Odense, 9. januar 2019

Jeppe Bundsgaard

Vedr. Evalueringen af nationale test januar, 2020.

Konklusioner som bør fremgå tydeligt af den tværgående sammenfatning

- Items i nationale test måler i alt for stort omfang forkert. Mere end halvdelen af alle items i dansk læsning og matematik har sværhedsgrader der afviger mere end en halv logit fra de i testen anvendte sværhedsgrader (STIL notat 4, s. 9).

- Betydelige andele (15-20 procent) link-items ændrer sværhedsgrad på tværs af lineære afprøvninger (STIL notat 4, s. 6). Det betyder at der er endnu færre link-items som faktisk anvendes til brobygningen mellem gamle og nye items.

- Fejl i item-sværhedsgrader har indflydelse på fordelingen af elever på en måde så det er synligt i fordelingsplots (STIL notat 4, s. 5, samt bilag 4.4). *Der er derfor grund til at tvivle på alle resultater fra analyser og forskning der hviler på nationale test.*

- For store andele (mellem 2 og 13 procent, STILs notat 3, s. 4) af eleverne er usikkerheden større end den fastsatte grænse.

- Usikkerheden (SEM på 0,55) er alt for stor til at elevernes resultater kan bruges på individniveau – og sandsynligvis også på klasse- og sågar årgangsniveau (egne analyser).

- Vi ved ikke noget om hvor mange elevers testforløb der ikke stemmer med Rasch-modellen. De lave test-gentest-korrelationer (STILS notat 3, s. 13) tyder på at det er betydelige andele.

- Der foregår en udbredt praksis med *teaching-to-the-test*: Fx for dansk: 35,6 procent af dansklærerne målretter undervisning frem mod obligatoriske test, 28,2 procent opprioriterer de områder der testes i, og 22,9 procent målretter undervisningen for at opnå gode resultater (bilagstabellerne 3.23-3.26)

- Betydelige andele af lærerne mener ikke at nationale test tester relevante kompetencer (60 procent af dansklærerne, 30 procent af matematiklærerne, 45 procent af engelsklærerne, 40 procent af fysik/kemi-lærerne, bilagstabellerne 3.16-3.19).

- Betydelige andele af skoleledere og lærere svarer ’ved ikke' på spørgsmål om hvordan data fra nationale test skal fortolkes (bilagstabeller 1.7 og 3.28. Forvaltningsmedarbejdere er ikke stillet sådanne spørgsmål, så vi ved ikke om også de mangler forståelse for hvordan data skal fortolkes.

- 46 procent af skolelederne mener at det er en god ide at afskaffe nationale test, og 64 procent af skolelederne mener at viden fra nationale test let kan afløses af anden viden.

- Nationale test opleves generelt ikke at understøtte skolernes evalueringskultur (VIVEs rapport 5, s. 10)

Konklusioner i VIVEs afrapportering der savner belæg eller er usande

Dette er en oversigt over de mest grelle fejl og misforståelser. Citaterne stammer overvejende fra rapporten med de tværgående konklusioner. Der er tilsvarende fejl i de øvrige rapporter som bør rettes.

> *s. 9: De er således ikke udtryk for elevers fulde kunnen inden for et fag, men et udtryk for deres kunnen inden for de områder, der testes i. Og det ved praktikerne godt.*

Der er ikke belæg for at (alle) ”praktikere” er klar over dette. Der er derimod belæg i delrapport 5, s. 70 for at undervisning tilpasses til testene: ”I nogle tilfælde er intentionen at identificere mønstre i, hvad eleverne i en klasse har vanskeligt ved, hvilket kan fungere som afsæt for en drøftelse af mulige tiltag på klasseniveau.”

