baseret på dette princip. Det tyder på, at test og prøver uden det adaptive princip sagtens kan være anvendelige. Mange af disse er i øvrigt langt mindre omstridte end de nationale test.

V. Diskussion om præcisionen i de nationale test bør sættes i sammenhæng med, hvilken præcision, der i øvrigt er i test, prøver og andre datakilder, som bruges i skolen. Også i folkeskolen findes en lang række test og prøver, som ikke er baseret på det adaptive princip. Den praktiske anvendelighed af en datakilde afhænger ikke af præcision alene, men også af professionel dømmekraft.

VI. De nationale test har ikke haft succes som pædagogisk redskab. En medvirkende årsag hertil er, at det ikke hidtil er lykkedes at give den enkelte lærer et hurtigt overblik over sammenhængen mellem opgavetyper og testresultater. Denne situation skyldes i et vist omfang det adaptive testprincip, som indebærer, at hver enkelt elev får sin egen kombination af testopgaver. Det adaptive princip hindrer derfor et let og hurtigt overblik over, hvordan man kommer pædagogisk videre på elev-niveau og klasseniveau. Lineære test kan medvirke til at afhjælpe dette problem, hvis en klasse får samme opgavesæt samtidig.

VII. I adaptive test har eleverne i en klasse besvaret forskellige opgaver. Dette gør det vanskeligt for lærerne at følge op. De konkrete svar fra eleverne kan give lærerne vigtig information om eleverne i klassen. En mere direkte indsigt i sammenhængen mellem opgavetyper og testresultater, som lineære test kan tilvejebringe, vil tillige medvirke til at fremme kollektiv meningsskabelse angående test og testresultater i dialogerne mellem lærere, fagkonsulenter, skoleledere, skoleejere, forældre, elever m.fl.

VIII. Selv om ethvert udvalg af testopgaver kan og skal gøres til genstand for kritisk diskussion og afprøvning, så afhænger udbredt anvendelse af testresultater i praksis af, at testopgaverne i hovedtræk opfattes som relevante og pålidelige. Med et lineært testprincip kan der – uanset om der kommer debat om en opgave eller to – opnås øget fælles demokratisk indsigt og kontrol med de opgaver, der stilles. Det kan fremme tilliden til relevansen af de opgaver, der stilles.

IX. Et lineært princip er langt mindre krævende end det adaptive med hensyn til det antal af testopgaver, der skal være fagligt relevante, gennemprøvede og validerede. Det medvirker til at gøre det sandsynligt, at man ved et lineært testprincip kan styrke testenes gennemskuelighed, accept og anvendelse uden at sætte væsentligt til, hvad angår testenes præcision og udviklingsomkostninger.

X. Lineære test kan implementeres på to måder. Umiddelbart kan man genbruge eksisterende opgaver fra de nationale test eller man kan konstruere helt ny opgaver. Denne anbefaling tager ikke stilling til hvilken af de to tilgange, der skal foretrækkes.

Argumenter for ikke at støtte

I. Generelt skal lineære test alt andet lige vare længere tid for at opnå tilfredsstillende reliabilitet.

II. Med lineære test vil lavt præsterende elever i mindre grad opleve en følelse af mestring, fordi et stort antal opgaver vil være for svære.

Rådgivningsgruppen anbefaler

ANBEFALING 10

Det anbefales at vie særlig opmærksomhed til den gruppe af elever, som føler ubehag og/eller har særlige vanskeligheder ved at deltage i de nationale test. Det kan fx omfatte en fast tidsramme for testen, at elever, der er færdige med de fastsatte opgaver, fortsætter med opgavelignende arbejdsopgaver, indtil tiden er gået, og at opgaverne er interessante, meningsfulde og udfordrende. Det kan desuden omfatte, at den konkrete test ikke tillægges så stor betydning af de voksne omkring eleverne, at eleverne oplever et pres for at klare sig godt.

Argumenter for at støtte

I. VIVE's evaluering og andre undersøgelser har vist, at omkring 10-20 procent af eleverne har en synligt dårlig oplevelse i forbindelse med test. Dette tal er uacceptabelt højt og unødvendigt.

II. En testsituation, som opleves som ubehagelig betyder, at eleverne ikke præsterer på deres almindelige niveau, og resultatet er derfor ikke et gyldigt udtryk for deres kompetencer.

III. De adaptive test resulterer i, at der er meget stor forskel på, hvornår eleverne er færdige med testen, og det skaber en uro i klassen som betyder, at de sidste elever kan begynde at svare mindre koncentreret eller decideret begynder bare at "klikke sig igennem". Derved er deres resultat ikke et sandt udtryk for deres dygtighed.

IV. De nationale test tillægges generelt stor betydning af de voksne omkring eleverne, og det kan betyde, at nogle elever oplever et unødvendigt pres for at klare sig godt.

ANBEFALING 11

Hvis obligatoriske test ikke besluttes, anbefales det at etablere national viden om norm og progression gennem en udtrækningsafprøvning af nationale test, hvor et mindre antal skoler pålægges at tage testen.

Argumenter for at støtte

I. En udtrækningsafprøvning giver grundlag for, at kommuner og skoler som ønsker det, kan sammenligne resultater fra kommunens eller skolens test med en repræsentativ national elevgruppe.

II. Antallet af elever, for hvem det er obligatorisk at tage testene, begrænses.

Argumenter for ikke at støtte

I. Der skal indgå mange skoler. For at sikre gyldige og generaliserbare resultater skal der indgå mindst 150 skoler for at opnå viden om det nationale niveau og forventeligt omkring 300 skoler for at kunne analysere diverse forskelle knyttet til geografi og socioøkonomi (dvs. potentielt op imod en fjerdedel af alle folkeskolerne). Forslaget vil altså stadig indebære et betydeligt tidsforbrug for de omfattede skoler, lærere og elever.

II. De udvalgte skoler, som obligatorisk skal deltage, kan opleve udvælgelsen som uretfærdig.

III. Sammenlignet med test, der omfatter alle elever, sikrer forslaget ikke imod uopdagede problemer på elev, klasse og skoleniveau.

Rådgivningsgruppen anbefaler

ANBEFALING 12

Det anbefales at imødegå tilfælde, hvor der systematisk testes i stof, som ikke er gennemgået på testtidspunktet – dog uden at indføre en mere striks regulering af, hvad der skal undervises i hvornår.

Argumenter for at støtte

I. Evalueringen har vist, at dette problem forekommer ved de nationale test i fysik i 8. klasse. Der er derfor grund til generelt at holde øje med problemet fremover.

Formidling og kommunikation

ANBEFALING 13

Det anbefales, at testresultater ikke må udgøre den eneste faktor i en tilbagemelding til forældre og ledelsesrepræsentanter. Grænserne for testens udsigelseskraft skal altid deklareres (hvad testen måler, og hvad den ikke måler).

Argumenter for at støtte

I. Resultater fra test udgør kun et mindre udsnit af, hvad elever har af kompetencer, så hvis dette resultat får en fremtrædende plads (fx gennem direkte tilbagemelding til forældre), så bliver det tillagt for stor betydning, og der bliver skabt opfattelser og taget for betydningsfulde beslutninger på denne baggrund.

ANBEFALING 14

Det anbefales, at formidling af resultater fra det nationale testsystem til forældre og elever overlades til elevens lærere, inden for de principper for tilbagemelding, som må være besluttet af skolebestyrelse, -ledelse og kommune. Lærerne bør altid supplere formidling af testresultater med formidling af lærernes andre vurderinger af elevens læring og trivsel, så et testresultat ikke står alene. Hvis forældre eksplicit beder om det, bør de kunne få udleveret det konkrete testresultat.

Argumenter for at støtte

I. Intet testresultat eller vurdering kan tegne det fulde billede af elevers læring og trivsel. Testresultaterne bør derfor ikke være konklusion på, men udgangspunkt for en dialog mellem skole og hjem om elevernes læring og trivsel. De standardbreve, som systemet i dag genererer til forældre, tager ikke hensyn til dette. Evalueringen viser, at de nationale test kun udgør en mindre del af skolernes samlede evalueringspraksis. Lærerne, der kender både elever og forældre, kan bedst vurdere, hvordan resultaterne bedst formidles, således at de sættes ind i en sammenhæng med elevens øvrige resultater, adfærd og trivsel.

II. Det ligger allerede i lovgivningen i dag, at skolebestyrelsen skal formulere principper for kommunikation til forældre om elevernes faglige udvikling. Nærværende anbefaling ændrer således ikke ved, at lærerne skal formidle resultaterne – kun måden.

III. Ved at lade læreren vurdere, hvordan resultaterne formidles, kan en for eksempel lav test score (i forhold til landsgennemsnittet) sættes i relation til elevens faglige progression i øvrigt, som kan være rigtig god (eller omvendt).

IV. Nogle elever kan være bekymrede for testresultaterne og måske tage dem for alvorligt i forhold til den usikkerhed, der er på alle testresultater og i forhold til de mange andre faktorer, der er væsentlige for elevernes udvikling. Andre elever kan måske tage for let på en dårlig faglig udvikling. Evaluering viser, at eleverne generelt er meget optagede af, hvordan de klarer sig i testene – blandt andet fordi forældre får resultaterne at vide. Ved at lade lærerne formidle resultaterne, kan lærerne vurdere, hvordan resultaterne formidles på en måde, der bedst tager hensyn til elevernes læring og trivsel generelt.

V. Evalueringen viser, at lærerne vurderer, at nogle forældre har vanskeligt ved at forstå de standardbreve, der i dag afrapporterer resultaterne fra testene. Sådanne forståelsesvanskeligheder kan afhjælpes, hvis formidlingen overlades til lærerne.

VI. Nærværende anbefaling kan kombineres med en særskilt anbefaling om udvikling af dialogredskaber – navnlig dialogredskaber, som kan understøtte lærernes dialog med forældre og elever. Nærværende anbefaling er således også forenelig med, at kommuner eller skoler vælger at have nogle fælles overordnede rammer for, hvordan formidling af resultater til forældre og elever skal foregå.

VII. I nogle situationer kan forældre have gavn af at kende de konkrete resultater af testene. Det kan for eksempel være, hvis lærere og forældre har meget forskellig vurdering af, hvordan eleven klarer sig fagligt. Da kan testene være et meget konkret udgangspunkt for en samtale. Denne indvending mod anbefalingen kan håndteres ved at give forældrene ret til at se de konkrete testresultater, hvis ikke lærerne selv præsenterer dem.

ANBEFALING 15

Når en tilfredsstillende gyldighed af de nationale tests er sikret, anbefales det, at der udvikles og testes dialogværktøjer til lokal anvendelse af testresultaterne. Udviklingen og tilpasningen sker i et samspil mellem nationalt og lokalt niveau under inddragelse af de relevante aktører, som på nationalt plan tæller bl.a. Skole og Forældre, Danske Skolelever, Danmarks Lærerforening, Skolelederforeningen, Børne- og Kulturchefforeningen, KL og Børne- og Undervisningsministeriet. På lokalt plan inddrages de aktører, der indgår i dialogen på det relevant niveau (fx forældre, elever og lærere på elevniveauet). Sammenhæng mellem elev-, klasse-, skole- og kommuneniveau er ønskværdig, men det vigtigste er en meningsfuld anvendelse på hvert af niveauerne.

Argumenter for at støtte

I. Evalueringsrapporten viser, at nationale tests ikke anvendes tilstrækkeligt konstruktivt i dag. Problemet er størst på lærerniveau efterfulgt af skoleledelsesniveauet, mens skolecheferne i højere grad allerede anvender resultaterne. En øget anvendelse kræver ejerskab og fleksibilitet i anvendelsesprocesserne, og det kræver inddragelse af de relevante aktører og deres nationale repræsentanter.

II. Hvis de nationale tests skal anvendes til både læring og styring/ledelse, skal der udvikles dialogværktøjer til flere forskellige fora, både inden for den enkelte skole og i skolens relation til kommunen (administrativ og politisk). Udvikling af fleksible delværktøjer til fri og fleksibel afbenyttelse på de enkelte niveauer kan gøre det nemt at tilpasse anvendelsen til den lokale kontekst.

III. Den nationale og lokale inddragelse bidrager til at forankre dialog og praksis for arbejdet med faglig kvalitetssikring i skolerne, som er fokuseret på elev-, klasse-, skole- og kommuneniveauerne. Udviklingen foregår såvel nationalt (udvikling af fleksible delværktøjer) som lokalt, hvorved den ultimative beslutningskompetence bliver henlagt til hhv. den enkelte lærer, skoleledelsen i samarbejde med lærerne, skolebestyrelsen, skolechefen og kommunalbestyrelsen, mens arbejdet med at lave processen gerne skulle være nemt og enkelt, fordi der er nationale værktøjer til rådighed at vælge mellem og tilpasse.

IV. Der er ikke tilrettelagt en anvendelsesstrategi på nationalt plan, da disse aktører opfordres til især at anvende tests, hvor Danmark bliver sammenlignet med andre lande. Derudover er der ingen ressourcebesparelse ved en fælles udarbejdelse af værktøjer på nationalt niveau, især da disse aktører vurderes at foretrække en fleksibel og dynamisk proces. Endelig er folkeskolernes kvalitet internt i Danmark kommunernes ansvar, mens Folketinget og regeringens opgave er at skaffe skolerne mindst lige så gode fælles rammer som i andre sammenlignelige lande.

V. Hvis man lykkes med fælles nationale forslag til dialogværktøjer til lokal tilpasning og anvendelse, vil det understøtte brugen af viden fra de nationalt udviklede test til at "lede opad" i organisationen, og dermed kvalificere andre personers beslutningsgrundlag.

VI. Et kvalitativt dialogmøde kan understøtte og bidrage til forståelsen og kendskabet til testens muligheder og styrke parternes analyser og omsætningsmuligheder.

Argumenter for ikke at støtte

I. Procesforslagene vedr. dialogværktøjer må ikke udvikle sig til at blive "anvendelse for anvendelsens skyld".

II. Inddragelsesforslagene vil tage tid for de involverede aktører.

III. Medmindre det aktivt modvirkes, kan processen give et uhensigtsmæssigt snævert fokus på de nationale test (aktørerne skal opfordres til at se testene som et element ud af mange i det samlede vidensgrundlag).

IV. Det vil være vigtigt at undgå et for snævert fokus på de nationale test, fordi en for instrumentel dialog, hvor parterne ikke har en åben og tillidsfuld dialog (ofte betegnet Open to learning conversations), ikke kan forventes at blive lærende og refleksiv og føre til nye handlinger/indsatser, hvis det er nødvendigt.

ANBEFALING 16

Det anbefales at inddrage nationale testresultater i skoleudvikling gennem dialoger mellem blandt andre skolens ejere, skolens ledelse, lærerne og faglige konsulenter, såvel som mellem lærere, elever og forældre. I dialogerne bør testresultater afdramatiseres og suppleres med lærernes observationer og andre data.

Argumenter for at støtte

I. Sådanne dialoger finder allerede sted på skoler, hvor der kan høres relativt positive vurderinger af test. Denne praksis kan yderligere udvikles, styrkes og udbredes.

ANBEFALING 17

Uanset om man følger anbefalingerne om fremover at anvende lineære test, eller om man fortsætter med adaptive test, anbefales det, at man kun rapporterer testresultater om enkelte elever, når dygtigheden er målt med tilstrækkelig sikkerhed, og når der ikke er tegn på, at testforløbet af en eller anden grund er slået fejl.

Argumenter for at støtte

I. Resultaterne fra både STIL og Bundsgaards og Kreiners analyser viser, at der er situationer, hvor dygtigheden måles med meget stor usikkerhed (fx SEM > 0,75). I disse situationer bør man undlade at bruge resultatet på elev-niveau. I stedet for at oplyse læreren om resultatet bør testsystemet oplyse læreren, at resultatet er så usikkert, at det ikke kan bruges. På populationsniveau, f.eks. i forbindelser med kortlægningen af det faglige niveau blandt samtlige elever, kan resultatet stadig benyttes.

II. Bundsgaards og Kreiners analyser af data fra 2017 viser, at der er situationer, hvor svarene ikke passer til den Rasch model, som de nationale test benytter, og hvor der er påfaldende mange fejlsvar, hvorved dygtigheden bliver systematisk undervurderet. I sådanne situationer skal testsystemet nægte at give et resultat.

III. I situationer, hvor svarene ikke passer til modellen, kan resultaterne heller ikke anvendes på populationsniveau. Bundsgaards og Kreiners rapport om læsning i 8. klasse 2017, viser imidlertid, hvorledes man kan beregne et validt men usikkert estimat af dygtigheden. Disse estimater kan bruges på populationsniveau og kan indgå i nationale opgørelser over det faglige niveau blandt danske elever. Det anbefales at Børne- og Undervisningsministeriet implementerer disse metoder.

ANBEFALING 18

Uanset om man følger anbefalingerne om fremover at anvende lineære test, eller om man beslutter at fortsætte med adaptive test, anbefales det, at de såkaldte percentil-scores og de nationale tests kriteriebaserede kategoriseringer erstattes af andre opgørelser af testresultaterne, der bedre kan tilgodese pædagogiske og styringsmæssige interesser.

Argumenter for at støtte

I. VIVEs evaluering af de nationale test pegede på store problemer med de såkaldte percentil scores.

II. Det er en tilbagevendende indvending fra lærere og skoleledere at testresultaterne er uforståelige og uanvendelige.

III. Problemerne kan reduceres ved at benytte transformerede logit scores fra Rasch modeller på samme måde, som PISA og andre internationale undersøgelser gør det. Tilsvarende transformationer sker i nationale test i fx Norge og USA.

IV. Man kan udvikle og dokumentere den form for kriterie-baserede proficiency kategorier, som kendes fra pædagogisk test-teori. De nationale tests kriteriebaserede scores var et skridt i den retning, men der er mange muligheder for at gøre dem mere nyttige for lærerne.

V. Børne- og Undervisningsministeriet benytter bl.a. gennemsnitlige percentil-score til at beskrive den faglige udvikling. Det bør ændres, således at man i stedet bruger gennemsnitsberegninger af de (transformerede) logit score, for at sikre, at udviklingen blandt de dygtigste og de mindst dygtige elever bliver beskrevet ordentligt.

ANBEFALING 19

Det anbefales at ændre Børne- og Undervisningsministeriets websider, så de i højere grad fokuserer på evalueringskultur. *Testogprøver.dk* ændres til kun at indeholde folkeskolens prøver, eller den lukkes helt ned. I stedet (gen)åbnes en evalueringsportal. Denne indeholder 1) udviklede og nyudviklede inspirationsmaterialer til brug for løbende interne evaluering i folkeskolen, 2) nationale test, samt 3) evt. folkeskolens prøver.

Argumenter for at støtte

I. De nationale tests formål er at styrke evalueringskulturen. Børne- og Undervisningsministeriets webkommunikation bør afspejle dette.

Rådgivningsgruppen anbefaler

Imødegåelse af teaching-to-the-test

ANBEFALING 20

Det anbefales i alle aspekter af nationale test (så som design, tilrettelæggelse, tidsmæssig placering og brug) at modvirke effekter i retning af teaching-to-the-test. Skoleejere og ledere har et særskilt ansvar for at sikre, at der ikke iværksættes tiltag med det primære formål alene at forbedre testresultater. Undervisning bør ikke tilrettelægges som træning i testlignende opgaver alene med henblik på forbedrede testresultater uden et begrundet pædagogisk sigte. Resultater fra nationale test alene bør ikke bruges til større beslutninger på lærer- eller elevniveau.

Argumenter for at støtte

I. VIVEs evaluering viser, at der er en udbredt teaching-to-the-test-kultur på danske skoler. Det betyder i realiteten, at de nationale test ikke måler elevernes dygtighed, men i hvilket omfang de er blevet forberedt på testene.

II. Teaching-to-the-test flytter opmærksomheden fra, at eleverne udvikler sig alsidigt fagligt og personligt, til at de bliver gode til at svare på spørgsmål inden for de snævre faglige områder som testene tester. Derfor skal teaching-to-the-test undgås.

III. Teaching-to-the-test øger elevernes orientering mod *præstation* i stedet for mod *mestring* af fagene og faglige praksisser.

Argumenter for ikke at støtte

I. Elever har behov og krav på at kender formen på opgaver, de testes i.

II. Anbefalingen kan tolkes i retning af, at der ikke må tages beslutninger på elev- og lærerniveau på en baggrund, hvori der indgår resultater fra de nationale test– heller ikke hvis det indgår i vurderinger sammen med andre evaluerings- og dokumentationsdata, hvilket er imod ånden i anvendelse af de nationale test pædagogisk.

ANBEFALING 21

Det anbefales, at nationale test som hovedregel tages i starten af skoleåret, således at undervisningen ikke indrettes med henblik på nationale test, men at test kan bruges til pædagogisk opfølgning.

Argumenter for at støtte

I. Ved at ændre tidspunktet for testene fra slutningen til begyndelsen af skoleåret, sender man et tydelig signal om, at de nationale test skal give information til støtte for læreres, skolers og kommunernes fremadrettede pædagogiske tiltag.

Argumenter for ikke at støtte

I. Anbefalingen kræver, at der testes i emner, der er gennemgået i skoleforløbet inden testtidspunktet.

Præcision og genberegninger

ANBEFALING 22

Det anbefales, at opgavernes sværhedsgrader genberegnes, og at de sværhedsgrader, som de nationale test fremover skal benytte, erstattes af korrekte sværhedsgrader. Den måde sværhedsgraderne skal beregnes på afhænger af, om de nationale test fremover skal være adaptivt eller lineært baseret på eksisterende opgaver.

Argumenter for at støtte

I. STIL's analyser viser, at de nationale test benytter forkerte sværhedsgrader, der stammer fra lineære afprøvninger af opgaverne, og at disse sværhedsgrader afviger fra sværhedsgraderne, når opgaverne bruges adaptivt. Af den grund er beregningerne af dygtigheden behæftet med systematiske fejl.