> *s. 10: Evalueringen viser, at de nationale test er usikre, når det kommer til den enkelte elevs resultat. Usikkerhed på elevniveau er forventeligt blandt lignende test. Der findes dog meget lidt viden om, hvor usikre andre test er på elevniveau, da området er relativt uudforsket.*

Det er ikke korrekt. SEM har været et centralt begreb i testteori i mindst 100 år. SEM udregnes rutinemæssigt ved Rasch-modeller. Nationale tests SEM er langt over såvel anbefalinger (jf. reviewerkommentarene) og almindelige SEM-niveauer ved tests hvor resultatet rapporteres på enkeltindividniveau. Udsagnet kan fint erstattes med et udsagn om at der ikke er noget der tyder på at der er nogen test der måler så dårligt som DNT.

> *s. 10: Det vil sige, at den tvivl, mange har om, hvorvidt man meningsfuldt kan anvende data på aggregeret niveau, når nu data er usikre for den enkelte elev, bør være afklaret. Det kan man godt, men selvfølgelig inden for de rammer, som er gældende for data af denne type.*

Der er ingen beregninger af usikkerhed på aggregeret niveau i rapporten, og dette er derfor en påstand helt uden belæg. Mine beregninger tyder på at gennemsnittet for en klasse vil have en SEM på omkring 0,2 (afhængig af profilområde), og for en årgang (3 klasser) vil ligge omkring 0,1. Det vil sige at en gennemsnitlig klasse vil have et konfidensinterval mellem nogle og 40 og nogle og 70 (fx 44-75), og en gennemsnitlig årgang vil have konfidensinterval mellem nogle og 40 og nogle og 50 (fx mellem 40-54).

Der er således *ikke* grundlag for at sige at data kan bruges på aggregerede niveauer uden angivelse af konfidensintervaller.

I denne påstand er der også set stort på at nationale test ifølge STILs beregninger måler forkert.

> *s. 10. Nogle elever oplever testsituationen positivt, mange oplever den som neutral og få oplever den negativt.*

Der er ikke indsamlet surveydata fra elever, så der er ikke noget belæg for udsagn om få eller mange. Der er interviewdata, men de kan ikke anvendes til udsagn om kvantitative forhold. Fra lærersurvey ved vi til gengæld at 80,1 procent af lærerne er helt eller overvejende uenige i at eleverne er glade for

nationale test og 63,2 procent af lærerne er helt eller overvejende enige i at nationale test giver eleverne en problematisk oplevelse af nederlag (bilagstabel 3.22).

> *s. 10. Testens adaptive princip er med til at forkorte testens længde, da det gør det muligt hurtigere at finde elevens niveau. Så en afskaffelse af det adaptive princip vil alt andet lige kræve en længere test for at opnå et lige så præcist resultat.*

Der er ikke gennemført undersøgelser af dette, og derfor ikke belæg for påstanden (det er korrekt i teorien, men har det betydning i praksis?).

> *Det vanskeliggør anvendelsen som et enkeltstående testresultat i det pædagogiske arbejde, om end lærerne oftest oplever, at elevernes resultater stemmer overens med lærerens opfattelse af elevens faglige niveau.*

Det er ikke korrekt at "lærerne" oplever at resultatet stemmer overens med deres opfattelse. Der er sådanne udsagn i interviewene, men ikke noget belæg for at det skulle gælde generelt. Med den store usikkerhed vil lærerne i øvrigt tage fejl, hvis de tror det.

> *s. 11. På styringsniveau og som ledelsesinformation er data dog pålidelige med høj ekstern validitet. Data bidrager særligt på kommunalt og nationalt niveau som et værdifuldt styringsredskab administrativt og i mindre grad politisk.*

Der er ikke beregnet usikkerhed på aggregerede niveauer, så denne konklusion har ikke belæg. Der er ikke givet argumenter for at der er tale om høj ekstern validitet. Der er korrelationer mellem DNT og andre målinger, men de er ikke høje.