II. Testresultater er ganske vist altid behæftet med en vis grad af usystematiske fejl, men systematiske fejl må ikke forekomme. Af den grund skal de nationale tests sværhedsgrader erstattes af korrekte sværhedsgrader.

III. Anvendelse af forkerte sværhedsgrader betyder, at de nationale test kan risikere at vælge opgaver der enten er alt for lette eller alt for vanskelige for eleverne. Det betyder, at usikkerheden kan ende med at være større end nødvendigt, samt at især de svageste elever kan risikere at have negative oplevelser ved at blive bedt om at løse opgaver, de ikke har nogen muligheder for at besvare korrekt.

IV. STIL's notat gør opmærksom på, at Børne- og Undervisningsministeriets praksis i forbindelse med afprøvning af nye opgaver er at ændre opgavernes sværhedsgrader, når der findes tegn på at de afviger fra dem der allerede bruges. Forslaget kan derfor opfattes som et forslag om at følge en allerede etableret praksis.

V. Hvis man følger anbefalingerne om at bruge lineære test i stedet for adaptive test og hvis de lineære test konstrueres ved hjælp af de opgaver, som allerede eksisterer, vil det være fornuftigt at starte med at benytte de eksisterende sværhedsgrader, fordi disse er estimeret ud fra lineære testforløb. Disse sværhedsgrader skal dog kontrolleres ved hjælp af data fra første obligatoriske brug af de lineære test.

VI. Hvis man foretrækker at fortsætte med adaptive test, bør sværhedsgraderne estimeres ved hjælp af data fra de obligatoriske test fra de sidste 3-4 år.

Udviklingsarbejde

ANBEFALING 23

Det anbefales at afprøve nye test, inden de tages i brug. Afprøvningen bør inddrage både teknikere og brugere på alle niveauer.

Argumenter for at støtte

I. Der eksisterer allerede megen viden om udvikling af test, men konkrete testopgaver skal udvikles (eller eksisterende skal genvurderes) for at vide, hvordan de fungerer både teknisk og for brugerne.

II. En afprøvning bør også gå på formidling af resultatet til brugerne, så det sikres, at test og formidling indrettes efter brugernes behov og kompetencer og ikke omvendt.

III. En afprøvning af et nyt testsystem vil også kunne vise forholdet mellem præcision i målingen og testens varighed på forskellige klassetrin.

Rådgivningsgruppen anbefaler

IV. En afprøvning vil give et kvalificeret beslutningsgrundlag i tilfælde, hvor en enkelt testform ikke kan opfylde alle ønsker på én gang (for eksempel begrænset tid til test (45 minutter) versus høj reliabilitet; måling af mange, adskilte aspekter af et fag (profilområder) versus reliabilitet og målingsvaliditet; fortolkbarhed/anvendelighed for lærere som i lineære test versus højere reliabilitet som i adaptive test).

V. Evalueringen har ikke peget på eksisterende test, som direkte kan bruges i stedet for de nuværende.

VI. Det vil øge accepten af de nye test på alle planer i skolesystemet, hvis testen er udviklet og afprøvet med både anvendelse og måleteknik for øje, under inddragelse af relevante parter, og hvis der opnås enighed om, at testene er så gode, som det kan lade sig gøre – givet de eventuelle afvejninger, man er nødt til at foretage.

VII. Afprøvningen af forskellige testformer kan i udviklingsfasen foregå sideløbende med og uafhængigt af det nationale testsystem, som i mellemtiden måtte være gældende.

VIII. Lodtrækningsforsøg vil til nogle aspekter (navnlig dem, der handler om anvendelse) være velegnede. Lodtrækningsforsøg kan ske blandt skoler, som melder sig frivilligt til at bidrage til udviklingen af testene.

IX. Når testene skal overgå fra lille skala til drift/stor skala, kan det overvejes at lade nogle skoler anvende gamle test og nogle nye test. Det vil give en yderligere afprøvning og en vis sikring mod IT-nedbrud og andre uforudsete hændelser med de nye testformer.

X. Nærværende anbefaling kan kombineres med særskilt anbefaling om, at de adaptive test udskiftes med lineære test. Jo mere sådanne nye test kan blive afprøvet, inden de tages i brug, desto større chance er der for, at nye lineære test bliver sammensat og afrapporteret bedst muligt. Det kan også gøre det muligt at undersøge, hvordan nye test korresponderer med de nuværende test.

XI. Nærværende anbefaling kan kombineres med særskilt anbefaling om, hvordan udvikling af nye test bør begynde med et conceptual framework.

Argumenter for ikke at støtte

I. Det vil tage tid at gennemprøve nye testformer.

II. Det vil øge udgifterne til udvikling af nye test med en grundig afprøvning.

ANBEFALING 24

Det anbefales, at alle de nationalt udviklede værktøjer evalueres løbende i samarbejde med de lokale brugere, og at de i anbefaling 15 nævnte nationale udviklingsaktører mødes mindst hvert andet år for at drøfte, om værktøjerne og den anbefalede brug heraf skal revideres baseret på denne løbende evaluering.

Argumenter for at støtte

I. Vi får løbende viden om, hvilke værktøjer der virker bedst i de forskellige fora, og der indbygges en proces til at bringe denne viden i spil i forhold til fremtidig udvikling og brug af værktøjerne.

Argumenter for ikke at støtte

I. Evalueringer af tests lokalt tager tid og andre ressourcer, og det samme gælder den løbende tilpasning på nationalt plan.

ANBEFALING 25

Det anbefales, at Børne og Undervisningsministeriet håndterer brugen og den fremtidige udvikling af de nationale test på en måde, der så vidt muligt lever op til den kvalitet og den standard, som professionelle testudbydere forsøger at leve op til.

Argumenter for at støtte

I. VIVE's evaluering dokumenterer en udbredt mistillid til de nationale test blandt lærere og skoleledere. Ministeriet har derfor en stor opgave med at genoprette tilliden. En måde at gøre det på vil være at sige, at den fremtidige implementering af de nationale test vil leve op til standarder som for eksempel er beskrevet i "STANDARDS for Educational and Psychological Testing" fra the American Educational Research Association and the National Council on Measurement in Education.[4]

II. Forslaget er en måde at fremtidssikre nytteværdien af testene på, fordi værdien af testene først og fremmest afhænger af lærernes brug af testene.

ANBEFALING 26

Det anbefales, at testopgaver skal udvikles ifølge internationale standarder (fx IMS Question and Test Interoperability specification, QTI).

Argumenter for at støtte

I. Opgaver udviklet af andre aktører (fx inden for FLIP+, der er et samarbejde mellem undervisningsministerier og forskningsinstitutioner i en række lande) vil kunne oversættes og tilpasses danske forhold.

II. Opgaver, der har været brugt i en national test, vil kunne anvendes sidenhen i frit tilgængelige test, der kan relateres til en norm (resultatet fra det år, opgaven blev anvendt).

III. Ved at følge en standard, knytter man sig ikke til et teknisk system, men kan skifte testsystem.

ANBEFALING 27

Det anbefales, at der stilles tydelige krav til de miljøer, der udpeges til at lave nye nationale test. De bør have anerkendt fagdidaktisk og testteoretisk kompetence. Endvidere bør testene udvikles i en proces med ekstern kvalitetssikring og dokumenterbarhed og med inddragelse af brugere.

Argumenter for at støtte

I. Analyser af det faglige indhold i de nuværende nationale test viser med al tydelighed, at testene ikke dækker fagenes indhold optimalt. Det er muligt at udvikle gode opgaver, som med større kvalitet tester elevernes kompetencer, men dette fordrer, at stærke faglige miljøer bliver overdraget opgaven.

II. Nationale test bør udfordre skoler og læreres vurderingspraksisser. Det er derfor vigtigt, at testene udover at understøtte gode målinger også reflekterer "state of the art" og "best practice" inden for feltet.

III. Ekstern kvalitetssikring og krav om dokumentation vil sikre troværdighed, transparens og et sundt kritisk rum for debat.

4 Svend Kreiners henviser til sit positionspapir, der giver eksempler på nogle af disse standarder.

Rådgivningsgruppen anbefaler

Argumenter for ikke at støtte

I. Der findes i dag kun få danske miljøer, som har en sådan kompetence. Det kan derfor tage tid at opbygge miljøer.

II. Mindre konkurrence, krav om høj kompetence og grundig dokumentation vil føre til dyrere testudvikling.

ANBEFALING 28

Det anbefales, at der tages initiativ til at etablere et nyt teknisk testsystem, der tager højde for de fejl og uhensigtsmæssigheder, der har vist sig gennem udviklingen og brugen af de nationale test.

Argumenter for at støtte

I. De nationale test anvender et teknisk system der ikke er up-to-date i forhold til standarder for test (fx IMS Global Learning Consortiums tekniske specifikationer for items).

II. De nationale test tester meget snævre tekniske aspekter af fagene, og de tekniske rammer for systemet (opgavetyper) kan ikke umiddelbart tilpasses de moderne formater, som muliggør test af mere avancerede aspekter af fagene.

III. Der eksisterer velafprøvede systemer (fx TAO som anvendes i Frankrigs årlige test af 8 millioner elever, og netop er introduceret i både Norges i forbindelse med nasjonale prøver og i en lang række andre lande), som er open source, og som lever op til moderne standarder for testudvikling.

IV. Det adaptive princip har vist sig at skabe flere problemer end det har løst, og dette princip er hårdkodet ind i de nationale tests testsystem.

Argumenter for ikke at støtte

I. Det er dyrt og vil tage tid at udvikle et system, der adresserer de problemer, som har vist sig med de nationale test.

II. De nationale test har vist at et nationalt testsystem fører til langt større problemer end det løser, og det bør derfor ikke være statens opgave at udvikle nationale test.

ANBEFALING 29

Det anbefales, at nationale test udvikles ud fra en tydelig specifikation, der beskriver hvilke centrale kvaliteter, som de endelige test skal efterleve: herunder 1) tydelige psykometriske kvalitetskriterier, 2) tydelige beskrivelser af hvad testen måler, inklusivt hvordan indholdet er relateret til Fælles Mål, 3) beskrivelse af typiske kendetegn for elevpræstationer på forskellige niveauer. Forud for tildelingen af opgaven om at udvikle nye nationale test, bør der derfor nedsættes en arbejdsgruppe som udarbejder begrundede psykometriske kriterier for alle testene. Der bør også nedsættes arbejdsgrupper for hvert fag, som udvikler specifikationer for de enkelte fag. Disse arbejdsgrupper skal være bredt sammensat med praktisk, fagdidaktisk og testteoretisk ekspertise, og de skal begrunde og dokumentere de specifikationer, de anbefaler.

Argumenter for at støtte

I. Uden en tydelig specifikation er der stor risiko for, at man efterfølgende ikke ved, hvad testene måler (eller hvad de ikke måler). Det vil derfor være vanskeligt at validere tolkningerne af testresultaterne.

II. En tydelig specifikation vil bidrage til, at der udvikles flere opgaver, som også fungerer godt empirisk. Dette er også omkostningsbesparende.

III. En tydeligt specifikation vil bidrage til bedre tolkninger af resultater og bedre kommunikation til alle involverede.

Argumenter for ikke at støtte

I. En tydeligt specifikation vil sætte fokus på et udsnit af fagets indhold. Dette kan efterfølgende opfattes som værende vigtigere end fagets andet indhold.

ANBEFALING 30

Det anbefales, at der på baggrund af de første års test gennemføres procedurer for *standard setting*, bl.a. med henblik på at synliggøre mulig progression for elever og lærere.

Argumenter for at støtte

I. Resultatet af standard setting er gode beskrivelser af, hvad der kendetegner elevernes præstationer på forskellige niveauer.

II. Gennem en standard setting vil man også få vigtig information til validering af tolkninger, og det giver information, som kan bruges til at justere rammeværket.

Argumenter for ikke at støtte

I. Standard setting er relativt omfattende logistiske processer som kræver tid og ressourcer.

ANBEFALING 31

Det anbefales, at der udvikles et design, som gør det muligt at koble resultater fra et års test til de senere års test i samme fag og på samme klassetrin. Dette kan eksempelvis opnås ved, at et udvalg af elever får en anden variant af testen med ankeropgaver.

Argumenter for at støtte

I. Det adaptive design varetager koblingen direkte. Hvis man indfører lineære test uden et sådan design, vil man ikke få information om, hvordan elevernes og skolernes præstationer udvikler sig over tid.

II. For at kunne evaluere mulige effekter af indsatser har skoler og ledere behov for information om, hvordan eleverne udvikler sig fra en årgang til den næste.

Argumenter for ikke at støtte

I. På skoleniveau kan sådan information være ustabil for mindre skoler.

II. Nogle få elever (måske 3000) vil hvert år få en test, som er delvist anderledes end testene for de resterende elever i klassen.

III. Et simpelt ankerdesign kan være udfordrende at implementere i test med læsning – eller andre test, som er afhængige af, at man kan implementere et sæt af opgaver, der relaterer sig til et fælles stimulusmateriale.

Rådgivningsgruppen anbefaler

ANBEFALING 32

Det anbefales, at der igangsættes udviklingsorienteret forskning for at afprøve designs, der kan følge de samme elever over tid. Det vil sige, at man udvikler nationale test på tværs af klassetrin, der bliver koblet sammen. Gennem et sådan design kan man rapportere scorer for elever på forskellige klassetrin på den samme skala.

Argumenter for at støtte

I. Hvis man lykkes med at udvikle fælles skalaer over klassetrin, vil man have et test-system, som gør det lettere at evaluere effekter af indsatser over tid. Både skoler, skoleledere og forskere vil hermed få et meget stærkt værktøj.

Argumenter for ikke at støtte

I. Dette kræver tid og ressourcer.

II. Det er metodisk komplekst og kræver specifik kompetence.

Navneændring

ANBEFALING 33

Det anbefales, at testen ændrer navn fra de nationale test. Navnet kunne eksempelvis være Fælles Test.

Argumenter for at støtte

I. Dette skal tydeliggøre forskellen mellem de nationale test og de nye test.

II. Dette skal bidrage til at sikre at den eventuelle skepsis over for testene hos nogle af brugerne af de nationale test ikke påvirker implementeringen af de nye test.

2.2 Der er bred tilslutning i rådgivningsgruppen til at anbefale

Principper for tilrettelæggelse af test og hvem der omfattes af test

ANBEFALING 34

Det anbefales, at der skal være standardiserede test i Danmark (forstået som test, der er ens i hele landet, og som omfatter alle elever på udvalgte årgange i skoleforløbet), dog skal dette være på betingelse af, at testen og resultatet kan bruges til informativ og konstruktiv feedback.

Argumenter for at støtte

I. Det overordnede argument til fordel for nationale test er, at de giver anledning til – i kommuner, skoler og klasser og i forhold til den enkelte elev – at reflektere over, om man lokalt er, hvor man ønsker at være. Dog er det altafgørende, at testen og resultatets udformning opleves af lærere og elever som et anvendeligt og ikke mindst vedkommende læringsværktøj.

II. Test, der omfatter alle elever, kan understøtte en balance mellem på den ene side at have viden om og anledning til at reagere på problemer for alle elever, klasser, skoler og kommuner og på den anden side at respektere det kommunale selvstyre samt give et tilstrækkeligt råderum hos lærere, skoleledere og kommuner. En begrænset mængde test, som alle deltager i på udvalgte tidspunkter i skoleforløbet, giver et fælles grundlag, der omfatter alle, men det afskriver ikke muligheden for derudover at have en testbank, som kommuner, skoler og lærere kan bruge fleksibelt.

III. Med obligatoriske nationale test er det ikke muligt at følge en tilskyndelse til at undgå tests hos kommuner, skoler og lærere, der kunne komme til at fremstå mindre positivt pga. dårlige testresultater. Ved at et begrænset antal tests omfatter alle elever på obligatorisk vis, bliver vidensgrundlaget om elevernes resultater fælles og dækkende, og det kan begrænse risikoen for, at alvorlige problemer ikke bliver opdaget, selvom det aldrig vil være en garanti.

Argumenter for ikke at støtte

I. En obligatorisk test forhindrer lokale forsøg med andre typer af evalueringsredskaber og test.

II. En obligatorisk test fratager skoleejerne muligheden for at skabe en evalueringspraksis, som er tilpasset de lokale forhold.

III. En obligatorisk test bliver let *high-stakes*, så der opstår en teaching-to-the-test-kultur.

IV. En obligatorisk test vil skabe unødig modstand mod et i øvrigt anvendeligt pædagogisk redskab.

V. Hvis hovedsigtet med nationale test er at understøtte lærernes pædagogiske arbejde, så har et obligatorisk testsystem kun marginal effekt i tilgift til det omfattende pædagogiske arbejde, der allerede finder sted, og som i vidt omfang er understøttet af en lang række eksisterende planlægnings-, dokumentations-, evaluerings- og testformer.

VI. Et nationalt testsystem, som er "gratis" for alle skoler og lærere, kan stille sig hindrende i vejen for udbud af test fra private aktører, der bidrager til et mangfoldigt og varieret udbud af testformer, der er målrettet forskellige behov hos brugerne.

VII. Tilslutning til et testsystem for alle elever skaber et dilemma angående elever i de frie skoler. Hvis man ønsker fortsat at tilgodese de frie skoler og deres bidrag til et mang-

Der er bred tilslutning i rådgivningsgruppen til at anbefale

foldigt skolemiljø i Danmark, og dermed stille dem frit vedrørende brugen af nationale test, kan obligatoriske test i folkeskolen udgøre et selvstændigt argument for at vælge de frie skoler for de mange elever og forældre, som ønsker et testfrit skolemiljø.

ANBEFALING 35

Det anbefales at fastholde test i 2. klasse, men testens længde og form tilpasses til dette klassetrin.

Argumenter for at støtte

I. Evalueringen viser, at nogle elever oplever testsituationen positivt, mange oplever den som neutral og et mindretal oplever den negativt. Særligt blandt de yngre elever er der udfordringer i forhold til længden af testen. Det kan imødegås ved, at eleverne på forhånd ved, hvor lang tid testen varer, og at dette ikke er længere, end hvad elever i de mindre klasser kan kapere.

II. Evalueringen viser samtidig, at testresultaterne i Dansk (læsning) på 2. klassetrin har en relativt stærk sammenhæng med elevernes karaktergennemsnit i folkeskolens afgangsprøver i dansk i 9. klasser syv år senere. Resultaterne fra 2. klasse er for eksempel et bedre pejlemærke for resultatet ved afgangsprøverne end alle de oplysninger, der ofte bruges i beregning af socio-økonomiske reference (forældres uddannelse, indkomst, indvandrerbaggrund m.v.) tilsammen.

III. Testene i de små klasser giver således en tidlig indikation på alle niveauer af, om der behov for særlig støtte til nogle skoler, klasser eller enkeltelever.

IV. Jo tidligere man kan sætte ind med særlig støtte, desto bedre chancer er der for, at støtten kan nå at virke for eleverne. Derfor er der en fordel i at fastholde test i de mindre klasser.

V. Nærværende anbefaling kan kombineres med særskilt anbefaling om indførelse af lineære test, da det formentlig vil give navnlig yngre elever en bedre forståelse af, hvordan testene fungerer.

VI. Nærværende anbefaling kan kombineres med særskilt anbefaling om, at det overlades til lærerne, hvordan resultaterne af testene formidles til elever og forældre. Det vil gøre, at lærerne, navnligt over for de yngre elever og deres forældre, kan tage hensyn til, hvordan testene bedst formidles, herunder at det understreges at testresultatet ikke er udtryk for fastlåst bedømmelse af deres evner, men en indikation af, om der er nogle områder, hvor de har særligt behov for støtte og arbejde.

VII. Nærværende anbefaling kan kombineres med særskilt anbefaling om afprøvning af nye test. Sådan afprøvning vil give et bedre vidensgrundlag for afvejning mellem længde af test og præcision i de mindre klasser og i forhold til om lineære test vil opleves mere eller mindre positivt i de mindre klasser. Hvor meget tiden skal afgrænses (for eksempel til 30 eller 45 minutter, eller om den skal opdeles på to lektioner med pause imellem), afhænger af en afvejning mellem elevernes trivsel under testen og præcision i testene.

Argumenter for ikke at støtte

I. Hvis testene gøres kortere, vil de alt andet lige blive mere upræcise. Dette må afvejes i forhold til hensynet om at kunne sætte tidligt ind med særlig hjælp og støtte.

ANBEFALING 36

Det anbefales, at aktører der bruger test, skal være, eller gives mulighed for at blive, i stand til at fortolke, vurdere og bruge test inden for de rammer testene har.

Der er bred tilslutning i rådgivningsgruppen til at anbefale

Argumenter for at støtte

I. Ifølge VIVE's evaluering er betydelige andele ikke i stand til at fortolke resultater af de nationale test. Det fører til forkerte opfattelser af eleverne, klasserne og skolerne, og det kan resultere i uhensigtsmæssige eller unødvendige beslutninger.

II. Lærere, skoleledere og kommunale konsulenter, der kan tolke testene korrekt, kan bruge dem hensigtsmæssigt i deres tilrettelæggelse af undervisning og indsatser over for enkelte elever og klasser.

Argumenter for ikke at støtte

I. Det vil kræve omkostningstung efteruddannelse af lærere, skoleledere og kommunale konsulenter for at sætte dem i stand til at tolke testresultater.

ANBEFALING 37

Det anbefales, at brugervenlighed prioriteres som det afgørende princip i designet af et nyt testsystem. Som konsekvens heraf bør der være mindre vægt på at forklare brugere om testresultaternes psykometriske og statistiske forudsætninger og egenskaber. Der bør i stedet lægges mere vægt på at forklare testudviklere, hvordan test og kommunikationen af testresultater tilrettelægges ud fra brugernes behov. Inspiration til denne vending kan hentes i design-tænkning, brugerdreven innovation og evaluering.

Argumenter for at støtte

I. Evalueringen viser, at det eksisterende testsystem producerer resultater, som ofte ikke er intuitivt forståelige. En del skoleledere og lærere føler sig koblet ud af diskussionen om brugen af de nationale test. Endvidere viser evalueringen, at lærere oplever det som meget tidskrævende at sætte sig ind i testresultaterne. Det tyder på, at det eksisterende testsystem har været baseret på uhensigtsmæssige design-principper, hvor man ikke har taget tilstrækkeligt hensyn til disse centrale brugere.

II. Der findes en række meget anvendte test- og prøveformer, som ikke kræver særlige statistiske forkundskaber. Accept og brugervenlighed kan opnås uden at forudsætte kendskab til statistiske usikkerhedsberegninger på brugerniveau.

III. Fremover bør brug af statistisk terminologi og kendskab til usikkerhedsberegninger derfor ikke opstilles som nødvendige forudsætninger for at kunne forstå og bruge testresultater, men bør alene være tilgængelige i de omfang, som de er efterspurgte blandt brugere.

ANBEFALING 38

Det anbefales, at testopgaver så vidt muligt skal være interessante at besvare og eksemplariske, så lærerne kan få inspiration til, hvordan de kan teste og vurdere elevernes dygtighed.[5]

Argumenter for at støtte

I. Der er et begrænset antal opgavetyper i de nationale test, som meget eksplicit handler om at svare rigtigt eller forkert. Det er muligt at udvikle opgaver der opleves som udfordrende og meningsfulde og hvor den der besvarer opgaven, ikke oplever at svaret er forkert.

5 Jeppe Bundsgaard henviser til sit positionspapir for yderligere baggrund for anbefalingen.

Der er bred tilslutning i rådgivningsgruppen til at anbefale

II. Meningsfulde opgaver findes allerede i et vist omfang i PISA, PIRLS, TIMSS og i særlig grad i ICILS, og i test udviklet inden for forskningsområdet *Assessment of 21st Century Skills* samt i regi af den internationale sammenslutning af nationale testudbydere FLIP+.

III. Der findes tekniske løsninger som muliggør udvikling af interessante opgaver, herunder udvikling af særligt tilrettede opgaveformater (fx TAO).

Argumenter for ikke at støtte

I. Det er en mere omfattende og kompliceret proces at udvikle og afprøve interessante opgaver.

II. Det kræver mere avancerede testsystemer end de nationale tests tekniske system.

ANBEFALING 39

En obligatorisk test skal også være obligatorisk for privatskoler, der modtager statstilskud.

Argumenter for at støtte

I. Forskning mangler viden fra op mod en femtedel af elevpopulationen.

II. Statslige myndigheder mangler indsigt i status for elever på privatskoler.

III. Forældre, der er imod nationale test, kan fravælge folkeskoler alene af den grund.

IV. Hvis ikke de private skoler tager de nationale test, er det ikke muligt for beslutningstagerne at følge den faglige udvikling i Danmark som helhed.

Argumenter for ikke at støtte

I. Man skal ikke tvinge privatskoler til at underlægge sig samme styringsregime som folkeskoler.

II. Forældre skal have et frit valg, herunder muligheden for at fravælge at deres børn skal deltage i test.

ANBEFALING 40

Det anbefales, at der på sigt udvikles to sidestillede testsystemer — et til pædagogisk brug og et, der kan fungere som styringsredskab. Hvis man ikke ønsker at have to forskellige testsystemer, kan man opnå nogle de samme fordele ved de to systemer ved at pege på, at de frivillige test kan benyttes til pædagogiske formål i starten af skoleåret og ved at lade lærerne fravælge profilområder, som de ser som mindre relevante i forbindelse med de frivillige test. (Jf. Anbefaling 43).

Argumenter for at støtte

I. Behovet for at kunne bruge de nationale test som styringsredskab har stillet sig i vejen for forbedringer af de nationale test, der kunne tilgodese de pædagogiske anvendelser. Det er internationalt set et velkendt problem, som begrunder, at man skiller tingene ad. Ved at tillade frivillige test i starten af skoleåret og ved at lade det være op til lærere og skoleledere selv at bestemme hvilke dele, der skal bruges, kan man opnå nogle af de samme fordele, som man ville opnå ved at have to forskellige systemer.

II. Begrundelsen for at placere de frivillige test i starten af skoleåret er, at det vil tilgodese pædagogiske hensyn uden at ødelægge noget for beslutningstagernes muligheder for at følge den faglige udvikling. Et tidligt tidspunkt vil også reducere risikoen for teaching-to-the-test.

Der er bred tilslutning i rådgivningsgruppen til at anbefale

Argumenter ikke for at støtte

I. Rådgivningsgruppen fremsætter øvrige forslag til ændringer, der understøtter brugen af testene som pædagogisk redskab inden for de kendte rammer af de nationale test, idet de støtter forståelsen af det enkelte testresultat og lader lærerne give meningsfuld feedback på den enkelte elevs testpræstation. Skolerne har allerede mulighed for at inddrage integreret testning af bredere pædagogiske kompetencer – efter eget valg – ved brug af nye undervisningssystemer, og formålsrettede pædagogiske test-redskaber, som er kommercielt tilgængelige fra danske udbydere, og som omfatter alle klassetrin. Tillæg af et nyt pædagogisk testsystem falder således uden for rammen af opgaveopdraget.

II. En udvidelse af de nationale test til også at omfatte didaktiske anbefalinger og undervisningsvejledning medfører en uklar opdeling mellem undervisnings- og kontroltiltag. En udvidet, fast pædagogisk testramme, udstukket af Børne- og Undervisningsministeriet, vil således begrænse skolernes hidtidige ret til lokal undervisningstilrettelæggelse og metodevalg. I den forbindelse bemærkes yderligere, at rådgivningsgruppen anbefaler, at de nationale test skal være obligatoriske for alle grundskoler i Danmark, som opererer med offentligt tilskud (både folkeskoler samt privat- og friskoler, som modtager tilskud).

III. Mere specifik anvendelse af pædagogiske test er relevant, når det giver en lærer mulighed for at undersøge bestemte aspekter af en elevs eller klasses skolefærdigheder, som kan afklare eventuelle udfordringer. Sådanne udredninger inddrager testredskaber, udvalgt af fagpersoner tæt på eleverne, som har følsomhed for de pågældende elevers udfordringer og faglige niveau. Landsdækkende testning i de nationale test med et generelt pædagogisk fokuseret testbatteri tilgodeser ikke disse behov.

IV. Et statsligt pædagogisk testsystem underbyder private aktører på området og gør nyudvikling og innovation inden for området urentabel. Ved introduktionen af et sådant system påtager Børne- og Undervisningsministeriet sig således en forpligtelse til fremadrettet selv at finansiere, vedligeholde og udvikle systemet og øvrige tiltag inden for området.

Formidling og kommunikation

ANBEFALING 41

Resultater fra test skal være meningsfulde (klart forklare hvad en elev med et givent resultat fagligt er i stand til, og endnu ikke mestrer). Det skal være klart for de professionelle brugere, hvad usikkerheden er på resultatet både for den enkelte elev og for aggregerede niveauer (klasse, skole osv.).[6]

Argumenter for at støtte

I. De nationale test har givet resultater på den såkaldte percentilskala og med såkaldt kriteriebaserede ord. Disse tal er et udtryk for elevens placering i forhold til elever i 2010, og de kriteriebaserede ord relaterede sig på tilsvarende måde til andre elever (jævn, fremragende). Dette gav ikke læreren viden om elevens faglige kompetencer og udviklingspotentiale. Percentilværdierne er ikke meningsfulde at regne på baggrund af (der er ikke lige langt mellem 1 og 2 og mellem 49 og 50).

II. Test kan give empirisk grundlag for at beskrive i detaljer, hvad elever med et givent resultat i en test er i stand til at løse af faglige opgaver, og hvad den nærmeste zone for udvikling er for eleven.

6 Jeppe Bundsgaard henviser til sit positionspapir for yderligere baggrund for anbefalingen.

Der er bred tilslutning i rådgivningsgruppen til at anbefale

III. De nationale test har i de senere år givet et 68-procentkonfidensinterval på den enkelte elevs resultat. Det bør være 95-procentinterval. På aggregerede niveauer har der ikke været beregnet usikkerhed på trods af, at den ikke er ubetydelig i hvert fald på klasse- og skoleniveau. Det har ført til forkerte opfattelser af forskelle mellem grupper.

Imødegåelse af teaching-to-the-test

ANBEFALING 42

Det anbefales ikke at knytte politiske målsætninger til udviklingen i testresultater over tid.

Argumenter for at støtte

I. Enhver test afspejler en balance mellem fordele og ulemper. Testresultater kan være vejledende strømpile. De kan vise, hvor man afviger fra et forventet resultat, f.eks. på landsplan. Men hvis testresultater gøres til politiske og ledelsesmæssige mål i sig selv, så øger man ulemperne, herunder at kommuner, skoler og lærere særskilt betoner de afgrænsede dele af fagene, som der testes i. Forslaget kan medvirke til at dæmpe teaching-to-the-test.

II. Endvidere medfører en strategisk satsning på forbedrede testresultater, at testene svækkes som måleredskab.

Argumenter for ikke at støtte

I. Det kan tolkes i retning af, at der ikke må knyttes politiske målsætninger til udviklinger over tid, hvori der indgår resultater fra de nationale test– heller ikke hvis det indgår i vurderinger sammen med andre evaluerings- og dokumentationsdata.

ANBEFALING 43

Det anbefales for frivillige tests vedkommende, at Børne- og Undervisningsministeriet indskærper over for lærere, skoleledere og skoleejere, at frivillige test kun skal bruges til pædagogiske formål på det tidspunkt, hvor lærerne finder det nyttigt, og at de frivillige test ændres, således at lærerne kan fravælge profilområder, de ikke har brug for.

Argumenter for at støtte

I. Uanset om man ønsker at fortsætte med adaptive eller lineære test, løser man en del af problemerne med sammenblandingen af de pædagogiske og styringsmæssige interesser i testene, hvis man lader lærerne bestemme, hvad der skal testes i, og hvornår det sker i forbindelse med de frivillige test.

II. Teaching-to-the-test er ødelæggende for det pædagogiske udbytte af testresultaterne. Forslaget vil reducere risikoen for, at det forekommer, især for de lærere, der tager de frivillige test tidligt i skoleåret.

III. Resultater fra frivillige test kan bruges til at definere, hvor de obligatoriske test skal starte, så man undgår, at de svageste elever starter med opgaver, der ligger langt over deres niveau.

IV. En ændring af de frivillige test således, at det er læreren, der styrer, hvad der skal ske, har ingen konsekvenser for de obligatoriske test sidst i skoleåret.

V. Fordelen ved at lade lærerne selv beslutte, hvilke dele de frivillige test skal indeholde betyder, at de kan fravælge mindre interessant profilområder for til gengæld at få mere sikre målinger af de centrale dele. F. eks. ved at fravælge afkodning og sprogforståelse i 8. klasse for at få sikre målinger af tekstforståelsen.

Der er bred tilslutning i rådgivningsgruppen til at anbefale

ANBEFALING 44

Det anbefales, at den enkelte elev kun én gang må tage den nationale test i det pågældende fag på det gældende klassetrin (Ønsker man en enkelt prøvekørsel forinden for at vænne sig til testformatet, kan man eventuelt tage en demo-version en enkelt gang).

Argumenter for at støtte

I. Der bør ikke lægges gentagen og vedvarende stor vægt på resultater af en enkelt test, da disse kun repræsenterer en begrænset del af folkeskolens formål og fagenes formål. Forslaget skal derfor ses som et blandt flere redskaber til at dæmpe tendenser i retning af teaching-to-the-test.

II. De nationale test kan ikke forventes at beskrive små fremskridt over kort tid for en enkelt elev meget præcist. Gentagne målinger på kort tid af en enkelt elev er den brug af de nationale test, der vil være mest præget af tilfældige udsving, hvilket kan bidrage til at svække tiltroen til nationale test generelt.

III. I samspillet mellem lærer, elev og klasse bør der følges op på testresultaterne, men det kan med stor fordel ske efter behov i det daglige pædagogiske arbejde uden indblanding af et nationalt testinstrument.

Argumenter for ikke at støtte

I. Følges denne anbefaling, kan man ikke benytte de nationale test til gentagne progressionsmålinger i et kortere tidsforløb.

Præcision og genberegning

ANBEFALING 45

Det anbefales, at de tal for dygtigheden, der er blevet beregnet i forbindelse med tidligere obligatoriske test fra 2010 til 2019 genberegnes vha. de estimater af sværhedsgrader, som er baseret på obligatorisk adaptive testforløb.

Argumenter for at støtte

I. Det er af afgørende betydning for Børne og Undervisningsministeriets muligheder for at beskrive udviklingen i det faglige niveau siden 2010 og for evalueringen af skolereformen, at tallene for dygtigheden beregnes uden systematiske fejl for at undgå, at systematiske fejlberegninger slører den udvikling, der har fundet sted.

II. VIVE's aktuelle evaluering af effekten af skolereformen er baseret på forkerte beregninger af dygtigheden. STIL's notater dokumenterer bl.a. at der er stor risiko for, at eleverne er blevet placeret i forkerte kategorier. Da VIVE basere deres analyser på de nationale tests kriteriebaserede kategorier er det vigtigt at minde om disse resultater. (jf. Anbefaling 18 der også handler om disse problemer).

III. Ovenstående begrundelse burde være mere end tilstrækkelig for Børne og Undervisningsministeriet. En anden begrundelse kunne være, at forskningsresultater baseret på tidligere resultater fra de nationale test bør kunne genberegnes, så man har mulighed for at kontrollere om konklusionerne holder.

Der er bred tilslutning i rådgivningsgruppen til at anbefale

ANBEFALING 46

Hvis man beslutter at fortsætte med adaptive test eller beslutter at ændre de nationale test uden at sætte brugen af de aktuelle adaptive test i bero, indtil de lineære er på plads, anbefales det at Børne- og Undervisningsministeriet øjeblikkeligt tager skridt til at forbedre præcisionen i testresultater fra de nationale test, således at testresultaterne også kan bruges på elev og klasseniveau.

Argumenter for at støtte

I. Usikkerheden i testresultater fra de nationale test er væsentlig større end usikkerheden i almindelige pædagogiske test og opfattes som et stort problem for lærere og skoleledere. Usikkerheden bør derfor reduceres.[7]

II. STIL har i deres notater allerede peget på en række muligheder for at forbedre sikkerheden, men der er flere muligheder, som bør forsøges.[8]

Udviklingsarbejde

ANBEFALING 47

Det anbefales, at de nationale test undergår en timeout indtil opgavesværhedsgraderne er korrekt beregnet, og at de nationale test derefter fortsætter ind til et nyt system er iværksat. Det meddeles skolerne, at resultaterne fra de nationale test er meget usikre på individniveau.

Argumenter for at støtte

I. Meget store andele af items (opgaver) i de nationale test har sværhedsgrader, der ikke passer i den adaptive afvikling. De nationale test giver derfor forkerte resultater for eleverne. Dette gælder både for den enkelte elev og for aggregerede niveauer (klasser, skoler, kommuner, nationalt).

II. Der er meget stor usikkerhed på den enkelte elevs resultat. Det gør ifølge VIVE's evaluering i praksis resultaterne uanvendelige på elevniveau. Det er uklart om usikkerheden også gør resultatet uanvendeligt på klasse- og skoleniveau.

III. Det kræver indgående forståelse af betydningen af usikkerhed samt konfidensintervaller at tolke – og særligt ikke at mistolke – resultaternes betydning. VIVE's evaluering viser, at en sådan forståelse mangler hos store andele af både lærere og skoleledere.

IV. En (ikke kendt, men sandsynligvis ikke ubetydelig) andel af eleverne svarer på en måde der ikke følger den statistiske model, og deres dygtighed er derfor ukendt. Det skal være muligt at identificere disse elever, så der ikke tages beslutninger på forkert grundlag.

V. Der er ikke forbindelse mellem test på forskellige klassetrin, så det er ikke muligt at identificere progression på tværs af klassetrin. Man kan derfor ikke se, om der er sket en udvikling.

VI. At genberegne opgavernes sværhedsgrader vil, hvis arbejdet prioriteres tilstrækkeligt højt, være muligt at fuldføre indenfor rimelig tid, således at de nationale test kan afholdes også i foråret 2020.

7 Svend Kreiners henviser til sit positionspapir, der indeholder en oversigt over sikkerheden og reliabiliteten for en række forskellige test, der i alle tilfælde giver væsentlig sikrere målinger end DNT.

8 Svend Kreiner henviser til sit positionspapir, der indeholder en oversigt over de muligheder for at forbedre sikkerheden.

Der er bred tilslutning i rådgivningsgruppen til at anbefale

Argumenter for ikke at støtte

I. Problemerne med de nationale test er så gennemgribende både i forhold til testenes indhold, måletekniske forhold, praktisk brug og styringsanvendelse, at det ikke giver mening at fortsætte med de nationale test, uanset at de måletekniske problemer løses.

II. Forslaget om at stille de nationale test i bero forudsætter et kriterium for, hvornår testene er gode nok til at blive sat i gang igen. Der er ikke klarhed og enighed vedrørende et sådant kriterium.

III. Blot fordi det eksisterende testsystem har svagheder, kan man ikke slutte, at det vil være bedre slet ikke at have nationale test.

ANBEFALING 48

Det anbefales at støtte udvikling af standardiserede og normerede test af høj kvalitet inden for en bred vifte af faglige områder.[9]

Argumenter for at støtte

I. Der er en tendens til, at test er udviklet inden for meget få og snævre faglige områder (særligt læsning, grammatik og regning), mens der er meget lidt viden om, hvad der er normalt og ønskværdigt inden for andre faglige områder.

II. Når test typisk tester få, forholdsvis tekniske aspekter af fagene, får disse aspekter en særlig opmærksomhed, som flytter fokus fra de mere avancerede og vigtigere kompetencer. Derfor skal lærere have instrumenter til at få øje og status på andre aspekter af fagene.

III. Der findes allerede internationale test og test udviklet til forskningsprojekter inden for bredere områder af fagene. Disse kan fungere som inspiration for udvikling af frit tilgængelige og normerede test.

Argumenter for ikke at støtte

I. Det er dyrt at udvikle sådanne test.

II. Der er ikke brug for flere, men færre test i skolen.

ANBEFALING 49

Det anbefales, at der udvikles et testsystem, som tager hensyn til, at forskellige test har forskellige formål. I tilknytning til de nationale test bør man udvikle en testbank, som bl.a. skal understøtte lærernes pædagogiske arbejde.

Argumenter for at støtte

I. Det er en fordel for såvel elevernes læring som handlerum for lærere, skoler og kommuner at have adgang til forskellige slags test i tilknytning til de nationale test. Eksempler omfatter: 1) Test af nye kompetencer, der er centrale i skolen, men hvor der ikke er etableret en god vurderingsform, 2) specifikke diagnostiske test, 3) screening-test for at identificere elever med specifikke læringsudfordringer, 4) test som kan hjælpe lærerenes karaktergivning, mv. Der er flere faglige områder, som ikke dækkes af de nationale test, men som er meget centrale for elevernes faglige dannelse. Et eksempel kan være demokrati og medborgerskab, et andet kan være IKT (informations- og kommunikationsteknologi). Der bør udvikles nogle nationale tests for disse fagområder for at synliggøre og stimulere god evalueringspraksis.

9 Jeppe Bundsgaard henviser til sit positionspapir for yderligere baggrund for anbefalingen.

Der er bred tilslutning i rådgivningsgruppen til at anbefale

II. Der er behov for diagnostiske test i sprogfagene, matematik og naturfag. Det er vigtigt, at man så tidligt som muligt retter opmærksomheden mod veldokumenterede hverdagsforestillinger og specifikke læringsudfordringer i disse fag.

Argumenter for ikke at støtte

I. Dette er test, som det kræver specifik kompetence at udvikle, men Danmark har allerede historisk tradition for at udvikle lignende tests i det daværende Pædagogisk psykologiske forlag. Statslig udvikling af sådanne test kan derfor ses som en offentlig understøttet aktivitet på et område, hvor der i dag findes private og kommercielle interesser.

II. Dette er test, som kunne være i brug i mange år. Der er derfor vigtigt, at lærerne, skolerne og skolelederne er klar over, at det er ødelæggende for praksis, hvis der opstår en forståelse af, at disse test er "high-stakes".

III. Rådgivningsgruppen fremsætter øvrige forslag til ændringer, der understøtter brugen af testene som pædagogisk redskab, inden for de kendte rammer af de nationale test, idet de støtter forståelsen af det enkelte testresultat og lader lærerne give meningsfuld feedback på den enkelte elevs testpræstation. Skolerne har allerede mulighed for at inddrage integreret testning af bredere pædagogiske kompetencer – efter eget valg – ved brug af nye undervisningssystemer, og formålsrettede pædagogiske testredskaber, som er kommercielt tilgængelige fra danske udbydere, og som omfatter alle klassetrin. Tillæg af et nyt pædagogisk testsystem falder således uden for rammen af opgaveopdraget.

IV. En udvidelse af de nationale test til også at omfatte didaktiske anbefalinger og undervisningsvejledning medfører en uklar opdeling mellem undervisnings- og kontroltiltag. En udvidet, fast pædagogisk testramme, udstukket af Børne- og Undervisningsministeriet, vil således begrænse skolernes hidtidige ret til lokal undervisningstilrettelæggelse og metodevalg. I den forbindelse bemærkes yderligere, at rådgivningsgruppen anbefaler, at de nationale test skal være obligatoriske for alle grundskoler i Danmark, som opererer med offentligt tilskud (både folkeskoler samt privat- og friskoler, som modtager tilskud).