> *s. 11. Evalueringen viser klart, at der er behov for data, der kan bruges pædagogisk af lærerne i skolerne, og data, der kan bruges som styringsredskab på højere niveauer. Og hvis de nationale test afskaffes, så vil der være behov for at udvikle et eller flere nye redskaber til at dække disse behov.*

Der er ikke belæg for at DNT bidrager med relevant data. Tværtimod indgår DNT ifølge skolederne kun i mindre omfang til at følge med i om skolen når sine mål og værdier (40 procent siger i temmelig høj eller høj grad, mens andre instrumenter (karaktergennemsnit, trivselsmåling, andre standardiserede test) alle indgår ifølge (mere end) halvdelen (hhv. 49,3, 88,0 og 67,2 procent) (bilagstabel 1.23).

> *s. 15. Dansktesten tester således i Fælles Mål-termer udelukkende kompetenceområdet læsning – og konkret halvdelen af dette kompetenceområdes seks færdigheds- og vidensområder. De tre områder fra Fælles Mål, der dækkes, svarer 1-1 til testens tre profilområder.*

Dette er ikke korrekt. Der er mange delelementer af de tre færdigheds- og vidensområder som ikke testes. Se et bud på hvad der faktisk testes, nederst i dette notat. VIVE har alene optalt testudviklernes angivelse af hvilke områder de mener at teste med de enkelte items. VIVE har ingen mulighed for at vide hvad der faktisk testes, da de ikke har undersøgt de konkrete items.

> *s. 18. Det er imidlertid dansklærerne, som vurderer de nationale test i dansk som mindst anvendelige sammenlignet med nationale test i andre fag. Deres skepsis kan hænge sammen med, at de nationale test i dansk ikke tester hele faget men alene delelementer.*

Der er ikke noget i evalueringen der underbygger denne hypotese.

> *s. 18. Særligt forvaltningschefer og til en vis grad skoleledere oplever, at de nationale test er et vigtig styringsredskab, som muliggør en dialog og opfølgning inden for og på tværs af kommuner og skoler.*

Det er blot 37 procent af skolelederne som er overvejende eller helt enige i at nationale test bidrager med viden der har stor værdi. 46 procent af skolelederne mener til gengæld at det er en god ide at afskaffe nationale test, og 64 procent af skolelederne mener at viden fra nationale test let kan afløses af anden viden. Påstanden ovenfor er således usand. Det er i øvrigt ikke kun forvaltningschefer der har svaret på spørgsmål til forvaltningen, men også konsulenter og andre medarbejdere.

> *s. 20. Nationale test opleves generelt ikke at understøtte skolernes evalueringskultur. Undersøgelsen peger på, at det hænger sammen med, at lærere og skoleledere er skeptiske over for validiteten af nationale test.*

Der er ikke noget i evalueringen som understøtter denne hypotese.

> *s. 23. For eksempel afhænger Standard Error of Measurement (SEM) af standardafvigelsen på en test, som igen afhænger af den skala, der måles på.*

Dette er ikke korrekt. VIVE forveksler sandsynligvis viden om klassisk testteori med Rasch-modellen.

> *s. 23. Ligeledes vil en test-retest kunne foretages på mange forskellige samples, som i større eller mindre omfang vil have betydning for korrelationen. Det vil sige, at det ikke uden en væsentligt dybere analyse er muligt at sammenligne de danske nationale tests reliabillitet med de fundne tests reliabilitet. Ud fra den information, der er indhentet på de 11 test, er der ikke noget, der tyder på, at de danske nationale test har en dårligere reliabilitet end andre test.*

Der er alene fundet oplysninger fra to andre tests, og disse er kun angivet på logit-skalaen for den ene. Denne tests SEM er tilsyneladende lavere end DNT. Der er således ikke belæg for påstanden. Dette burde være undersøgt ordentligt. Fx er Smarter Balanced omkring 0,3, PISA omkring 0,4, talblindhed omkring 0,3, Gyldendals webprøver omkring 0,25. Altså i alle tilfælde lavere end DNT (bemærk at PISA ikke opgiver resultater på individniveau).

> *s. 10f. Som pædagogisk redskab på klasseniveau er der bedre muligheder for at anvende data [dette er en uunderbygget påstand]. Men der er uklarhed om, hvordan man omsætter den viden, som testene potentielt bidrager med, til praksis. Uklarheden kan både bygge på manglende viden, manglende tid, og at diskussionen om testenes validitet har fyldt så meget, som den har.*

Der er ikke belæg (heller ikke i evalueringsrapporterne) for at testene ville kunne bruges pædagogisk – og derfor er den mest rimelige hypotese ikke at der mangler viden, tid eller at der har været diskussion. Det er at DNT ikke kan bruges pædagogisk og derfor ikke bliver det.

> *s. 11 (se også s. 18). Den eksterne validitet er også med til at forhøje den informationsværdi, skoleledelserne kan have, for de ledere, der formår at forene deres styring med den pædagogiske praksis.*

Der er ikke noget belæg for at der er høj ekstern validitet, og der er ikke noget belæg for at denne ikke-underbyggede validitet ville give højere informationsværdi. Der er ikke i evalueringen undersøgt om der er forskel på skoleledere der ”formår” at forene styring med pædagogisk praksis, og dem der ikke gør.

> *s. 11. Hvis man ikke afskaffer de nationale test, er der behov for at arbejde med reliabiliteten og den interne validitet samt med at gøre det nemmere for lærere og skoleledere at arbejde konstruktivt med testene – eksempelvis gennem bedre vejledninger og mere handlingsorienteret oversættelse af data til pædagogisk anvendelse – ligesom der bør arbejdes med fortællingen om, hvad de nationale test egentlig kan og skal måle, og hvad de ikke kan og skal måle.*

Der er ikke noget i evalueringen der understøtter påstande om at vejledningerne kunne være bedre. Der er arbejdet med vejledninger igennem alle de ti år testen har eksisteret. Tværtimod er lærere og skoleledere ganske tilfredse med disse (bilagstabel 1.30 og 3.63). Man kan ikke vejlede sig ud af at der ikke er noget at bruge testene til pædagogisk og ledelsesmæssigt på skoleniveau. VIVEs forslag savner belæg i den gennemførte evaluering.

> *s. 17. Der er ikke belæg i analysen for at sige, at de nationale test skaber hverken mere eller mindre ubehag eller glæde hos eleverne end andre test.*

Dette er et uunderbygget udsagn da der ikke er gennemført sammenligninger med andre undersøgelser. En rå analyse af data fra ICILS 2018 viser at 53 procent af de danske elever synes nationale test tester noget relevant, mens 58 procent synes at ICILS-testen tester noget relevant og 66 procent af eleverne synes at test som lærerne har taget med i klassen tester noget relevant. I samme undersøgelse svarer 11 procent at det en spændende udfordring at svare på spørgsmålene eller løse opgaverne i nationale test, mens det er 30 procent af eleverne der synes ICILS var en spændende udfordring, og 20 procent som synes test deres lærere tager med i klassen, er en spændende udfordring.

Det faglige indhold

Der er ikke gennemført undersøgelser af det faglige indhold, men alene af opgavekommisionernes opmærkning af hvilket fagligt indhold de mener et givet item relaterer sig til. En simpel undersøgelse af dansk baseret på de tilgængelige items viser at kun små dele af læsning testes, og at sprogforståelsesitems ikke har noget modsvar i Fælles Mål. Se nedenstående figur.

> *s. 15. En del af forklaringen kan være, at testformatet i de nationale test (it-baseret og multiple choice) ikke egner sig til at teste kompetencer og kun i nogen grad færdigheder. Disse dele af fagene dækkes derfor enten slet ikke eller i lav grad af testene. Det drejer sig eksempelvis om områder som ’Kommunikation’ og ’Modellering’.*

Det er muligt at teste matematiske kompetencer it-baseret og (delvis) multiple choice – det sker i både TIMSS og i PISA. Kompetencer testes desuden i fx ICILS, Smarter Balanced, GBL21-forskningsprojektet og mange andre Rasch-skalerede tests.