V. Mere specifik anvendelse af pædagogiske test er relevant, når det giver en lærer mulighed for at undersøge bestemte aspekter af en elevs eller klasses skolefærdigheder, som kan afklare eventuelle udfordringer. Sådanne udredninger inddrager testredskaber, udvalgt af fagpersoner tæt på eleverne, som har følsomhed for de pågældende elevers udfordringer og faglige niveau. Landsdækkende testning i de nationale test med et generelt pædagogisk fokuseret testbatteri tilgodeser ikke disse behov.

VI. Et statsligt pædagogisk testsystem underbyder private aktører på området og gør nyudvikling og innovation inden for området urentabel. Ved introduktionen af et sådant system påtager Børne- og Undervisningsministeriet sig således en forpligtelse til fremadrettet selv at finansiere, vedligeholde og udvikle systemet og øvrige tiltag inden for området.

Der er bred tilslutning i rådgivningsgruppen til at anbefale

ANBEFALING 50

Det anbefales at udvikle test, der tester et bredt udsnit af faglige områder.

Argumenter for at støtte

I. De nationale test tester afgrænsede dele af fagene. Det betyder, at undervisningen retter sig mod disse dele, og derved nedprioriteres hele kompetenceområder.

II. Der findes test af tekniske aspekter af fagene i vidt mål (stavning, afkodning, regning, glosekendskab, grammatik), mens mere avancerede aspekter er mindre udbredte (eksamenssæt fra tidligere år er en undtagelse, men disse er ikke normerede). Det er muligt (PISA, TIMSS, PIRLS, ICILS gør det), men svært at udvikle test af mere avancerede aspekter, og der skal derfor gøres en særlig indsats for at sådanne test bliver alment tilgængelige for den enkelte lærer.

2.3 Et mindretal i rådgivningsgruppen anbefaler

Principper for tilrettelæggelse af test og hvem der omfattes af test

ANBEFALING 51

Det anbefales at afskaffe de nationale test i 2. klasse.

Argumenter for at støtte

I. Børn på dette klassetrin har særligt vanskeligt ved at forstå meningen med testen og principperne for testens afvikling.

II. Det er vanskeligt at standardisere de praktiske betingelser for afvikling af testen.

Argumenter for ikke at støtte

I. En udsættelse af test til lidt ældre klassetrin vil medføre en udsættelse af muligheden for at reagere på resultaterne fra de nationale test.

II. Også i starten af skoleforløbet er det et legitimt ønske/behov for kommuner og nationalt at kunne følge udviklingen af resultater over tid, så man kan sætte relevante indsatser ind tidligt.

III. Udfordringer med elevernes forståelse for mening og principper kan dels håndteres ved at afskaffe det nuværende standardbrev og lade lærerne vælge, hvordan resultater skal formidles, dels vil et eventuelt lineært testsystem gøre principperne lettere forståelige.

IV. Trods eventuelle vanskeligheder ved at standardisere betingelserne for afviklingen af test i 2. klasse, har resultaterne en relativt stærk sammenhæng med afgangsprøven i 9. klasse. Resultaterne er for eksempel et bedre pejlemærke end alle de oplysninger, der ofte bruges i beregning af socio-økonomiske indeks (forældres uddannelse, indkomst, indvandrerbaggrund m.v.) tilsammen.

ANBEFALING 52

Det anbefales at lade det være op til lokale beslutningstagere (kommuner, skoleledere, lærere) at afgøre, hvorvidt eleverne skal tage test.

Argumenter for at støtte

I. Et decentralt styringssystem åbner mulighed for afprøvning af nye tilgange og metoder som kan fungere bedre end et centralt bestemt system, som er svært at ændre og tilpasse til lokale forhold. Eksempler på andre tilgange og metoder kan fx være andre test, andre måleredskaber eller dialogiske tilgange.

II. Ikke alle kommuner og skoler er ens. Nogle kommuner og skoler kan have tilstrækkelig viden om skolerne fra andre, lokale metoder og instrumenter, således at de ikke behøver at belaste skolerne med yderligere test.

III. Styring kan virke demotiverende, hvis aktørerne oplever den som kontrollerende og illegitim. Det er derfor afgørende vigtigt, at de lokale aktører oplever medindflydelse.

IV. På det nationale niveau giver de internationale undersøgelser (PISA, PIRLS, TIMSS, ICILS, ICCS, TALIS) allerede vigtig og omfangsrig information om det nationale skolesystems status og udvikling. Med de internationale undersøgelser følger betydelig indsigt i kontekster for kompetencer (opnået gennem spørgeskemaer) som muliggør sammenligninger med andre relevante landes resultater.

Et mindretal i rådgivnings-gruppen anbefaler

Argumenter for ikke at støtte

I. Det er et behov for, at alle i styringskæden kan få information om, hvad der foregår på skoler.

II. Forskere har ikke adgang til data fra alle folkeskoler.

III. At have fælles nationale test er kun et lille indgreb i kommuners, skolers og læreres autonomi. Disse aktører har inden for Fælles Mål fuld frihed til at indrette undervisningen.

IV. Hvis det er frivilligt at deltage i fælles nationale test, vil en række analyser på nationalt niveau være vanskeligt at gennemføre. Fx vil det være sværere at dokumentere regionale forskelle, forskelle mellem forskellige kommunetyper, forskelle mellem store og små kommuner etc. Dette er en form for analyser, som også kan baseres på stikprøveanalyser, men i så fald må stikprøverne være betydeligt større end 150-200 skoler.

V. Selvom understøttelse af det kommunale selvstyre, samt det lokale råderum, er vigtige forhold, vejer hensynet til sikkerhed for, at der er anledning til at tage eventuelle problemer op for alle elever samt hensynet til at få et fælles og sammenligneligt vidensgrundlag tungere, såfremt der er tale om et begrænset antal test, der giver gyldig viden uden at belaste eleverne på urimelig vis.

ANBEFALING 53

Det anbefales at lade det være op til lokale beslutningstagere at afgøre, hvornår på skoleåret test skal afvikles.

Argumenter for at støtte

I. Test kan bruges både som undersøgelse af resultatet af undervisning (og vil så skulle tages ved slutningen af et forløb, summativt) og som bidrag med viden om udgangspunktet for undervisning (og vil så skulle tages ved starten af et forløb, formativt).

II. Styringskæden kan have brug for information på forskellige tidspunkter på året – fx som input før budgetlægning, eller som kvalitetsindikator.

Argumenter for ikke at støtte

I. Det er vanskeligere at relatere et testresultat til et forventet niveau, hvis ikke den tages på samme tidspunkt som normen er taget.

II. Det faglige indhold i en test kan være for svært eller ikke undervist i, hvis testen tages tidligt.

ANBEFALING 54

Det anbefales at udvikle et testkoncept, der udnytter forhåndsviden om elevens dygtighedsniveau, hvor læreren, eleven eller dem begge i fællesskab, før testen indplacerer eleven i én af et på forhånd fastlagte dygtighedsgrupper, og hvor eleven får præsenteret et antal opgaver, som forventes at matche elevens forhåndsvalgte dygtighedsniveau. Udvalget af opgaver kan være tilfældigt som i adaptive test eller være opgaver i lineære test beregnet for elevens niveau. Afrapporteringen omfatter først og fremmest tal oplysninger om procent rigtigt løste opgaver.[10]

10 Peter Allerup henviser til sit positionspapir, der uddyber denne anbefaling.

Et mindretal i rådgivnings-gruppen anbefaler

Argumenter for at støtte

I. Testkonceptet tillader simpel procent-rigtig udregning som mål for elevens dygtighed. Da eleven kun har modtaget opgaver, der forventes at passe til hans eller hendes dygtighedsniveau, kan antal rigtige i testen beskrives med en simpel statistisk binomial-fordeling, som medfører, at det er relevant at bruge procent rigtige som mål for elev-dygtigheden.

II. Dette simple design vil modvirke nogle misforståelse og 'myter' skabt af det adaptive system, ved at benytte et fast antal opgaver i testen og ved at være meget gennem-skueligt.

III. Konceptet tydeliggør på simpel måde, hvor sikker vurderingen af eleven er. Usikker-heden på udregningerne belyses med to slags beregninger: (a) udregn et forventet antal rigtige for hele testforløbet (skal være ca. halvdelen af antallet af stillede opgaver pga. match-teknikken) og (b) vurdér den faktiske procent-rigtige i et 95% konfidensinterval for parameteren i binomialfordelingen med elevens 'tænkte' dygtighedsmål som parameter.

IV. Hvis de opgaver eleven skal besvare kommer fra lineære test med opgaver, der på forhånd er kendt af læreren, giver det meget bedre mulighed for formativ feedback til eleven, fordi læreren ved, hvilke opgaver eleven har besvaret.

Argumenter for ikke at støtte

I. Trods de beskrevne fordele ved dette system mister test deres objektive præg, når lærernes indplacering af den enkelte elev kommer ind i processen.

II. Der vil kunne rejses tvivl om testresultaternes troværdighed på tværs af klasser, skoler og kommuner, når man ikke ved, om resultaterne er en funktion af forskelle i elevernes dygtighed eller af forskelle i lærernes måde at indplacere elever på.

III. Hvis en elev eller dennes forældre udtrykker utilfredshed med en elevs indplacering, kan der opstå et ønske om en ny testafvikling baseret på en anden indplacering. Hvis den fører til nye testresultater, der afviger fra de tidligere, så vil det give næring til ny debat om eventuel utroværdighed i nationale test.

IV. Et argument imod at støtte udnyttelse af forhåndsviden om elevens dygtighedsniveau til, at læreren, eleven eller dem begge før testen indplacerer eleven i én af et på forhånd fastlagte dygtighedsgrupper, hænger sammen med risikoen for stigmatisering og fejl i indplaceringen. Hvis læreren på den ene side laver en meget præcis indplacering, erstatter indplacering nærmest testen, og hvis indplaceringen bliver baseret på lærerens forhåndskendskab, risikerer fx meget generte elever at blive placeret for lavt. Dertil kommer risikoen for, at eleverne oplever det som værende negativt at blive stemplet på forhånd som værende mindre dygtige.

V. Det vil indsnævre muligheden for, at en usikker, måske forkert, lærervurdering af en elevs niveau i et profilområde i et fag kan bliver korrigeret.

3 Bilag: Positionspapirer

Der er udarbejdet fem positionspapirer, her listet alfabetisk efter forfatternavn:

- Peter Allerup
- Lotte Bøgh Andersen, Simon Calmar Andersen, Stine Hertz, Thomas Dandanell Nielsen og Jakob Ryttersgaard med input fra Leanor Dall, Rasmus Edelberg og Signe Tofft
- Jeppe Bundsgaard
- Rasmus Edelberg, Skole & Forældre
- Svend Kreiner

Disse positionspapirer er inkluderet som bilag til rådgivningsgruppens anbefalinger.

Rådgivningsgruppen for evaluering af de nationale test

Anbefalinger, januar 2020

Design: BGRAPHIC

Børne- og Undervisningsministeriet
Frederiksholms Kanal 26
1220 København K

Positionspapir fra Peter Allerup professor emeritus Aarhus Universitet 23.01.2020

I forlængelse af diskussioner af de nationale test, specielt i forhold til at forstå resultaterne (scorerne) fra testen og formidle idéen med den adaptive proces ved testen foreslår jeg, at man rekonstruerer testen til en ny testtype, som kan afholdes individuelt på hvilket som helst tidspunkt af året og/eller bruges obligatorisk gældende for samtlige elever i Danmark – eventuelt forvaltet i praksis som en repræsentativ stikprøve.

Det anbefales at udvikle et testkoncept der udnytter forhåndsviden om elevens dygtighedsniveau, hvor læreren, eleven eller dem begge i fællesskab før testen indplacerer eleven i én af et på forhånd fastlagte dygtighedsgrupper. Eleven har forventninger om sin egen 'dygtighed' som i denne situation kommer til at svare til at deltage i mange velkendte spil, hvor eleven på forhånd vælger 'niveau' – enten ud fra en slags 'det prøver jeg' – eller 'det tror jeg' opfattelse.

Afhængig af valgte niveau får eleven præsenteret et antal opgaver, som er samstemt med elevens forhånds-valgte dygtighedsniveau, således, at der er omkring 50% sandsynlighed for at svare rigtigt, altså samme sandsynlighed, som det adaptive system i øjeblikket leder eleven ind i efter de indledende opgavesvar. Udvalget af opgaver kan være tilfældigt som i adaptive test eller være opgaver i lineære test beregnet for elevens niveau. Afrapporteringen omfatter første og fremmest taloplysninger om *procent rigtigt løste opgaver*.

1. Testkonceptet tillader simpel procent-rigtig udregning som mål for elevens dygtighed. Da eleven kun har modtaget opgaver, der forventes at passe til hans eller hendes dygtighedsniveau, kan antal rigtige i testen beskrives med en simpel statistisk binomialfordeling, som medfører at det er relevant at bruge procent rigtige som mål for elevdygtigheden.
2. Dette simple design vil modvirke nogle misforståelse og 'myter' skabt af det adaptive system, ved at benytte et fast antal opgaver i testen og ved at være meget gennemskueligt.
3. Konceptet tydeliggøre på simpel måde, hvor sikkert vurderingen af eleven er. Usikkerheden på udregningerne belyses med to slags beregninger: (a) udregn et forventet antal rigtige for hele testforløbet (skal være ca. halvdelen af antallet af stillede opgaver pga. match-teknikken) og (b) vurdér den faktiske procent-rigtige i et 95% konfidensinterval for parameteren i binomialfordelingen med elevens 'tænkte' dygtighedsmål som parameter.
4. Hvis de opgaver eleven skal besvare kommer fra lineære test med opgaver, der på forhånd er kendt af læreren giver det meget bedre mulighed for formativ feedback

til eleven fordi læreren ved, hvilke opgaver eleven har besvaret og kender deres didaktiske 'indhold'.

I øjeblikket bestemmer man antallet af stillede opgaver sammen med en løbende beregning/justering af elevdygtigheden. Erfaringerne tyder på, at netop denne løbende justering via det adaptive system har været én af de uudryddelige misforståelser mht. at forstå 'hvor længe' og 'hvor mange opgaver' eleven skal sidde og besvare. Dette forslag peger på muligheden af på *forhånd at bestemme antallet af opgaver.* Ved dette antal (eller fast tid) stoppes testen og eleven og læreren kan nu beregne (eller overlade til maskinen at beregne) *procent rigtigt løste opgaver*. Antallet af opgaver afspejles i de grænser for procent-rigtige, som binomialfordelingen leverer og som sammenligned med den procent-rigtige, som eleven har præsteret.

Bekymringerne vedrørende usikkerheden på udregningerne løses altså under dette forslag kan illustreres ved følgende eksempel. Først udregnes et forventet antal rigtige for hele testforløbet ud fra elevens valgte/forventede niveau, dvs. omkring halvdelen af antallet af stillede opgaver pgr. match-teknikken. Dernæst illustreres resultatformidlingen ved følgende tænkte elevs tilbagemelding på testen: "*Ja, du løste 45% af de stillede opgaver korrekt, men med din placering i den forventede/planlagte dygtighedsgruppe havde vi forventet, at du løste 53% rigtigt. Vi kan beregne, at du lige så godt kunne have løst et sted mellem 35% og 65% af opgaverne rigtigt -det er rene tilfældigheder, som afgør, om du ender i den ene eller anden af værdierne i det interval. Alt i alt betyder det, at din placering i gruppen den valgte gruppe er i orden, og du har klaret testen med et antal rigtigt løste opgaver som er forventet i forhold til din placering* ".

Grænserne 35% og 65% fastlægges som de grænser man som statistiker anvender (fx nederste 2.5% og øverste 97.5% fraktil), når man laver et statistisk test for, om den opnåede 45% er forenelig med (dvs. 'godkender') en påstand om at elevens dygtighedsmål svarer til (Rasch) skalaværdien for 'god' . Proceduren er den samme for *alle* dygtighedsgrupper, og det forventede antal rigtigt løste opgaver er for *alle* elever ca. halvdelen af antallet af opgaver. Altså en ret simpel måde at formidle test-usikkerheden på!

Hvis en elev, der startede testen i den tænkte kategori faktisk kun løste 10% af opgaverne rigtigt må man konstatere, at 10% ligger uden for intervallet [35%,65%] og i statistisk forstand forkaster man, at eleven har den udgangsværdi 'god', som var grundlaget for valg af matchende opgaver. Elevens faktiske dygtighedsniveau er altså (signifikant) *under* dent planlagte/forventede. I fortsættelse af det tænkte eksempel ville en passende melding i dette tilfælde måske være "*Ja, du løste 10% af de stillede opgaver korrekt, men med*

din placering havde vi forventet, at du løste 53% rigtigt. Du har altså et lavere resultat end forventet og skal måske overveje at tage testen én gang til på et lavere niveau ".

Det fremhæves at der under dette forslag er bedre muligheder for at give formativt feedback til eleven, ud over den besked at 'alting er OK,' som det skete i det første eksempel.. Men både her og ved det andet eksempel, hvor man forkastede en antagelse om elevens dygtighedsniveau, er der behov for at vide, *hvad* det var eleven kunne og ikke kunne, set fra et formativt feedback synspunkt. Med den foreslåede procedure præsenteres eleven udelukkende for opgaver med (næsten) samme sværhedsgrad. I sådanne tilfælde er det klart en mere overkommelig opgave at beskrive det faglige indhold i opgaverne ud fra didaktiske kriterier end tilfældet er i dag. Det var jo en del af selve konstruktionen af opgavebanken og medfølgende forståelse/fortolkning af opgavesværhedsgrad!

Med forslaget forsøges det hidtidige testsystem vendt lidt på hovedet. I stedet for 'ikke at vide noget som helst om eleven' og sætte testapparatet til at finde ud af, hvor dygtig eleven er, så undersøger den beskrevne fremgangsmåde 'om eleven ligger i den dygtighedsgruppe' som man på forhånd antager - eller eleven vil måske bare 'udfordre' et bestemt dygtighedsniveau? I hvert fald kan det gennemføres på 'frivillig basis' – som det er anført i punkt 6 – på de tidspunkter der passer eleven og læreren.

Afslutningsvist skal det bemærkes, at et sådant testsystem løbende kan opsamle viden om elevernes 'dygtighed' og herfra danne de sædvanlige normer, hvis der er behov for at sammenligne eleven faktiske præstation med "hvad elever på dette tidspunkt af året" sædvanligvis kan

Positionspapir: Baggrund vedrørende dialogredskaber og det at have en fælles test

Lotte Bøgh Andersen, Simon Calmar Andersen, Stine Hertz, Thomas Dandanell Nielsen og Jakob Ryttersgaard med input fra Leanor Dall, Rasmus Edelberg og Signe Tofft

Det er vigtigt at have en *fælles* skole med plads til lokal tilpasning og valgfrihed til elever og forældre. I Danmark sker det ved, at Folketinget sætter rammerne for både Folkeskolen og de frie grundskoler, mens kommunerne har ansvaret for driften af folkeskolerne, og forældrene vælger mellem forskellige folkeskoler og frie grundskoler. Denne balance mellem fælles elementer og råderummet for tilpasning i kommunerne og på skolerne samt valget for de enkelte brugere er vigtig. Det er uhensigtsmæssigt med stram national styring, men et totalt fravær af fælles rammer og fælles viden om resultaterne vil også forhindre, at vi samlet udvikler den danske grundskole.

Når man skal finde denne balance mellem fælles elementer og lokal selvbestemmelse, er det vigtigt at være opmærksom på betydningen af datainformeret ledelse på alle niveauer. Som Ledelseskommissionen skriver, er solid viden om resultaterne et nødvendigt grundlag for dialog. Det handler om at kunne træffe og udmønte beslutninger til gavn for eleverne på alle niveauer. Ledere på alle niveauer (skoleledere, skolebestyrelser, forvaltningsledere samt lærerne i deres funktion som klasseledere) har alle brug for at have et stærkt kendskab til den praksis, de er ledere for. Det kræver, at viden er relevant og gyldig, og at denne viden bliver anvendt på de relevante niveauer. Viden kan stamme fra såvel kvalitative som kvantitative datakilder, og ingen af dem kan stå alene.

Evalueringen af de nuværende nationale test viser tydeligt, at viden fra disse test ikke anvendes i særlig høj grad, særligt ikke af niveauerne tæt på eleverne. Når reviderede fælles test kommer til at give gyldige og sikre målinger af elevernes læring (og dermed har potentiale for at bidrage til beslutninger og handling på alle niveauerne), er ejerskab til anvendelsesprocessen det næste vigtige skridt. Det skal være så nemt som muligt at anvende resultaterne fra de fælles tests på en måde, der passer til den enkelte elev, klasse, skole og kommune.

Vi ved fra forskningen, at især medarbejdernes opfattelse af diverse styringstiltag er vigtig. Opfattes tiltag som understøttende eller som kontrollerende? Det har betydning for såvel anvendelsen som medarbejdernes generelle motivation. Sidstnævnte er vigtig i sig selv, og forskningen viser også, at lærernes motivation påvirker forhold så som målopfyldelse (fx elevers afgangskarakterer) og sygefravær. Nogle styringstiltag er af mange blevet set som en kontrolforanstaltning. Det gælder fx kravet om elevplaner for alle børn i alle fag. Andre tiltag er blevet mere positivt modtaget. Det gælder fx pædagogiske læreplaner i daginstitutionerne, hvor opfattelsen over tid også har ændret sig i understøttende retning. Det hænger sammen med, hvor fleksible og lokalt tilpassede styringstiltagene er. Det er et argument for at sikre stor inddragelse i planlægningen af anvendelsen af test. Vi ved desuden fra forskningen, at medarbejdere som lærerne med høj grad af professionalisme har særlig meget brug for at arbejde med tiltagene for at få ejerskab til dem.

Det betyder, at forslagene til dialogredskaber ikke skal ses som opskrifter på, hvordan man skal gøre lokalt. Der skal snarere udvikles fleksible redskaber, der kan tilpasses og kombineres lokalt. Det er også i det lys, man skal se de nedenstående eksempler fra forfatternes egen praksis på klasse-, skole- og kommuneniveau. De er ikke ideelle glansbilleder, men snarere eksempler på hvordan viden fra test kan anvendes konstruktivt.

I Gentofte Kommune og dermed fx på Skovshoved skole har der de seneste 10 år været stor vægt på dialog ud fra klassekonferencer. Disse konferencer blev i begyndelsen af en del lærere oplevet som kontrol, men over tid og med opmærksomhed på god facilitering fra skolernes ressourcepersoner er disse konferencer i dag vigtige og konstruktive i dialogen om den enkelte klasses lærere og pædagoger. Resultater fra de nuværende nationale tests indgår i konferencerne på lige med anden testning samt helhedsbilledet af barnet. Skovshoved Skole arbejder ud fra et systemisk perspektiv og lægger derfor stor vægt på at se barnet i relation til de omgivelser, barnet indgår i. Når der fremadrettet skal udvikles et dialogredskab, er det vigtigt at give plads til et systemisk børnesyn og til variationen fra kommune til kommune og mellem forskellige skoler. Det er også vigtigt at være opmærksom på, at dialogen ikke kommer af sig selv: Konstruktiv anvendelse på lærerniveau kræver en positiv invitation og sparring fra ledelsen.

En vigtig pointe er, at testresultater ofte kun er et enkelt element i den samlede viden om barnet og klassen. Et eksempel fra Nørre Alslev Skole illustrerer, at viden fra nationale test sagtens kan indgå konstruktivt i en dialog, der starter med lærernes overordnede vurdering af, hvordan det går i klassen. Det nedenstående skema bruges som forberedelse til klassekonferencerne. Skemaet udfyldes på forhånd af lærerne i fællesskab, og det er muligt at være uenig og således sætte forskellige krydser. Uenigheden i sig selv er interessant i den efterfølgende dialog. Det illustrerer fleksibiliteten i opbygning af dialogredskabet. Under klassekonferencen kigger lærere og ledelse på tidligere skemaer samt bruger diverse testmateriale til at underbygge skemaet. Der er meget sjældent uenigheder mellem testresultaterne, og hvad lærerne har krydset af. Selvom skemaet har vist sig som en god dialogstarter, er skolen opmærksom på, at det stadig kan forbedres, og det er også allerede sket løbende. Skemaet er således blevet brug i ca. 5 år, hvor det løbende er blevet forenklet i erkendelsen af, at en god dialog kræver en mindre kompleks forberedelsesstruktur, end man ofte tror.

I Gentofte Kommune bruges lignende skemaer, og kommunen har heller ikke standardiseret en skabelon. Hver skole udvikler og tilpasser til deres behov. Det styrker involveringen. I andre tilfælde kan der være fordele ved fælleskommunale redskaber. Eksempelvis har den meget fastlagte struktur, som Gentofte Kommune arbejder med omkring fællesskaber (kaldet Fællesskabsmodellen), givet et fælles sprog og en fælles måde at tale om børn på. Det har gjort arbejdet med Fællesskabsmodellen til et fælleskommunalt anliggende, og implementeringen har været gennemført med løbende opfølgning og efteruddannelse af de ressourcepersoner, som har skulle kunne facilitere brugen af den på de enkelte skoler. Det taler for, at udviklingen af dialogredskaber ift. anvendelsen af nationale tests skal adresseres på både lærer-, skole- og kommuneniveau. Dialogbaserede værktøjer kræver facilitering, og det er vigtigt, at der tænkes i implementering over en længere periode. Når det lykkedes, så rykker det ude i skolen.

Et eksempel på et sådan arbejde er Aalborg Kommunes udvikling af materiale vedr. Læringssamtaler. Det beskriver kæden af samtaler omkring børnenes læring og trivsel og kan findes her: https://www.nogetathavedeti.dk/laeringssamtaler samt https://drive.google.com/file/d/0B9Vwk3dZfMm6WlFHcXROeFZzclU/view. Dette materiale eksemplificerer dialogværktøjer på kommuneniveau, der kan bruges fleksibelt på skole- og klasseniveauerne. Dialogen giver blandt andet mulighed for en fælles undersøgelse af muligheder udfordringer og nye handlingsveje i arbejdet omkring elevernes læring og trivsel. Derfor er det også vigtigt at have et bredt blik på såvel de kvalitative som kvantitative data, så dialogen aktivt modvirkes til at få et for snævert fokus på en datakilde – eksempelvis De Nationale Test.

En af de vanskelige opgaver, når man anvender test og andre typer data er netop at gøre det på en måde, der gavner børnenes læring og trivsel. I bestræbelserne på at anvende data på en ordentlig og nuanceret måde kan man lade sig inspirere fra den øvrige offentlige sektor. Aarhus Kommune (Sundhed og Omsorg) har fx sammen med professor Simon Calmar Andersen og adjunkt Jakob Majlund Holm udviklet en spilleplade til brug for datainformeret ledelse. Spillepladen, der er vist nedenfor, illustrerer, at enkle værktøjer kan bidrage til en bedre datainformeret dialog.

Sammenfattende kræver konstruktiv anvendelse af fremtidige fælles test (1) dialog mellem de vigtigste aktører og (2) fleksibilitet i værktøjerne til dialogerne, så delværktøjerne nemt kan sættes sammen og tilpasses på en måde, der passer til konteksten. Målsætningen er at udvikle og teste elementer af dialogværktøjer, der skaber og øger ejerskab og dermed understøtter, at test bliver anvendt på en måde, der øger læring og trivsel hos både enkeltelever og grupper af elever. Det er vigtigt at finde en balance mellem, at dialogerne på den ene side ikke kommer til at følge faste og ufleksible opskrifter, men det lokale tidsforbrug til processen på den anden side ikke bliver for stort. Dialogen kan med fordel understøttes i følgende fem relationer:

- Mellem lærere, forældre og elever (her er fokus på den enkelte elev)
- Mellem lærere og skoleledelsen (her er fokus på klasserne)
- Mellem skoleledelsen og skolebestyrelsen (her er fokus på skolen i lokalområdet)
- Mellem skoleledelsen og skoleforvaltningen (her er fokus på skolen i kommunen)
- Mellem kommunalpolitikerne, skoleforvaltningen, skolelederne og den lokale lærerkreds (her er fokus på skolevæsnet i hele kommunen)

Eksempler på dialogredskaber fremgår på de følgende sider.

At arbejde data-informeret…

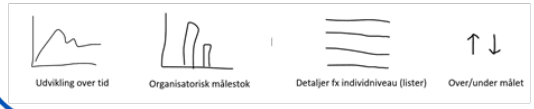
1. Vi modtager eller producerer data:

Hvad 'ved' vi *allerede*?
- Om borgerne?
- Om medarbejderne?
- Om lederne?

Nu kommer der så *nye* data. Hvilken slags viden er det? Og hvad kunne det *ellers* være godt at vide?

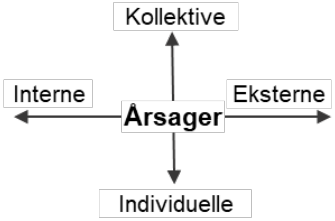
2. Vi fortolker:

Er der tendenser i tallene, som vi bør undre os over?
Kan de forstås på mere end én måde?
Er resultaterne tilfredsstillende?
Ud fra hvilket sammenligningsgrundlag?



3. Vi finder forklaringer:

Hvordan kan vi forklare de resultater, som undrer os eller som vi ikke synes er tilfredsstillende?
Skyldes det især:
- Eksterne forhold – dvs. forhold, som vi *ikke* har direkte indflydelse på
- Interne forhold – dvs. forhold, som vi *kan* påvirke
- Kollektive forhold, som påvirker alle
- Individuelle forhold, som knytter sig til den enkelte

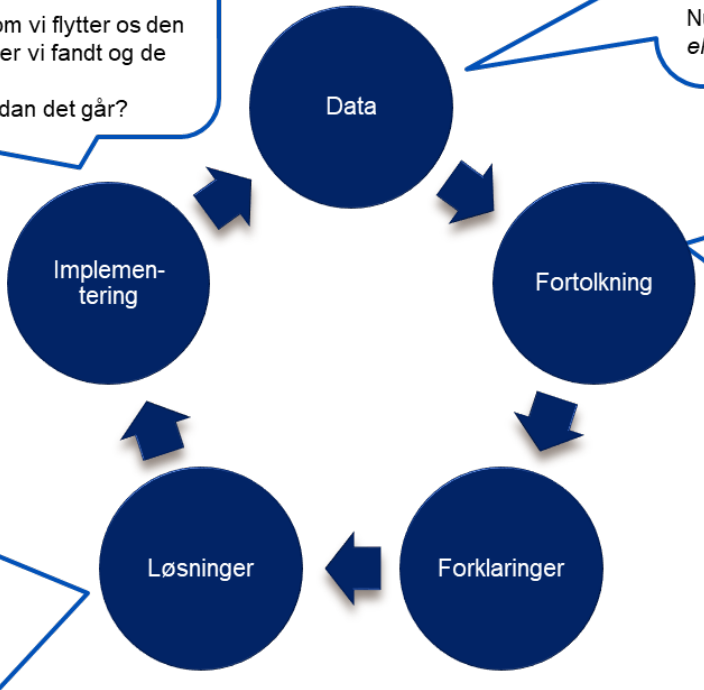


4. Vi vælger løsninger og handlinger:

Med afsæt i politisk vedtagne visioner og mål, organisationens strategi og vores faglige baggrund – hvad vil så være de bedste løsninger?

5. Vi implementerer og følger op:

Hvem gør nu hvad?
Hvordan vil vi i implementeringsprocessen se, om vi flytter os den rigtige vej ift. de udfordringer vi så, de forklaringer vi fandt og de løsninger vi valgte?
Hvilke data skal vi bruge? Hvem skal vide, hvordan det går?



Klassekonference - Efterår 2019

Klasse:

Navn	Socialt: Relationer til børn			Socialt: Relationer til voksne			Arbejds-vaner			Udbytte af fagligt potentiale			De gyldne 8	Fravær	Udv. forældresamarb.	Visitation	Suppl.uv/støtte	Indstillinger	GUSA	Underretning	UPV (Udskoling)	Bemærkninger

Udfyldt af:

Gennemgået på klassekonference den:

- God udvikling
- Kræver opmærksomhed
- Bekymrende

Odense, d. 22. januar 2019

Positionspapir til Rådgivningsgruppen for Evaluering af Nationale Test

De fire byggesten i test og evaluering

Jeppe Bundsgaard, DPU, Aarhus Universitet

Professor i psykometri fra University of California, Berkeley, Mark Wilson har introduceret fire grundlæggende principper for udvikling af test og evaluering og fire tilhørende byggesten til brug i udviklingen af test og evaluering. Samlet kalder han principper og byggesten for *BEAR Assessment System* efter *Berkeley Evaluation and Assessment Research Center* som han er leder af.

De fire principper lyder

1. Test og evaluering skal bygge på et udviklingsperspektiv for læring.
2. Det der undervises i og det der testes og evalueres, skal være i overenstemmelse.
3. Lærere skal kunne håndtere og bruge data.
4. Test og evaluering i skolen skal leve op til sunde standarder for validitet (gyldighed) og reliabilitet (pålidelighed) (Wilson, 2009).

Det første princip understreger at test skal have til hensigt at give læreren indsigt i hvad hendes elever er i stand til inden for et fagligt område – og derfor også hvad de ikke er i stand til, men skal arbejde hen imod. Resultater fra test og fra lærerens øvrige evaluering skal med andre ord kunne bruges til at støtte eleverne i det videre arbejde i undervisningen.

Det næste princip bygger på den erfaring at indholdet og formen på test og evaluering virker tilbage på det der undervises i, og at test og evaluering derfor skal være i overenstemmelse med det man ønsker der undervises i – og den måde der undervises på. Tilsvarende skal test og evaluering kunne bruges konstruktivt af læreren i den videre undervisning, og derfor skal opgaverne have en form og et indhold der kan bruges i eller relateres til undervisning.

Det tredje princip påpeger at det er lærere der skal bruge data, og at data derfor skal have en form der er overskuelig og let at fortolke. Med 25 elever nytter det ikke noget hvis der er forbundet en masse fortolkningsarbejde eller arbejde med at få adgang til den enkelte elevs resultater. Hvis læreren skal vurdere åbne besvarelser, skal kriterierne tilsvarende være enkle og overskuelige.

Test og evaluering måler noget der ikke er synligt, nemlig elevenes faglige kompetencer. Det fjerde princip siger derfor at det er afgørende vigtigt at test og evaluering faktisk måler det det hævdes at der måles, og at resultatet er korrekt og så præcist som det kan blive.

På baggrund af disse fire principper formulerer Mark Wilson fire byggesten for udvikling af test og evaluering af elevers viden og kompetencer. De fire byggesten er:

1. Beskrivelse af hvad der måles og hvad den faglige progression er.
2. Udarbejdelse af opgaver der relaterer sig til alle dele af den faglige progression.
3. Udvikling af kriterier for hvad der betragtes som rigtige svar (”giver point”).
4. Fastsættelse af en metode til at omsætte pointene til en målestok for elevernes dygtighed i relation til den faglige progression.

Jeg vil her tilføje yderligere to byggesten (den første er faktisk en del af Wilsons fjerde byggesten):

5. Forståelig og brugbar kommunikation af resultatet til dem der skal bruge det (læreren).
6. Vejledning i og inspiration til hvordan resultatet og testen bruges fagligt-pædagogisk.

I det følgende gennemgår jeg de fire plus to byggesten i relation til højt strukturerede tests som skal bruges til at teste elever i folkeskolen.

Hvad måles – den faglige progression

Når vi måler vægt, ved vi at det er et udtryk for hvor meget kraft der skal bruges for at flytte den målte genstand. Men hvad måler vi når vi fx måler læsning? Måler vi alene om eleven kan omsætte bogstaver til lyd? Eller måler vi også om eleven forstår ordene i teksten? Måler vi hvor hurtigt eleven kan læse en tekst? Eller måler vi sågar om eleven kan følge et argument i en tekst, kan identificere modstridende oplysninger, kan vurdere lødigheden af teksten eller troværdigheden i lyset af kommunikationsituationen? PISA-undersøgelserne fokuserer mere på de sidstnævnte af disse aspekter, mens De Nationale Test mere har fokus på de førstnævnte (ingen af dem måler i øvrigt læsehastighed, selv om PISA har forsøgt). Så det er i en forstand ikke det samme de to test tester, og læreren vil derfor kunne bruge de to test til forskellige ting.

Derfor er det afgørende at det er grundigt beskrevet hvad det er hensigten at testen måler – og også hvad den ikke måler. Det gøres typisk i såkaldte specifikationer eller frameworks før testen udvikles[1]. Mark Wilson understreger at det ikke er nok blot at beskrive det faglige område i bredden, men at det også skal beskrives i højden, dvs. det skal beskrives hvad der kendetegner en begynder og hvad der kendetegner en ekspert (relativt til den gruppe af elever der måles) – og et antal trin derimellem.

Selv om Fælles Mål og lignende fagbeskrivelser er et udtryk for en beskrivelse både i bredden og i højden af de faglige områder i fagene, vil de ofte ikke være tilstrækkeligt præcise og konkrete til at de kan fungere som faglige progressionsbeskrivelser for en test. For det første vil sådanne fagbeskrivelser som oftest være alt for brede til at en enkelt test kan komme omkring det hele, og for det andet vil de ofte være for indforståede både i bestemmelsen af bredden og af højden.

En faglig progressionsbeskrivelse indledes derfor ofte med en ganske kort definition af hvad der forstås ved det faglige område efterfulgt af beskrivelser af delaspekter af det faglige område og afsluttet med beskrivelser af grader af mestring. I PISA lyder definitionen af læsning fx:

> *At være i besiddelse af læsekompetence vil sige, at man kan forstå, bruge, vurdere, reflektere over og engagere sig i tekster så man kan opnå sine mål, udvikle sin viden og sine muligheder og deltage aktivt i samfundslivet (Bremholm & Bundsgaard, 2019, s. 19).*

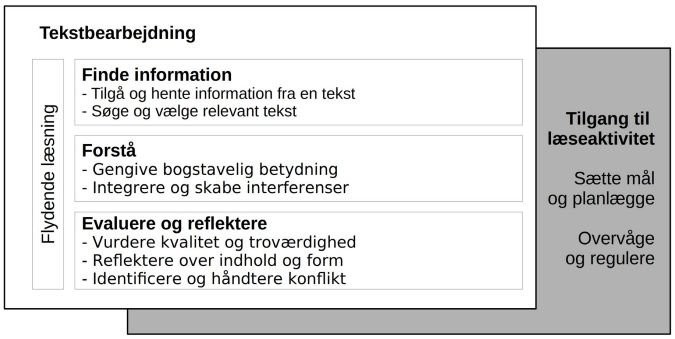
I beskrivelsen af det faglige område udarbejdes ofte modeller der kort og præcist beskriver sammenhængen mellem de centrale faktorer som udgør det faglige område. I PISA 2018 er figur 1 således en blandt flere centrale redskaber til at beskrive hvad der måles.

Hver af de centrale faktorer bliver dernæst defineret og beskrevet – i PISA 2018 fx således:

1 Men et sådant framework findes ikke for De Nationale Test.

2.1.2.6 At finde information

Denne kognitive proces består i at lede efter og finde bestemt information i tekster. Læseren behøver ikke nødvendigvis at læse hele teksten, men skal kunne danne sig overblik over teksten og afpasse læsemåde og -hastighed til at finde relevante tekstpassager. I digitale tekstformater skal læseren kunne overskue og håndtere bl.a. søgeresultater og hjemmesider.



Figur 1. PISA 2018. Kognitive aspekter af læseprocessen (Bremholm & Bundsgaard, 2019, s. 23).

Denne kognitive proces kan opdeles i følgende to delprocesser:

- *At tilgå og uddrage information i et stykke tekst: Læseren er i stand til at skanne- og overblikslæse en enkelt tekst eller uddrag af en tekst for på den måde at finde konkret information, der er direkte udtrykt i teksten i form af bestemte ord, en sætning eller tal.*

- *At søge efter og udvælge relevant information: Læseren er i stand til at finde frem til relevante tekstdele og informationer blandt et udvalg af tekster. Denne kognitive proces er især vigtig i forbindelse med digitale tekstmiljøer, hvor læseren ofte skal forholde sig til et større antal tekster. I denne proces er brugen af tekstelementer som overskrifter, kildeinformation (afsender, medie, udgivelsestidspunkt) og linkinformation (ved resultatsider ved netsøgning) ofte særlig vigtig (Bremholm & Bundsgaard, 2019, s. 23f.).*

Og til slut er det beskrevet hvordan den forventede eller almindelige progression er for den samlede læsekompetence. I PISA ser progressionsbeskrivelsen fx ud som i tabel 1. Som det fremgår, beskrives det hvad læsere er i stand til på det givne niveau – og også hvad de ikke er i stand til. På den måde kan en lærer få inspiration til hvad hun skal arbejde med i forhold til den gruppe af elever der ligger på det givne niveau – og hvad hun ikke behøver har fokus på længere for disse elever.

Tabel 1. PISA 2018. Beskrivelse af hvad elever kan og ikke kan på udvalgte kompetenceniveauer i læsning (Bremholm & Bundsgaard, 2019, s. 57f).

Kompetenceniveau	Hvad eleverne kan på dette niveau	Hvad eleverne ikke kan på dette niveau
Under niveau 2	Eleverne kan forstå den bogstavelige betydning af sætninger eller korte passager. De kan genkende hovedtemaer i et tekststykke om et velkendt emne. Og de kan skabe forbindelse mellem flere tæt forbundne informationer i en tekst.	Eleverne har svært at læse tekster der ikke er korte. De har svært ved at skabe sammenhæng selv i afgrænsede dele af teksten hvis informationerne ikke er udtrykt direkte. Og de har svært ved at læse tekster med forstyrrende eller modstridende information.
Niveau 2	Eleverne kan også identificere hovedpointen i et tekststykke af moderat længde. De kan forstå sammenhænge i afgrænsede tekstpassager ved at drage simple følgeslutninger. De kan håndtere tekster med forstyrrende information, hvis den søgte information er udtrykt direkte. De kan forholde	Eleverne har svært ved at læse moderat lange eller lange tekster med et ikke-velkendt indhold, og tekster, der har en moderat eller høj kompleksitet. De har svært at inddrage flere træk ved tekterne til at opnå en dybere tekstforståelse. Og de har svært ved at forstå tekster hvis den nødvendige information ikke er

	sig til simple visuelle og typografiske træk ved teksten. Og de kan søge og udvælge tekst og information blandt en samling af tekster baseret på eksplicitte anvisninger.	direkte udtrykt, hvis teksterne rummer anden i sammenhængen ikke-relevant eller kontraintuitiv information.
Niveau 4	Eleverne kan også forstå lange komplekse tekster med en form og et indhold, som ikke er velkendt. De kan læse på tværs af flere tekster og sammenholde udsagn og påstande, som er eksplicit udtrykt i flere tekster. De kan vurdere troværdigheden af en informationskilde baseret på fremtrædende træk. Og de kan inddrage sproglige nuancer i en tekst i deres forståelse og vurdering af teksten.	Eleverne har svært ved at læse og forstå flere lange tekster, hvor de skal sammenstille og modstille information. De har svært ved at skelne mellem indhold og formål og mellem fakta og holdning i komplekse tekster indeholdende abstrakte og kontraintuitive udsagn. Og de har svært ved at vurdere neutralitet og partiskhed hvis udsagnene, der indikerer troværdigheden ikke er eksplicitte.
Over niveau 4	Elever på niveau 5 kan også læse og forstå flere lange tekster hvor de skal sammenligne og modstille information. De kan forstå og håndtere begreber som er abstrakte eller kontraintuitive. De kan vurdere neutralitet og partiskhed baseret på eksplicitte eller implicitte tegn på troværdighed knyttet til indholdet og/eller til afsenderen af informationen. Og de kan overskue og gennemføre flere trin i deres læsning for at nå et overordnet mål. Elever på niveau 6 kan også forstå lange og abstrakte tekster, hvor den søgte information kun er indirekte til stede. De kan sammenligne, modstille og integrere information som repræsenterer flere forskellige, muligvis modsatrettede perspektiver. Og de kan drage følgeslutninger på tværs af informationer der befinder sig langt fra hinanden i teksterne.	Elever på niveau 5 har svært ved at forstå lange, abstrakte tekster, hvor den søgte information kun er indirekte til stede. De har svært ved at sammenligne, modstille og integrere information som repræsenter flere forskellige, modsatrettede perspektiver hvis informationerne de skal bruge befinder sig langt fra hinanden i teksterne.