Dansk

Færdigheds- og vidensmål (Læsning)

Klassetrin	Kompetencemål	Faser	Færdigheds- og vidensmål											
			Finde tekst		Forberedelse		Afkodning *		Sprogforståelse		Tekstforståelse		Sammenhæng	
Efter 2. klassetrin	Eleven kan læse enkle tekster sikkert og bruge dem i hverdagssammenhænge	1.	Eleven kan vælge en tekst ud fra et mindre udvalg	Eleven har viden om teksters sværhedsgrad	Eleven kan forberede læsning gennem samtale i klassen	Eleven har viden om måder til at skabe forforståelse	Eleven kan læse ord i tekster til klassetrinnet sikkert	Eleven har viden om bogstavernes kontekstbetingede udtale	Eleven kan identificere ukendte ord i tekst og tale	Eleven har viden om ord og udtryk i instruktioner og opgaver	Eleven kan gengive hovedindholdet af tekster til klassetrinnet	Eleven har viden om fortællende og informerende teksters struktur	Eleven kan forbinde tekstens emne med egen viden, erfaring og ideer	Eleven har viden om samspil mellem tekst og læser
		2.	Eleven kan finde tekster ved at navigere på alderstilpassede hjemmesider	Eleven har viden om sideopbygning på hjemmesider	Eleven kan anvende enkle førlæsestrategier	Eleven har viden om enkle førlæsestrategier		Eleven har viden om stavemåde og betydning af ord i tekster til klassetrinnet	Eleven kan forstå betydningen af indholdsord i kontekster	Eleven har viden om forskelle og ligheder i ords betydning	Eleven kan kombinere tekst og baggrundsviden til at skabe sammenhængende forståelse	Eleven har viden om samspillet mellem teksterns informationer og læserens viden	Eleven kan forholde sig til teksters emne	Eleven har viden om enkle refleksions-spørgsmål
			Finde tekst		Forberedelse		Afkodning *		Sprogforståelse		Tekstforståelse		Sammenhæng	
Efter 4. klassetrin	Eleven kan læse multimodale tekster med henblik på oplevelse og faglig viden	1.	Eleven kan navigere ud fra søgespørgsmål på aldersvarende hjemmesider og på biblioteket	Eleven har viden om hjemmesiders struktur	Eleven kan strukturere sin baggrundsviden	Eleven har viden om metoder til strukturering af viden	Eleven kan læse ord i tekster til klassetrinnet hurtigt og sikkert	Eleven har viden om regler for sammensætning af ord	Eleven kan anvende ordbøger og opslagsværker til afklaring af ords betydning	Eleven har viden om funktion og opbygning af opslagsværker og ordbøger	Eleven kan identificere elementer i teksten, som skaber sammenhæng	Eleven har viden om tekstbånd	Eleven kan samtale om teksters budskaber	Eleven har viden om teksters påvirkende funktion
		2.	Eleven kan vurdere hjemmesiders relevans i forhold til søgespørgsmål	Eleven har viden om enkle kildekritske metoder på internettet	Eleven kan formulere enkle læseformål	Eleven har viden om oplevelseslæsning og faglig læsning		Eleven har viden om ordklasser og regler for bøjning af ord	Eleven kan anvende over- og underbegreber til at skabe sammenhængende forståelse af teksten	Eleven har viden om over- og underbegreber	Eleven kan håndtere problemer med at forstå teksten	Eleven har viden om læserforståelses-strategier	Eleven kan gengive sin forestilling om teksterns situationer og sammenhænge	Eleven har viden om visualiseringsformer
			Finde tekst		Forberedelse		Afkodning		Sprogforståelse *		Tekstforståelse		Sammenhæng	