Hvad kan måles

Der er en mere end hundredårig tradition for at teste læsning – særlig de mere tekniske aspekter af den faglige progression. Det er ikke så mærkeligt for læsning er jo grundlæggende for at tilegne sig viden inden for alle faglige områder – og læsning er i øvrigt forholdsvis enkel at teste så længe man holder sig til de mere tekniske aspekter. Men dansk er jo så meget mere end læsning. Andre vigtige spørgsmål som læreren kunne ønske svar på kunne være:

- Hvor gode er mine elever til at analysere og fortolke litteratur?
- Hvordan står det til med mine elevers skriftlig fremstillingskompetencer?
- Hvor gode er mine elever til at søge og forholde sig kritisk til det de finder på nettet?
- Hvor meget ved mine elever om det danske sprog?
- Hvor meget indsigt har mine elever i den danske kulturhistorie?
- Hvor gode er mine elever til at udvise og skabe interkulturel forståelse? Osv.

Tilsvarende kan lærere inden for andre fag pege på mange aspekter af deres faglige områder som de kunne ønske at have mere viden om, så de bedre kan tilrettelægge deres undervisning i forhold til forudsætningerne hos deres elever.

Og tilsvarende kan man pege på en lang række områder som går på tværs af fag, men som er afgørende for elevernes nutid og fremtid. Det drejer sig fx om de områder som internationalt ofte kaldes det 21. århundredes kompetencer eller tværgående (*transversal*) kompetencer:

- Hvor gode er mine elever til at få ideer og omsætte dem i løsninger?
- Hvor gode er mine elever til at samarbejde om løsning af problemer?

- Hvor gode er mine elever til at løse konflikter i samarbejde og deres sociale liv?
- Hvor gode er mine elever til at planlægge og gennemføre processer? Osv.

Alle sådanne spørgsmål kan måles gennem test – men nogle er noget lettere end andre. Det er vigtigt at tage stilling til hvad man vil måle, ikke bare ud fra overvejelser over hvad der er let tilgængeligt, men også ud fra hvad der er vigtigt at vide noget om. Uanset hvordan man vender og drejer det, så fører øget testning inden for et område, også til øget fokus på dette område.

Hvordan måles: Opgaverne

Når vi tænker på test, tænker vi nok ofte på en lang række af opgaver hvor vi enten skal skrive et enkelt ord eller tal eller hvor vi skal vælge mellem flere valgmuligheder (*multiple choice*). Sådanne test er forholdsvis lette at udarbejde og lette at vurdere. Til gengæld kan det være svært at måle mere avancerede aspekter af de kompetencer vi ønsker at vide noget om. Og det kan ærlig talt være dræbende kedeligt at svare på.

Figur 2. Rammefortælling fra Demonstrationsskoletest (Bundsgaard, 2018a)

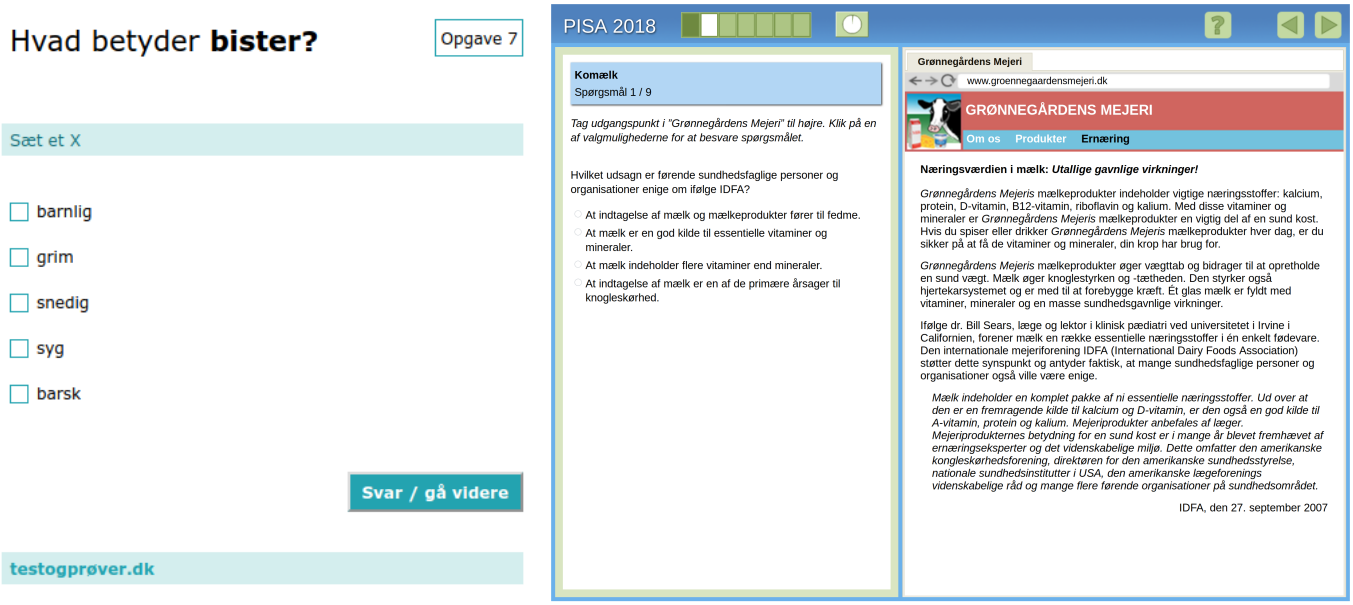
Men sådan behøver det ikke være. For det første behøver opgaverne ikke ligge som usammenhængende perler på en snor. Det har vist sig at give eleverne en mere meningsfuld oplevelse hvis opgaverne sættes ind i en fælles ramme i form af en fortælling. Et eksempel fra den kompetencetest vi udviklede til Demonstrationsskoleprojekterne, ses i figur 2. Her introduceres eleverne til rammen som er at de skal holde et loppemarked for skolens elever og deres familie. Selve opgaverne består bl.a. i at planlægge dagen, løse konflikter og til sidst udarbejde et storyboard til en reklamefilm for arrangementet.

Sådanne rammefortællinger findes både i ICILS og i PISA. I ICILS skal eleverne fx planlægge en skoleudflugt, deltage i et socialt netværk, formidle viden om åndedrætssystemet til yngre elever og arrangere en bandkonkurrence.

Opgaverne

Som sagt er opgaver i test ofte ganske simple multiple choice-opgaver hvor eleven måske skal læse en kort tekst, får et kort spørgsmål og skal vælge mellem 3-5 mulige svar. Der er ikke i sig selv noget galt med multiple choice-opgaver, men det kan som sagt blive meget ensformigt, og der er også grænser for hvad man kan teste med dette format. Multiple choice-opgaver kan dog antage meget forskellige former, og hvis der er tænkt grundigt over valgmulighederne, kan selv forkerte svar give interessante

oplysninger om elevens viden og kompetencer. Figur 3 viser to ganske forskellige eksempler på opgaver fra De Nationale Test og PISA.



Figur 3. Multiple choiceopgaver fra Nationale Test og PISA

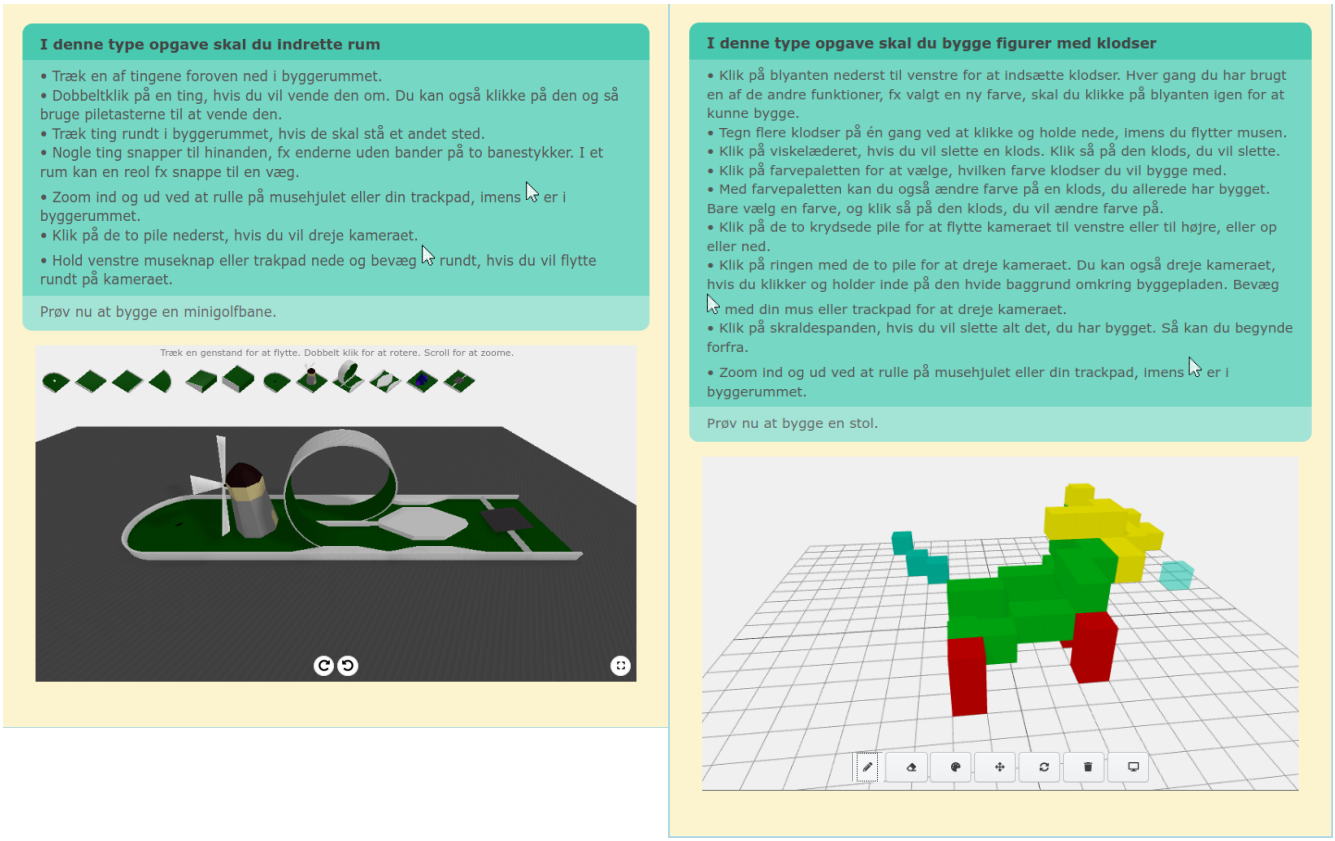
Hvis man vil have en computer til at bedømme elevernes svar, har man den udfordring at det – al hype om kunstig intelligens til trods – skal være ganske veldefineret hvad der er rigtige og hvad der er forkerte svar. Det betyder at fx svar på regnestykker og ord skrevet i en diktat godt kan gå, men så snart eleverne bliver bedt om at svare med sætninger, og man ikke bare vil vide om de er skrevet korrekt, har man udfordringer med at få computeren til at bedømme svaret. Der arbejdes i mange sammenhænge – fx i forbindelse med PISA-undersøgelserne – på at udvikle programmer der kan vurdere teksters indhold, men det fordrer for det første at man har meget tekst at "træne" computeren på, og for det andet vil der være tekster som computer ikke tilstrækkeligt sikkert kan vurdere korrektheden af, således at der stadig er behov for menneskelige øjne til at vurdere korrektheden. Men besparelsen på at måske 80 procent af besvarelserne vurderes af en computer, og de resterende overlades til mennesker, kan være betragtelig i forbindelse med en undersøgelse som PISA.

Man må derfor tage stilling til om man, arbejdsbyrden til trods, vil udforme en vis mængde opgaver som man overlader det til mennesker at bedømme. En sådan beslutning har man taget i PISA og i IEA-undersøgelserne (TIMSS, PIRLS og ICILS) hvor nogle af opgaverne for eksempel består i at svare med sætninger. I forbindelse med Demonstrationsskoleprojekterne og forskningsprojektet Game Based Learning in the 21st Century (GBL21) har vi desuden udviklet

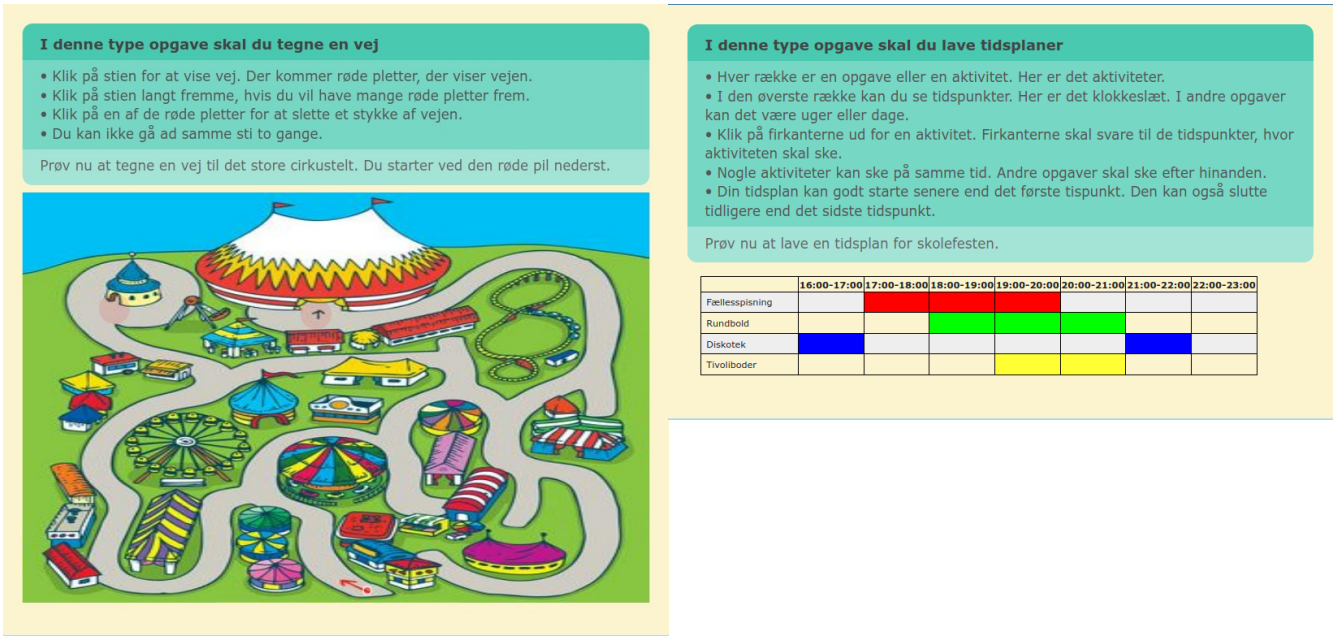
Figur 4. Eksempel på en elevbesvarelse på en opgave fra ICILS hvor eleverne skal udarbejde en profilside for et band (Bundsgaard et al., 2019).

en opgavetype hvor eleverne deltager i en brainstorm og hvor deres bidrag vurderes af mennesker. I ICILS-undersøgelsen er der sågar opgaver hvor eleverne skal udarbejde en hjemmeside, en plakat eller et slideshow (se et eksempel i figur 4). Sådanne opgaver kræver meget grundige beskrivelser af kriterier for hvad der betragtes som gode og mindre gode svar. For at sikre at der tages nuanceret stilling til alle relevante aspekter af besvarelsen, deler man ofte bedømmelsen op på flere delaspekter. I den viste opgave, skal ”koderne” som dem der bedømmer, kaldes, fx tage stilling til placering af billeder, valg af indhold, farvesammenspil med mere.

Men med computerteknologien er der utallige muligheder for at stille opgaver som er udfordrende, og som man ikke kan gætte sig til løsningen af, men som man alligevel kan få computeren til at vurdere. I figurerne 5 og 6 ses eksempler på fire sådanne opgaver fra kompetencetesten der er blevet udviklet til GBL21-projektet. Opgaverne tester for de tre første opgavers vedkommende elevernes kompetencer til rumlig forståelse og rumlig modellering og Gantt-kortopgaven tester deres kompetencer til at overskue og planlægge processer. Alle opgavetyperne kan vurderes automatisk af computeren ved fx at teste om der er valgt et passende antal elementer til minigolfbanen, om det er muligt at få bolden igennem banen osv. Nogle 3D-opgaver kan gives point automatisk – i særlig grad hvis opgaven er at tegne en figur om eller gengive en 2D-figur i 3D osv. For rute-opgavers vedkommende undersøges det om eleven følger instruktionerne, om hun vælger relevante veje (fx ikke gå på hovedvej, ikke køre på sti/gågade) og om hun finder den korteste rute som ikke indeholder uhensigtsmæssige dele.

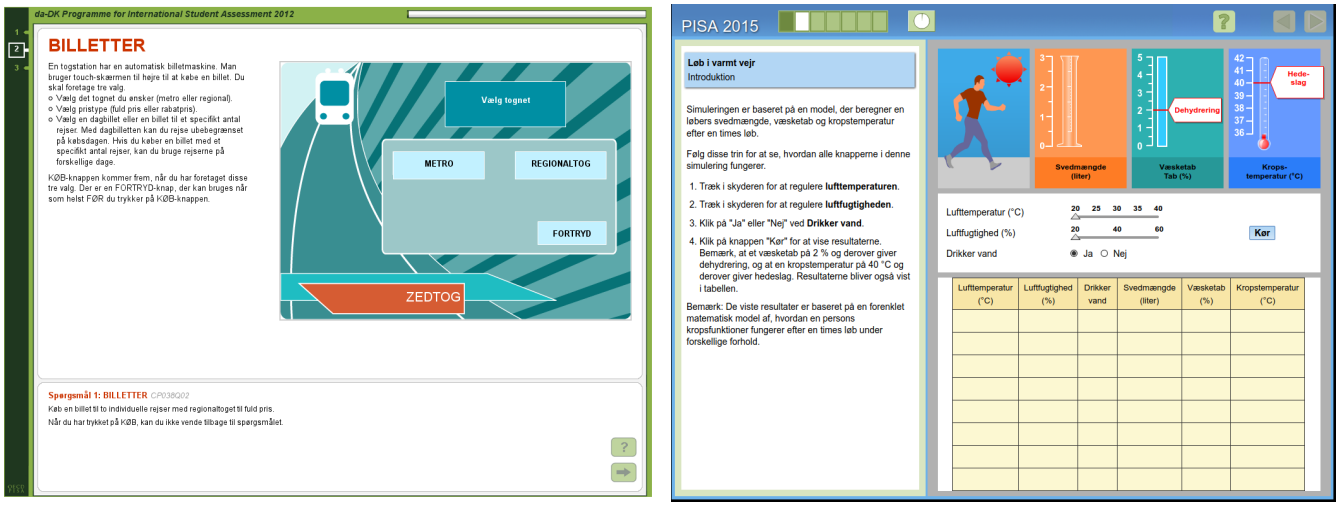


Figur 5. Skærmbilleder fra vejledningsmodulet til kompetencetesten i GBL21-projektet (Rusmann & Bundsgaard, 2019).



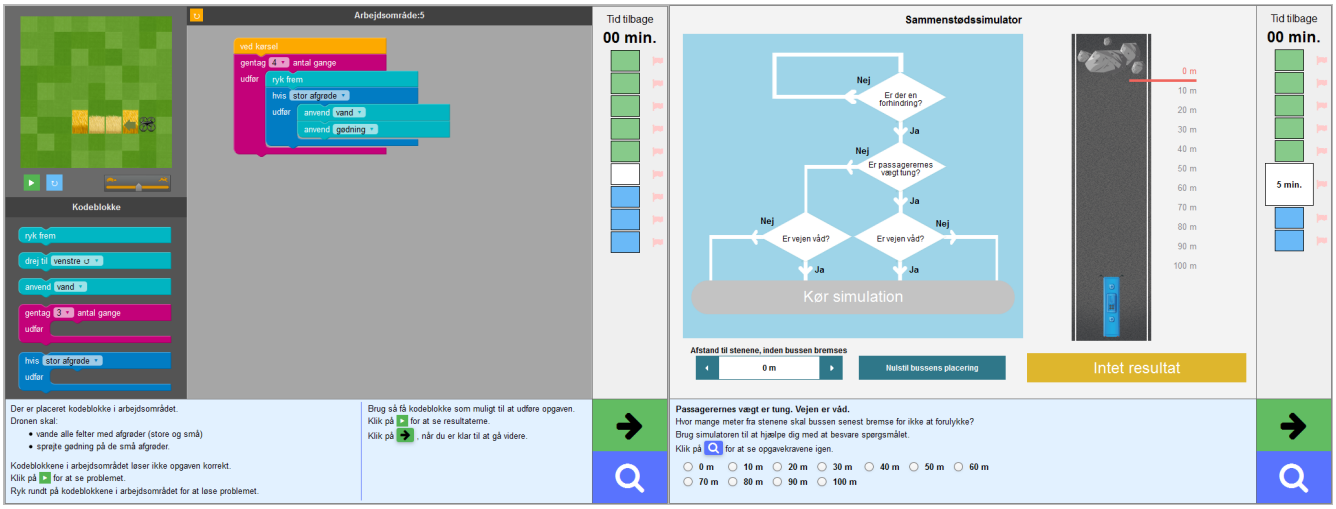
Figur 6. Skærmbilleder fra vejledningsmodulet til kompetencetesten i GBL21-projektet (Rusmann & Bundsgaard, 2019).

I figur 7 ses eksempler fra PISAs test af problemløsning og af naturfag i 2012. Disse opgaver beder eleverne om at interagere med en interaktiv teknologi – i det ene tilfælde en billetautomat og i det andet tilfælde en simulering af en løbers sved og kropstemperatur. Efter at have interageret med teknologierne, skal eleverne så svare på konkrete spørgsmål. I vurderingen af spørgsmålene indgår også registreringer af hvordan eleverne har interageret med teknologierne.



Figur 7. Skærmbillede fra PISA 2012 Problemløsning (Egelund, 2014) og fra PISA 2012 Naturfag (Egelund, 2013).

I figur 8 ses to opgaver fra ICILS 2018 hvor eleverne skal programmere en drone ved hjælp af et visuelt blokprogrammeringssprog og simulere et automatisk bremsesystem i en bus. Også disse opgaver kan vurderes automatisk af computeren.



Figur 8. Skærmbilleder fra ICILS 2018-testen af datalogisk tænkning (Bundsgaard et al., 2019).

Som det fremgår af de mange eksempler på opgaver, er det muligt at udvikle ganske avancerede og udfordrende opgaver som eleverne i vidt omfang oplever som meningsfulde og interessante at besvare. Det kendetegner også flere af disse opgaver at eleverne ikke oplever at de svarer rigtigt eller forkert.

Grænser for test

Men uanset hvor innovative opgaver man kan udvikle, så er der også grænser for hvad test kan. Vi betragter det således ifølge Folkeskolens Formålsparagraf og Fælles Mål som vigtigt at eleverne kan deltage aktivt og med medbestemmelse i skolens daglige liv, vi værdsætter at de kan argumentere for en sag i en skriftlig tekst, at de kan kommunikere mundtligt med personer fra andre kulturer, at de kan opleve og indleve sig i litteratur. Ikke alle disse ting giver det mening at teste med testopgaver som dem jeg har omtalt her. Det er således vigtigt også at give eleverne opgaver hvor de skal formulere sig på skrift i længere tekster, og hvor de skal tale deres sag mundtligt.

Og også det kan vi blive bedre til – både i forhold til at stille meningsfulde og engagerende opgaver, og i forhold til at vurdere og give mere kvalificeret feedback.

Hvordan vurderes

Den tredje byggesten i Mark Wilsons BEAR Assessment system er spørgsmålet om hvordan opgaverne vurderes. Det er afgørende for en pålidelig vurdering at det er meget præcist og grundigt beskrevet hvad der kendetegner en god besvarelse. Det er forholdsvis enkelt når der er tale om multiple choiceopgaver og andre typer af opgaver hvor der er et enkelt korrekt svar. Men selv i multiple choiceopgaver kan der være flere rigtige svar eller svar som er delvis rigtige. Hvis lærere skal bruge opgaverne, skal disse svar og begrundelserne for deres (delvise) korrekthed og forkerthed forklares.

Når der er tale om mere avancerede opgaveformater, selv om de skal vurderes af computeren, er det nødvendigt at skille delelementerne i besvarelsen ad og begrunde hvorfor der er forskel på forskellige besvarelser – altså fx hvorfor noget gives 1 point og noget andet 2. Dette er nødvendigt både af hensyn til transparensen i testen (så andre kan vurdere om det der er vurderet rigtigt, faktisk er rigtigt) og af hensyn til lærere der skal bruge (elevernes besvarelser af) opgaverne i deres undervisning.

Når opgaver skal vurderes af mennesker, skal man være endnu mere grundig. I sådanne tilfælde udarbejder testdesignerne en kodeguide som for hver opgave beskriver hvad opgaven består i, hvad ideen med den er, hvad den tænkes at teste – og derefter beskriver i detaljer hvordan forskellige typer besvarelser skal kodes. Typisk gives også eksempler på elevbesvarelser der har besvaret opgaven på forskellige måder (både forkerte og rigtige).

Her er nogle eksempler på hvori kriterierne kan bestå:

- Brainstorm: nævnes på forhånd bestemte typer indhold og aktiviteter.
- Gantt-kort: er rækkefølgen i orden, bruges der tilstrækkelig og ikke for megen tid til de enkelte opgaver.
- Udarbejdelse af multimodale tekster (hjemmesider, plakater, slideshowpræsentationer): Er billeder og tekst velvalgte, er elementerne placeret meningsfuldt, er størrelsesforhold i orden osv.

Også ved vurdering af skriftlige tekster og lignende omfattende og selvstændige produkter, kan man anvende disse principper. Her handler det om at se tegn på forskellige aspekter af kompetencer i elevernes produkter. Når en lærer således vurderer en skriveopgave, kan det være formålstjenligt at dele iagttagelserne op i delaspekter og så kun fokusere på dem i sin vurdering. Det kan fx være aspekter som:

- Stavning
- Sætningskompleksitet
- Tekststruktur
- Fremstillingsform
- Genre-overholdelse
- Målgruppefokus
- Indhold (er det rigtigt, er det nuanceret, er det velvalg og fokuseret osv.)

Hvordan udregnes resultatet

Den sidste af Mark Wilsons byggesten er spørgsmålet om hvordan resultatet af testen beregnes. En simpel – og klassisk – måde at gøre det på, er ved at tælle antallet af point og give det som resultatet.

En lidt mere avanceret, men stadig klassisk måde at gøre det på er ved at dividere antallet af rigtige med det maksimale antal mulige point. Derved fås procent rigtige – et tal som de fleste forstår. Men problemet er at ikke alle opgaver er lige svære, og hvis der er otte lette opgaver i en test og to svære, så er der meget større forskel på at få 80 og 90 procent korrekte, end der er på at få 70 og 80 procent korrekte. Dem der har 100 procent korrekte er derfor meget bedre, end dem der får 80 procent korrekte. Mens dem der får 70 procent ikke er meget bedre end dem der får 50 procent korrekte. Man kan med andre ord ikke rigtig afgøre hvor dygtige eleverne er, ud fra et procenttal.

Derfor er der udviklet mere korrekte og brugbare måder at måle dygtighed på. En af modellerne til dette er den såkaldte Rasch-model som omsætter antallet af rigtige til et tal man faktisk kan regne med og på, og hvor der er er lige langt mellem pointene på skalaen. Rasch-modellen har en række andre fordele som gør at man kan kvalitetssikre testen. Men det skal vi ikke komme nærmere ind på her.

Hvordan kommunikeres resultatet

Uanset hvilken model man vælger til at beregne dygtigheden med, så skal resultaterne formidles til folk der skal bruge disse resultater. Med de eksisterende nationale test valgte man at omregne resultaterne til såkaldte percentilniveauer der er et udtryk for hvor i fordelingen af elever, den enkelte elev ligger. Da de fleste elever er cirka lige dygtige, betyder det at der i midten af skalaen kan være ganske store forskelle i resultaterne for elever der er næsten lige dygtige – og omvendt at store forskelle i dygtighed ude i enderne, ikke ser så store ud.

Når man måler noget man ikke direkte kan røre ved eller se – som det er tilfældet med dygtighed og viden, så vil man altid kun have et estimat, ikke et præcist tal. I de første år af nationale tests levetid blev resultatet angivet uden angivelse af at der var en vis usikkerhed forbundet med resultatet. Efter en del kritik valgte ministeriet omkring 2016 at angive et såkaldt 68-procentkonfidensinterval på resultatet. Det betyder at i knap et ud af tre tilfælde vil eleven have et resultat uden for dette interval, og derfor er det mere almindeligt at angive et såkaldt 95-procentkonfidensinterval (hvor kun hver tyvende ligger uden for intervallet).

I de internationale undersøgelser (PISA, ICILS, PIRLS, TIMMS mv.) og i mange landes nationale test har man i stedet valgt at omregne Rasch-scoren til et tal hvor eleverne i gennemsnit har 500 point, og hvor de fleste elever ligger inden for 400-600 point. Dette gør det muligt at regne med tallene og at sammenligne eleverne mere korrekt.

Men uanset hvordan man vælger at angive tal – hvis man overhovedet gør det – så er det væsentligt vigtigere at formidle resultatet af testen i forhold til den faglige progression man ønsker at måle indenfor. Til det er udviklet såkaldte *proficiency scales* som kan oversættes med kompetence- eller mestringsniveauer. Sådanne mestringsniveauer beskriver hvad elever med forskellige grader af dygtighed er i stand til. I tabel 1 præsenterede jeg et uddrag af de kompetenceniveauer som PISA bruger til at præsentere hvad elever på forskellige niveauer er i stand til. Sådan en skala kan i første omgang beskrives på teoretisk grundlag, men når elever har taget den udviklede test, så kan skalaen underbygges empirisk ved at undersøge hvad der kendetegner opgaver som elever på de forskellige niveauer er i stand til at løse af opgaver (Bundsgaard, 2018b).

En forudsætning for at man kan lave en meningsfuld beskrivelse af kompetenceniveauer, er at de opgaver der er i testen, er meningsfulde fagligt set. Et af problemerne med opgaverne i nationale test er at opgaverne ikke har klart defineret og varieret fagligt indhold. Fx er det uklart hvad det betyder fagligt at man kan sætte to korrekte streger i merskumspibeselvhenterlykke, men ikke i orienteringsløbfestligholdelsesamfundsomstyrter. De opgaver som jeg har givet eksempler på i figurerne ovenfor, er derimod fyldt med fagligt indhold som kan beskrives og udfoldes og omsættes til meningsfulde faglige kompetenceniveauer. Til glæde både for lærere og for andre der har brug for viden om hvad elever på forskellige kompetenceniveauer er i stand til.

Faglig brug

Når testen er taget og resultatet er modtaget, så er det naturlige næste spørgsmål: Hvad kan jeg som lærer gøre i forhold til de forskellige grupper af elever? Kompetenceniveaubeskrivelserne er naturligvis et godt udgangspunkt for at sige: Hvis denne gruppe elever er i gang med at udvikle kompetencer på dette niveau, så vil denne type opgaver og aktiviteter være velegnede.

Men der er god grund til også at udvikle materiale der kan støtte lærerne i arbejdet med det faglige område testen tester. Det kan være lærervejledninger til hvordan testens resultater kan tolkes og bruges, med eksempler på opgaver inden for de forskellige niveauer, med forslag til initiativer der kan sættes i gang i forhold til specifikke grupper – og måske også med opfølgende tests der kan nuancere og uddybe resultaterne.

Tilsvarende kan resultaterne i form af definitionen og beskrivelsen af frameworket for den faglige progression bruges sammen med kompetenceniveau beskrivelserne i læremidler både til eleverne og til læreruddannelse og -efteruddannelse.

Og endelig kan resultaterne bruges af faglige vejledere der kan være specialister i test og kan bidrage til opkvalificering af testpraksis, tolkning og beslutningstagning på baggrund af test.

Næste generation af nationale test

Jeg har i dette positionspapir kort beskrevet hvordan kvalificeret testudvikling bygger på en række principper og bygges af en række byggesten. Erfaringerne med ”de gamle” nationale test viser med al tydelighed at det er en god ide at følge sådanne principper og være meget grundig og transparent i udviklingen af test.

Det er mit håb og min opfordring til beslutningstagere og fremtidige udbydere og udviklere af ”nye” nationale test, at sådanne principper og byggesten udgør en grundpille i udviklingen og en referenceramme for vurderingen af om testen har et tilstrækkeligt højt niveau – og at de forsøger at nå til det højeste internationale niveau i udviklingen af tidssvarende, engagerende og informative tests. Hvis det bliver tilfældet kan vi forhåbentlig nå en situation hvor både lærere og skoleledere mener at testen er både højt kvalificeret og en hjælp i det daglige arbejde med at undervise og styre og udvikle skolen.

Og til allersidst vil jeg minde om at uanset hvor gode test man udvikler, så er der behov for en meget mere nuanceret evalueringskultur. Test løser en specifik opgave i form af objektive input om elevernes kompetencer, færdigheder og viden, men der er brug for mange flere nuancer.

Litteratur

Bremholm, J., & Bundsgaard, J. (2019). Læsning i PISA 2018. I V. T. Christensen (Red.), *PISA 2018: Danske unge i international sammenligning* (s. 18–69). VIVE - Det Nationale Forsknings- og Analysecenter for Velfærd.

Bundsgaard, J. (2018a). Det 21. Århundredes kompetencer. I J. Bundsgaard, M. Georgsen, S. Graf, T. I. Hansen, & C. K. Skott (Red.), *Skoleudvikling med it: Forskning i tre demonstrationsskoleprojekter I* (s. 143–165). Aarhus Universitetsforlag.

Bundsgaard, J. (2018b). Pædagogisk brug af test. *Sakprosa*, *10*(2), 1–40.

Bundsgaard, J., Bindslev, S., Caeli, E. N., Pettersson, M., & Rusmann, A. (2019). *Danske elevers teknologiforståelse. Resultater fra ICILS-undersøgelsen 2018*. Aarhus Universitetsforlag.

Egelund, N. (2013). *PISA 2012—Danske unge i en international sammenligning*. KORA.

Egelund, N. (2014). *PISA problemløsning: Danske unge i en international sammenligning*. Dafolo.

Rusmann, A., & Bundsgaard, J. (2019). Developing a Test to Measure Design Thinking. *The Proceedings of the 13th International Conference on Game Based Learning ECGBL 2019*, *13*, 587–595. https://doi.org/DOI: 10.34190/GBL.19.071

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, *46*(6), 716–730.

26. januar 2019

POSITIONSPAPIR FRA SKOLE OG FORÆLDRE

Overordnet position

Skole og Forældre tilslutter sig, at der fortsat er nationalt udviklede, standardiserede og obligatoriske test i folkeskolen. Testene kan bruges til at følge den faglige udvikling på klasse-, skole- og nationalt niveau og til forskning omkring folkeskolen.

Formålet med nationalt udviklede faglige test

Formålet med testene er at sikre kvalitet i skolernes arbejde med faglighed i form af en standardiseret undersøgelse, der giver mulighed for at få øje på specifikke udfordringer i fagligheden på enkelte skoler eller klasser undervejs i skoleforløbet (og ikke først ved afgangseksamen). Desuden anerkender vi, at der er et politisk behov i stat og kommuner for at vide, hvordan det står til med den faglige kvalitet på landets skoler.

Skole og Forældre anbefaler

Skole og Forældre anbefaler, at nationalt udviklede test fremover indrettes til at vurdere grupper af elever, klasser, skoler og kommuner på samme niveau som fx Den Nationale Trivselsmåling, i erkendelse af at ambitionen om at forene styringsbehov med pædagogiske og didaktiske værktøjer til undervisningen har vist sig vanskelig at gennemføre med tilfredsstillende kvalitet i praksis. Dermed kan de bruges både til lokal, kommunal og national kvalitetsopfølgning, men ikke til at få viden om den enkelte elevs faglige niveau og udvikling eller til indsatser over for det enkelte barn.

Alternativt, hvis ovenstående kollektive tilgang ikke er politisk mulig, skal nationalt udviklede test gøres 100 % kompatible med formålet om at understøtte den pædagogiske og didaktiske praksis i klassen, på linje med en række af de øvrige test, der normalt anvendes som led i undervisningen i alle fag. Det vil i så fald være bedst at:

- testene gøres lineære i stedet for adaptive, så lærerens skalering i praksis og feedback for enkelte elever i klassen sker på et sammenligneligt grundlag.
- der stilles hårde grænser op for at minimere risikoen for dels udvikling af 'teaching to the test' og dels belastning af eleverne og det didaktiske rum.
- Sværhedsgrad og validitet skal være på et passende niveau.

Teaching to the test kultur skal minimeres

Skole og Forældre mener, at undervisningens situationen primært skal være et 'pædagogisk læringsrum', hvor eleverne har mulighed for at øve sig, gøre sig umage, turde fejle og udvikle sig i respekt for, at læreprocesser er komplekse, og at der kan være mange veje og forskellige mål undervejs. Derfor er det vigtigt at begrænse omfanget af teaching to the test-situationer. Vi tilslutter os derfor anbefalingen om, at indholdet i de nationalt udviklede test ligger tæt op ad fælles mål knyttet til fagbekendtgørelsen i det enkelte fag. Skole og Forældre anbefaler derfor at testenes profilområder genbesøges og sværhedsgraderne løbende opdateres eller revideres ift. fælles mål.

Pædagogiske og didaktiske værktøjer skal prioriteres
Skole og Forældre mener, at det er vigtigt, at lærerne har mulighed for at måle på elevernes faglige niveau, så de kan tilrettelægge og differentiere undervisningen i forhold til den enkelte elev, og at eleverne skal have mulighed for at kunne arbejde med deres egen faglige udvikling. Derfor anbefaler vi, at skolerne har adgang til en række pædagogiske og didaktiske værktøjer af faglig høj kvalitet, der kan bruges som udgangspunkt for lærernes tilrettelæggelse af undervisningen. Disse skal indgå meningsfuldt i forældresamarbejdet og underretningen af forældrene om elevens udbytte af undervisningen, jf. § 13 i folkeskoleloven. Desuden vil den type af værktøjer give forbedrede muligheder for, at det pædagogiske personale – i samarbejde med vejledere og ledere – har mulighed for at udvikle deres pædagogiske og didaktiske praksis.

Skolebestyrelsens grundlag for at udføre sin tilsynsrolle skal styrkes
Skole og Forældre mener, at det bør være op til den enkelte skole og skolebestyrelse at fastlægge rammer og principper – inden for lovgivningen – for brugen af disse værktøjer. Det kan evt. ske i samarbejde med den kommunale skoleforvaltning, så der sikres ensartet systematisk opfølgning på kvaliteten af kommunens skoler. Samtidig mener vi, at skoleledelsen i kvalitetsrapporter og overfor bestyrelsen skal kunne forpligtes på at orientere om, hvordan der arbejdes med konkrete pædagogiske og didaktiske indsatser i de enkelte klasser på baggrund af de nye nationale test.

Derfor bør det sikres, at skolebestyrelser er klædt på og uddannet til at varetage deres vigtige opgave med at formulere skolens fælles principper og føre tilsyn med disse, så bestyrelsernes beslutninger og vurderinger indgår i en fælles referenceramme med skoleledelsen, kommunen og staten, som træffer beslutninger omkring folkeskolen.

Underretning af forældrene om elevens udbytte af undervisningen
Forældrene har berettigede forventninger om at få god underretning fra skolen om elevens udbytte af undervisningen. Dette sker både gennem løbende dialog og samarbejde samt ved skole-hjem-samtaler. Hertil kommer skriftlige orienteringer fra det pædagogiske personale.
I erkendelse af at test, prøveresultater, adgang til læringsplatform eller elevplan aldrig kan stå alene og ikke nødvendigvis synliggør elevens faglige progression skal skriftlige underretninger/test følges op af dialog mellem skole og hjem.

Faglig trivsel i et elevperspektiv
Skole og Forældre arbejder for at styrke gode børnefællesskaber, og vi mener derfor, det er vigtigt at være opmærksom på balancen mellem fokus på individet og fællesskabet, hvis vi ønsker at skolen skal fremme alle børns alsidige udvikling og forberede dem til deltagelse, medansvar, rettigheder og pligter i et samfund med frihed og folkestyre. Vi mener, at nationalt udviklede test, der måler på kollektive praksisser, vil være med til at skabe bedre balance i det forhold og fjerne en del af det uhensigtsmæssige pres som mange børn og unge oplever.

Datasikkerhed og anonymisering
Skole og Forældre anbefaler, at data anonymiseres umiddelbart efter indsamlingen, og at der tages særlige hensyn til beskyttelse af børns data som beskrevet i persondataforordningen.

Nationale test på det specialiserede område

På de ikke-prøveafholdende skoler er der ingen alternative standardiserede test til at følge den faglige udvikling. Men også på specialområdet er der brug for at kunne følge den faglige udvikling på klasse-, skole- og nationalt niveau. Dette vil kunne hjælpe med at styrke en tidlig indsats ift. eleverne med fysiske og psykiske handicap. Samtidig vil det give forældre et kvalificeret indblik i kvaliteten af undervisningen og dermed et stærkere grundlag for en løbede dialog mellem kommuner, skoler og forældre om den specialiserede indsats.

Dette positionspapir består i en uddybende begrundelser for en række anbefalinger mht. de nationale test.

Positioner vedrørende anbefaling om generelt genberegning af sværhedsgrader

Hvis man ikke kender til detaljerne i analyser ved hjælp af Rasch modeller er det vanskeligt at gennemskue, at brugen af forskellige sværhedsgrader for opgaverne kan føre til forskellige resultater, når dygtigheden beregnes. For at begrunde anbefalingen er det derfor nødvendigt at forklare helt konkret, hvad der sker nå dygtigheden beregnes.

For at gøre det så enkelt som muligt vil jeg antage at alle opgaver er dikotome, hvor svarene scores som enten forkerte (0) eller korrekte (1), og hvor sandsynligheden for et korrekt svar på en opgave afhænger af forskellen på elevens dygtighed (θ - theta) og opgavens sværhedsgrad (β - beta).

Antag at en elev har besvaret k forskellige opgaver med sværhedsgrader lig med (β_1,…,β_k) og at eleven tilsammen har opnået S korrekte svar.