Efter 6. klassetrin	Eleven kan læse og forholde sig til tekster i faglige og offentlige sammenhænge	1.	Eleven kan vurdere relevans af søgeresultater på søgeresultatsider	Eleven har viden om søgerelaterede læsestrategier	Eleven kan orientere sig i teksterns dele	Eleven har viden om rubrikker, billeder, diagrammer og grafik	Eleven kan læse ukendte ord ved umiddelbar genkendelse af de mest almindelige orddele	Eleven har viden om morfemer i danske ord	Eleven kan anvende overskrifter og fremhævede ord til at skabe forståelse af tekster	Eleven har viden om ord og udtryk, der forklarer nyt stof	Eleven kan gengive hovedindholdet af fagtekster	Eleven har viden om fagteksters struktur	Eleven kan vurdere teksters perspektiv på et emne	Eleven har viden om metoder til sammenligning af teksters perspektiver
		2.	Eleven kan gennemføre billed- og fuldtekstsøgning	Eleven har viden om teknikker til billed- og fuldtekstsøgning	Eleven kan sammenholde teksters formål og indhold med læseformål	Eleven har viden om teksters formål og om læseformål	Eleven kan læse ukendt ord i fagtekster	Eleven har viden om stavemåde og betydning af ord i fagtekster	Eleven kan udlede dele af ords betydning fra konteksten	Eleven har viden om ordforståelses-strategier	Eleven kan anvende grafiske modeller til at få overblik over teksters struktur og indhold	Eleven har viden om grafiske modeller	Eleven kan vurdere teksters anvendelighed	Eleven har viden om kriterier for teksters anvendelighed
			Finde tekst		Forberedelse		Afkodning		Sprogforståelse		Tekstforståelse		Sammenhæng	
Efter 9. klassetrin	Eleven kan styre og regulere sin læseproces og diskutere teksters betydning i deres kontekst	1.	Eleven kan kildekritisk vurdere bruger- og ekspertproduceret indhold	Eleven har viden om afsenderforhold og genrer på internettet	Eleven kan vurdere teksters afsender og målgruppe	Eleven har viden om afsenderforhold og målgruppe	Eleven kan variere læsehastighed bevidst efter læseformål og ordkendskab i teksten	Eleven har viden om sammenhæng mellem ordgenkendelse og læsehastighed	Eleven kan vurdere teksters sproglige virkemidler	Eleven har viden om sproglige virkemidler	Eleven kan sammenfatte informationer fra forskellige elementer i teksten	Eleven har viden om tekstelementers opbygning og funktion	Eleven kan sætte tekster ind i sammenhæng	Eleven har viden om sammenhæng mellem tekst og kontekst
		2.	Eleven kan planlægge og gennemføre faser i informationssøgning	Eleven har viden om faser i informationssøgning	Eleven kan skaffe sig overblik over multimodale teksters opbygning	Eleven har viden om genretræk og multimodalitet	Eleven kan læse komplekse danske og lånte ord hurtigt og sikkert	Eleven har viden om morfemer i låneord	Eleven kan vurdere betydningen af ord og begreber i relation til teksters opridelse	Eleven har viden om sociolekter og formelt sprog	Eleven kan sammenfatte informationer fra flere tekster	Eleven har viden om metoder til sammenstilling af informationer fra flere tekster	Eleven kan diskutere mulige udfald af situationer beskrevet i tekster	Eleven har viden om metoder til opstilling af scenarier
		3.	Eleven kan gennemføre målrettet og kritisk informationssøgning	Eleven har viden om kildekritisk søgning	Eleven kan afgøre, hvordan en tekst skal læses	Eleven har viden om førlæsestrategier	Eleven kan læse komplekse tekster hurtigt og sikkert	Eleven har viden om stavemåde og betydning af ord i alle tekster	Eleven kan anvende ord og udtryks betydning til at forstå komplekse tekster	Eleven har viden om ordvalgets betydning for budskabet	Eleven kan forstå komplekse tekster	Eleven har viden om metoder til vurdering af teksters formål og perspektiv	Eleven kan kritisk vurdere teksters udsagn på baggrund af kontekst	Eleven har viden om metoder til systematisk undersøgelse af tekster

Testes (måske)

Testes delvis/måske ikke

Testes (måske) ikke

Testes (måske), men under tekstforståelse