Når det drejer sig om opgaver fra den såkaldte Rasch model er det ligegyldig hvad det er for opgaver eleven har besvaret korrekt. Vurderingen af hvor dygtig eleven er er den samme, når der svares korrekt på de første S opgave og når der svares rigtigt på de sidste S opgaver. Resultatet vil være det samme.

DNT benytter det såkaldte maksimum likelihood estimat af theta som et mål for dygtigheden. For at beregne dette estimat er DNT nødt til at løse nedenstående relativt kompliceret ligning, der både afhænger af hvor mange korrekte svar eleven havde og af sværhedsgraderne på de opgaver som eleven har besvaret.

$$S = \frac{\sum_{s=0}^{k} s e^{s\theta} g_s\left(\beta_1,\ldots,\beta_k\right)}{G\left(\theta,\beta_1,\ldots,\beta_k\right)}$$

hvor $G(\theta,\beta_1,...,\beta_k) = \sum_{i=0}^{k} e^{i\theta} g_i(\beta_1,...,\beta_k)$ og hvor funktionerne $g_s(\beta_1,...,\beta_k)$ er de såkaldte symmetriske polynomier af værdierne $e^{-\beta_1}$,…, $e^{-\beta_k}$, dvs.

$$g_0(\beta_1,...,\beta_2) = 1,$$
$$g_1(\beta_1,...,\beta_2) = e^{-\beta_1}+e^{-\beta_2}+...+e^{-\beta_k} = \sum_i e^{-\beta_i}$$
$$g_2(\beta_1,...,\beta_2) = e^{-\beta_1}e^{-\beta_2}+e^{-\beta_1}e^{-\beta_3}+...+e^{-\beta_{k-1}}e^{-\beta_k} = \sum_{(i,j)} e^{-\beta_i}e^{-\beta_j}$$
…
$$g_k(\beta_1,...,\beta_2) = e^{-\beta_1}e^{-\beta_2}...e^{-\beta_k} = \prod_i e^{-\beta_i}$$

Formålet med at vise disse formler er ikke at imponere eventuelle læsere med, at jeg er i stand til at skrive uskønne formler. Formålet er kun at vise, at man skal bruge en kompliceret funktion af den samlede score og af item parametrene for at beregne dygtigheden og at det ikke er let at gennemskue, hvad der sker, hvis man kommer til at bruge nogle forkerte sværhedsgrader. Hvad der f.eks. sker hvis man bruger sværhedsgrader fra lineære afprøvninger til at beregne dygtigheden fra adaptive testforløb.

STILs notater dokumenterer at forskellen på lineære og adaptive sværhedsgrader kan være meget store i alle fag og på alle klassetrin. Heraf følger det logisk, at dygtigheden fra starten er blevet forkert beregnet. Om forskellen på beregninger af dygtigheden ved hjælp af adaptive sværhedsgrader og dygtigheden beregnet ved hjælp af lineære sværhedsgrader er store eller små siger STILs notater ikke meget konkret om, men Bundsgaards om min rapport om læsning i 8. klasse giver konkrete eksempler på forskelle som er særdeles store. Af den grund kan det kun anbefales, at STIL genberegner sværhedsgraderne, således at risikoen for systematisk forkerte beregninger reduceres. Hvordan det skal gøres afhænger af, om DNT fremover skal være et adaptivt eller lineært testsystem. Et spørgsmål som dette positionspapir ikke kan komme ind på. Uanset beslutningen skal sværhedsgraderne svare til den måde opgaverne bruges på.

I VIVEs tales der om at beregning af sværhedsgrader baseret på lineære testforløb og beregning af sværhedsgrader baseret på adaptive testforløb er to *måder* at beregne sværhedsgraderne på. At sværhedsgraderne altså må være de samme, men at de bare estimeres på vha. forskellige metoder.

Det er en fundamental misforståelse. Det er velkendt fra pædagogisk testning, at den måde en opgave præsenteres på påvirker opgavens sværhedsgrad. At der er forskel på sværhedsgraden,

hvis opgaven præsenteres i en papir-og-blyant test eller hvis de præsenteres i en IT-baseret test. Om ikke af andre grunde, så fordi læsning på papir og læsning på skærme er forskellige måder at læse på. Af den grund har teorien omkring adaptive test været fokuseret på de problemer, der opstår, når man overfører opgaver fra papir-og-blyant test til IT-baserede adaptive test. At sværhedsgraderne afhænger af den måde opgaven præsenteres på.

Det STIL har påvist er, at det samme er tilfældet med IT-baserede opgaver, der præsenteres lineært i en naturlig rækkefølge og en i meningsfuld kontekst, og IT-baserede opgaver, der præsenteres adaptivt i en ikke-naturlig rækkefølge i en mindre meningsfuld kontekst, hvor opgaver fra forskellige profilområder præsenteres i en for eleven tilfældig rækkefølge.

Både STIL og vi bruger de samme *måder* til at beregne forskellige sværhedsgrader på. VIVEs formulering er enten sjusket eller udtryk for en misforståelse.

STIL har ingen mulighed for at forklare hvorfor adaptive og lineære sværhedsgrader er forskellige. Det vil være nyttigt med en sådan viden, men mit og sikkert også STILs synspunkt er, at det er mere vigtigt at DNT fremover benytter de rigtige sværhedsgrader, og at årsagen til forskellene er en forskningsopgave som andre forhåbentlig vil tage sig af.

Positioner vedrørende anbefaling om genberegning, hvis testene forbliver adaptive

Der er to grunde til at de eksisterende resultater for DNT giver et sløret billede af den faglige udvikling. Den ene er, at beregningerne er behæftet med systematiske fejl pga. anvendelsen af sværhedsgrader fra lineære test, som STILs analyser viser er forkerte når der er tale om adaptive testforløb.

STILs analyser antyder, at problemerne er størst for de dygtige elever og for sværhedsgraderne af de vanskelige opgaver, og det kan derfor forventes at det især af udviklingen af færdigheder blandt de dygtigste elever, hvor udviklingen vil være sløret. Hvorvidt denne antagelse er korrekt kan man bruge tid på at undersøge nærmere, men anbefalingen er et forslag om ikke at spilde tid på dette og i stedet genberegne dygtigheden på den rigtige måde. En måde at løse problemet der både er mere overkommelig og mere sikker.

Forskningsresultater baseret på forkerte tal for dygtigheden kan også være behæftet med fejl. Hvorvidt de berørte forskere skal reanalysere deres data med de korrekte tal for dygtigheden må være op til forskerne. Jeg finder det naturligt at gøre det, men anbefalingen inkluderer ikke noget krav om det.

Positioner vedrørende anbefaling om forbedring af DNT's præcision

VIVES evaluering konstaterer, at der ikke er noget der tyder på at DNT måler mindre præcist end andre test men VIVE kunne med samme ret have konkluderet at der ikke er noget der tyder på, at der er andre test der måler lige så upræcist som DNT. VIVE har - formodentlig pga. tidspresset - ikke fundet nogle konkrete oplysninger om, hvor godt eller dårligt andre test måler.

I stedet kunne man have henvendt sig til forskere, som har arbejdet med og været med til at udvikle pædagogiske test. Nedenstående tabel viser, hvad undertegnede uden at bruge mere end en times tid kunne have trukket frem. Tabellen indeholder både tal for den bedste præcision, som testene kan levere (søjle 3 med "Target SEM") og den gennemsnitlige præcision i de elevpopulationer, hvor testene er blevet afprøvet. I samtlige tilfælde er præcisionen langt bedre end i DNT og i mange tilfælde lige så præcise, som man oprindelig ønskede, at DNT skulle være. I betragtning af at man valgte at DNT skulle være adaptiv for at opnå den højest mulige grad af præcision, kan det ikke være tilfredsstillende at alle eksemplerne på lineære test er meget mere præcise end DNT.

Udover tallene for præcision, indeholder tabellen også de samme tal for reliabiliteten, som STIL beregner.

Target SEM = value of SEM where SEM is minimized
Average SEM = average SEM in the study population

Test specific information				Population/sample dependent			
Test	Number of items	Max Score	Target SEM	Average SEM	Reliability	Test-retest reliability	PSI
PISA 2009 math.	22	40	0.33	0.36	0.86	0.87	0.87
PISA 2006 reading	9	33	0.36	0.40	0.82	0.83	0.83
CHIPS - bhkl	40	40	0.32	0.36	0.85	0.81	0.85
CHIPS - kl2	40	40	0.32	0.36	0.85	0.85	0.85
CHIPS - kl5	40	40	0.32	0.42	0.81	0.86	0.77
Matematikprofilen 1	42	69	0.22	0.26	0.80	0.80	0.73
Matematikprofilen 2	40	68	0.24	0.29	0.90	0.91	0.88
Matematikprofilen 3	40	68	0.22	0.24	0.85	0.86	0.84
RoS/Test kl4	40	40	0.31	0.37	0.90	0.90	0.87
RoS/Test kl5	40	40	0.31	0.40	0.90	0.90	0.89
RoS/Test kl6	40	40	0.31	0.44	0.90	0.91	0.90

STILs notater indeholder en række muligheder man kunne forsøge sig med. Nedenstående liste indeholder de muligheder, som jeg ville forsøge mig med, hvis jeg skulle løse opgaven. Listen indeholder formodentlig muligheder, som STIL også selv ville nævne.

Første mulighed er at lade brugeren (læreren) afgøre, hvor DNT skal vælge opgaver i starten af forløbet for at sikre, at de svageste og stærkeste elever ikke får for mange ikke-informative opgaver i starten. Mig bekendt har STIL selv undersøgt denne mulighed og har konkluderet at gevinsten som helhed er meget beskeden. Det er korrekt, men der vil være en gevinst for de allersvageste og de allerstærkeste, hvis man startede testforløbet som foreslået her. For de svageste vil det også løse problemet med en urimelig og alt for udfordrende start på testene. Dette betyder ikke noget for de stærkeste, men her kan man i stedet pege på at STILs analyser afslører, at problemet med sikkerheden er størst for de dygtigste. Set med mine øjne er der kun fordele ved at implementere denne mulighed, så den bør forsøges.

Anden mulighed er at undersøge om der rent faktisk er belæg for at beregne tre forskellige profilområder inden for hvert fag. STIL har allerede taget de første skridt til at foretage en empirisk vurdering af belægget for at have tre usikre mål i stedet for et eller to mere sikre mål, men er ikke komme ret langt.

Dette arbejde må intensiveres. Hvis der ikke kan påvises mere end en eller to bagvedliggende latente variable, skal der udvikles en samlet Rasch model for hver af de profilområder, som måler forskellige aspekter af den samme færdighed. Det vil føre til en væsentlig forbedring af sikkerheden. Det skal samtidig bemærkes, at det ikke behøver at ændre noget ved hvordan de adaptive eller måske lineære test i øvrigt fungerer på. Man kan og bør stadig vælge opgaver i algebra og opgaver i geometri som tidligere for at sikre indholdsvaliditeten, men sværhedsgraderne skal kalibreres i forhold til hinanden, så man efterfølgende kan beregne et samlet estimat af dygtigheden.

Mit forslag er i øvrigt at man gør dette i samarbejde med fagfolk (dvs. opgavekommisionerne). Det første må være at spørge dem om de insisterer på at de forskellige profilområder måler kvalitativt forskellige ting og at de af den grund vil modsætte sig at profilområderne slås sammen. Jeg kan forestille mig begge dele. At man f.eks. vil insistere på at sprogforståelse og tekstforståelse er helt forskellige (om end statistisk korrelerede) færdigheder. Men jeg forventer også også det

modsatte, fordi beslutningen om *præcis* tre profilområder i *alle* fag og på *alle* klassetrin aldrig har været fagligt motiveret.

Eller med andre ord: Det må være på tide at stoppe med at snakke om det. STUK og STIL bør sætte sig sammen og løse denne opgave.

Tredje mulighed er at intensivere brugen af polytome-testlet opgaver, hvor der stilles flere spørgsmål til det samme emne. Der er to forskellige årsager til at det er en god ide. Den første er, at det er mulig at stille flere enkelt spørgsmål på denne måde og måske også lettere at konstruere sådanne opgaver. Den anden årsag er, at vi tidligere har vist, at polytome opgaver er mere informative end et tilsvarende antal dikotome opgave, og i visse tilfælde meget mere informative.

Det er værd at bemærke at STILs resultater tydeligt viser at dette er korrekt. Der forekommer flere polytome opgaver i fysik, og usikkerheden er dernede, hvor man oprindelig gerne ville have den. For mig, er der ingen diskussion. Jeg ville intensivere brugen af polytome opgaver i alle fag og gøre endnu mere ud af at finde ud af hvilke former for polytome opgaver, der er mest intensive.

Jeg forudser, at den fjerde mulighed måske ikke er spiselig for ministeriet, men den bør alligevel nævnes.

Hvis ovenstående muligheder ikke slår til er der kun en udvej. Man bør fjerne et eller flere profilområder fra de obligatoriske test. Beslutningen om hvad der kan undværes og hvad der ikke kan undværes i de obligatoriske test skal træffes af fageksperter i samarbejde med brugere, men ikke af STIL.

Bemærk, at udeladelsen af et profilområde i forbindelse med de obligatoriske test ikke betyder, at de også skal udelades af de frivillige test. Jeg vil på den ene side forvente, at der er nogen, der vil mene, at det ikke har mening at teste alle elever i 8. klasse i afkodning, og som ville foretrække mere sikre målinger af tekstforståelse. At man af den grund ville fjerne afkodning på dette klassetrin.

På den anden side vil der også være situationer, med meget svage læsere, hvor det vil være nyttigt for læreren at få noget at vide om problemet bl.a. skyldes problemer med afkodningen. I sådanne tilfælde skal læreren naturligvis have lov til at bede om frivillige læsetest, hvor fokus er på afkodningen og ikke på tekst- og sprogforståelse. Forslaget om at fjerne et profilområde fra de obligatoriske test er ikke et forslag om at fjerne et profilområde fra de frivillige test. Det skal kun ske hvis fageksperterne siger at det er en fuldstændig misforståelse at profilområdet overhovedet er dukket op.

Positioner vedrørende anbefaling om erstatning af percentil-scoren og den kriteriebasering

Brugen af percentil-scores til at beskrive den faglige udvikling i Danmark er uhensigtsmæssig, fordi store fremskridt blandt de bedste elever og de svageste elever kun afspejler sig i små fremskridt på perecentil-skalaen. I stedet for at benytte percentil-skalaen bør man benytte en lineær transformation af den såkaldte logit-skala fra Rasch modellen, således som PISA og andre internationale undersøgelser gør det.

Det er sådanne skalaer, der er de rigtige at bruge til styringsmæssige formål, men argumentet mod dem er, at mange oplever, at det er vanskeligt at tolke betydningen af tal på logit-skalaen. At de er uanvendelige til pædagogiske formål. I stedet bør man udvikle kriteriebaserede kompetence-beskrivelser, som tolker værdierne af logit-skalaerne på en måde, der gør dem nyttige i pædagogiske sammenhænge. Ministeriet har allerede foretaget det første skridt i en sådan retning, men ministeriets kriteriebaserede skala kan og skal forbedres, hvis tanken er den skal bruges pædagogisk.

Positioner vedrørende at begrænse afrapportering på individniveau til de tilfælde, hvor dygtigheden er målet med stor sikkerhed

Målinger med for stor SEM på elevniveau kan godt benyttes til ministeriets opgørelser på populationsniveau. At beslutte ikke at fortælle læreren og forældrene hvad dygtigheden ser ud til at være pga. usikkerhed har ingen styringsmæssige konsekvenser.

Med hensyn til de fejlslagne forløb, er der stadig et spørgsmål, der skal besvares. Hvor stor en andel af testforløbene er karakteriseret ved lange kæder af fejlsvar på opgaver, som eleven burde kunne klare, og hvor det er tydeligt at slutresultatet handler om noget andet end den færdighed, der måles. Bundsgaard og mine analyser viser at de forekommer, men ikke hvor hyppigt. Før der taget stilling til om det er et problem også på populationsniveau og hvad der skal gøres ved det, bør STIL foretage analyser, der kortlægger problemets omfang.

Positioner vedrørende anbefaling om professionelle standarder for testudvikling

Standarder for pædagogiske test.

Hvis ministeriet levede op til de standarder, som findes beskrevet i beskrevet i "STANDARDS for Educational and Psychological Testing" ville en lang række af problemer som praktikkerne peger på have været undgået. Jeg kan kun anbefale at man læser dem og tager dem alvorligt.

Oversigten over krav til udbydere af test definerer standarder for følgende punkter

- Validity
- Reliability/Precison and errors of Measurement
- Fairness of testing
- Test Design and Development
- Scores, Scales, Norms, Score Linking, and Cut Scores
- Test Administration, Scoring, Reporting, and Interpretation
- Supporting Documentation for tests
- The Rights and Responsibilities of Test takers.
- The Rights and Responsibilities of Test
- Psychological Testing and assessment
- Workplace Testing and Credentialing
- Educational Testing and Assessment
- Uses of Tests for Program Evaluation, Policy Studies, and Accountability

Nedenstående eksempler der stilles til seriøse udbydere af test, en rolle som ministeriet (STUK og STIL) har påtaget sig. Eksemplerne er ikke på nogen måde et udsagn om at jeg generelt mener at ministeriet ikke lever op til alle krav eller at de skal leve op til alle krav, men de er et udsagn om at der er punkter, hvor standarden af DNT klart kunne forbedres. Mange af eksemplerne relaterer sig i øvrigt logisk til mange af de vores anbefalinger. Anbefalinger, der ville have været overflødige, hvis man havde sat sig bedre ind i det der kræves af pædagogiske test.

Standards for Scores, Scales, Norms, Score Linking, and Cut scores.

Cluster 1. Interpretation of Scores.

Standard 5.1

Test users should be provided with clear explanations of the characteristics, meaning, and intended interpretation of scale scores, as well as their limitations.

Standard 5.2

The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly.

Standard 5.3

If there is sound reason to believe that specific interpretations of a score scale are likely, test users should be explicitly cautioned.

Standard 5.5

When raw scores or scale scores are designed for criterion-referenced interpretation, including the classification of examinees into separate categories, the rationale for recommended score interpretation should be explained clearly.

Standard 5.6

Testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability on which scores are reported.

Cluster 4. Cut Scores.

Standard 5.21

When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.

Standard 5.23

When feasible and appropriate, cut scores defining categories with distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to relevant criteria.

Educational testing and Assessment

Cluster 1. Design and development of Educational Assessment

Standard 12.1

When educational testing programs are mandated by school, district, state. or other authorities, the ways in which test results are intended to be used should be clearly described by those wo mandate the test. It is also the responsibility of those who mandate the use tests to monitor the impact and to identify and minimize potential negative consequences as feasible. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test developer and/or user.

Standard 12.2

In educational settings, when a test is designed or used to serve multiple purposes, evidence of validity, reliability/precision, and fairness should be provided for each intended use.

Standard 12.4

When a test is used as an indicator of achievement in an instructional domain or with respect to specified content standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domain should be provided. Both the tested and the target domains should be described in sufficient detail for their relationship to be evaluated. The analyses should make explicit those aspects of the target domains that the test represent, as well as those aspects that the test fails to represent.

Standard 12.5

Local norms should be developed when appropriate to support test users' intended interpretations.

Standard 12.6

Documentation of design, models, and scoring algorithms should be provided for tests administered and scored using multimedia or computers.

Cluster 2. Use and Interpretation of Educational Assessments

Standard 12.7

In educational settings,, test users should take steps to prevent test preparation activities and distribution of materials to students that may adversely affect the validity of test score inferences.

Standard 12.10

In educational settings, a decision or characterization that will have major impact on a student should take into consideration not just scores from a single test but other relevant information.

Standard 12.11

When differences or growth scores are used for individual students, such scores should be clearly defined, and evidence of their validity, reliability/precision, and fairness should be reported.

Standard 12.12

When an individual student's scores from different tests are compared, any educational decision based on the comparison should take into account the extent of overlap between the two constructs and the reliability or standard error of the difference score

Standard 12.14

In educational settings, those who supervise others in test selection, administration, and score interpretation should be familiar with the evidence for the reliability/precision, the validity of the intended interpretations, and the fairness of the scores. They should be able to articulate and

effectively train others to articulate a logical explanation of the relationships among the tests used, the purposes served by the tests, and the interpretations of the test scores for the intended uses.

Standard 12.15

Those responsible for educational testing programs should take appropriate steps to verify that the individuals who interpret the test results to make decisions within the school context are qualified to do so or are assisted by and consult with persons who are so qualified.

Cluster 3. Administration, Scoring, and reporting of Educational Assessment

Standard 12.16

Those responsible for educational testing programs should provide appropriate training, documentation, and oversight so the individuals who administer and score the test(s) are proficient in the appropriate test administration and scoring procedures and understand the importance of adhering to the directions provided by the test developer.

Standard 12.17

In educational settings, reports of group differences in test scores should be accompanied by relevant contextual information, where possible, to enable meaningful interpretation of the differences. Where appropriate contextual information is not available, users should be cautioned against misinterpretation.

Standard 12.18

In educational settings, score reports should be accompanied by a clear presentation of information on how to interpret the scores, including the degree of measurement error associated with each score or classification level, and by supplementary information related to group summary scores. In addition, dates of test administration and relevant norming studies should be included in score reports.

Standards for uses of test for program evaluation, policy studies, and accountability

Cluster 1. Design and Development of Testing Programs and Indices for Program Evaluation, Policy Studies, and Accountability Systems

Standard 13.2

When change or gain scores are used, the procedures for constructing the scores, as well as their technical qualities and limitations, should be reported.

Standard 13.3

When accountability indices, indicators of effectiveness in program evaluations or policy studies, or other statistical models are used, the method for constructing such indices, indicators, or models should be described and justified, and their technical qualities should be reported.