De nationale test måler udmærket og bidrager til vigtige forskningsresultater

Som forskere inden for uddannelsesområdet finder vi det glædeligt, hvis de nationale test kan gøres endnu mere præcise. Men i en ny evaluering af de nationale test giver Jeppe Bundsgaard og Svend Kreiner en misvisende fremstilling af de nationale tests målesikkerhed og drager urimelige konklusioner heraf.

"Forskning baseret på de nationale test – kan vi stole på den?", spørger Jeppe Bundsgaard og Svend Kreiner i deres rapport om de nationale tests måleegenskaber. I rapporten, og i en kronik i Politiken d. 2. april konkluderer de, at det kan vi ikke uden fornyet revision. Ifølge de to forskere er resultaterne af de nationale test behæftet med usikkerhed og fejl, blandt andet fordi opgavernes sværhedsgrad har ændret sig, siden sværhedsgraden blev fastlagt i 2010 og 2014. Konsekvensen er ifølge de to forskere, at de nationale test må droppes, og at "alle beslutninger foretaget på baggrund af nationale tests må tages op til revision. (…) Og det gælder forskningsresultater, som bygger på de nationale tests."

Som forskere inden for uddannelsesområdet vil vi gerne slå fast, at vi sætter stor pris på enhver analyse, der kan bidrage til at forbedre de nationale test. Bundsgaard og Kreiner har udført et stort stykke arbejde med henblik på at vurdere, om de nationale test regner rigtigt, når de måler børns læseniveau i 8. klasse. Vi værdsætter dette arbejde, men vi mener af flere grunde, at deres konklusioner er forhastede og i nogle tilfælde helt forkerte. Vi er ikke tilhængere af test for testenes skyld, men de nationale test måler evner hos eleverne, som har tæt sammenhæng med, hvordan de klarer sig videre i uddannelsessystemet. Og forskning baseret på de nationale test har bidraget til at give ny vigtig viden om for eksempel betydningen af frikvarterer, indsatser for tosprogede elever, effekten af at have to lærere i klassen og meget mere. Netop derfor er det vigtigt at gå Bundsgaard og Kreiners konklusioner og analyser efter i sømmene for at vurdere, hvor holdbare de er, og hvad de betyder for anvendelsen af de nationale test i pædagogisk og forskningsmæssigt øjemed.

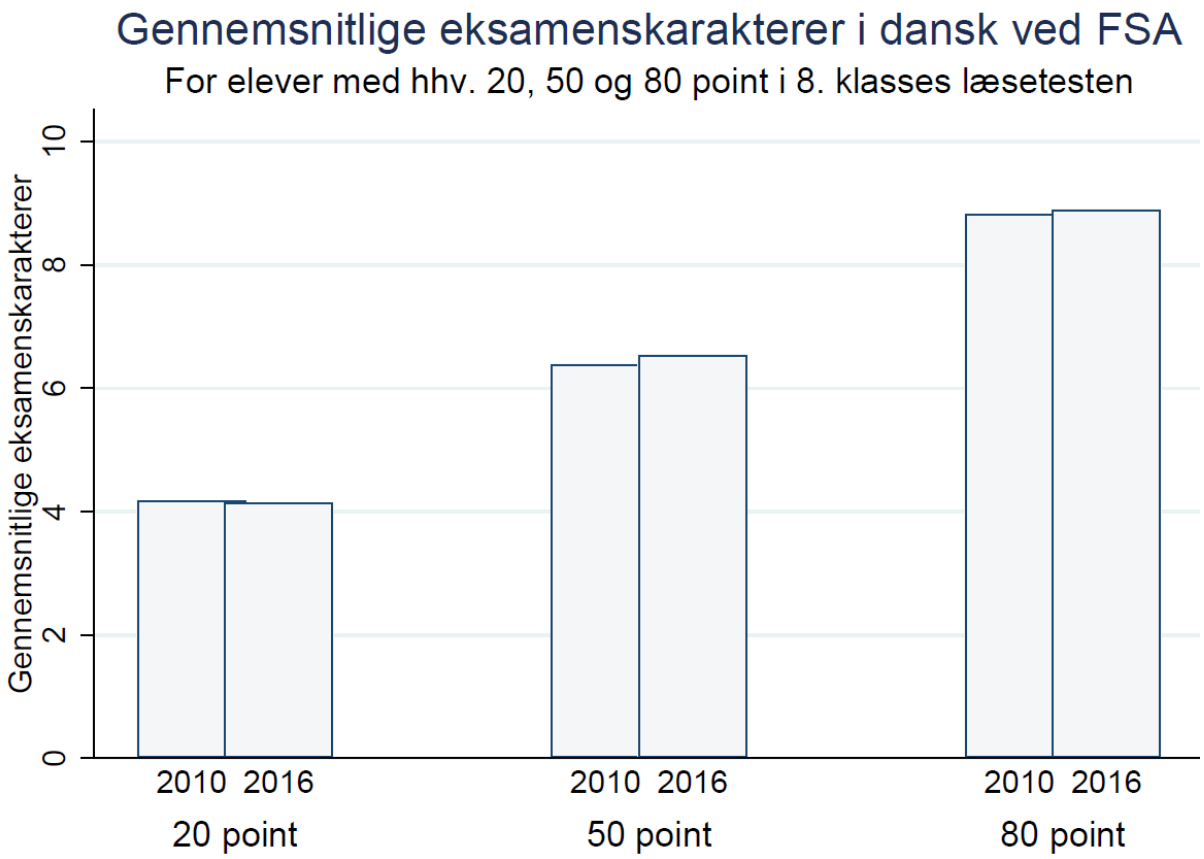
EN AF BUNDSGAARD OG KREINERS vigtigste konklusioner er, at de "nationale test for mange elevers vedkommende ikke måler så præcist som lovet". Det er en ofte fremført kritik af de nationale test, at en elev kan tage den samme test to gange uden at få det samme resultat. Dette gælder imidlertid enhver pædagogisk test. Det overrasker næppe nogen, hvis en golfspiller, som spiller den samme bane flere gange, ikke bruger præcist lige mange slag på hver runde. Det burde heller ikke overraske folk med kendskab til test af elevers dygtighed, at det samme gælder de nationale test.

Bundsgaard og Kreiners fremstilling af den statistiske usikkerhed er imidlertid misvisende. De skriver nemlig: "Når læreren får at vide, at en elev scorer 50 point på en skala fra 1 til 100, så er usikkerheden så stor, at man med sikkerhed kun kan sige, at resultatet ligger mellem 20 og 80". Denne påstand er fundamentalt forkert på flere måder. Man kan ikke med sikkerhed sige, at resultatet ligger mellem 20 og 80. Eleven, som scorer 50, ligger med størst sandsynlighed på 50. Det er rigtigt, at eleven kunne score 20 – eller 19 for den sags skyld – det er bare ikke nær så sandsynligt. Bundsgaard og Kreiner giver således det indtryk, at testresultatet på 50 lige så godt kunne være 20 eller 80. Det er forkert. En test vil sandsynligvis aldrig ramme helt præcist, men testresultaterne viser, hvad der er det mest sandsynlige niveau, og jo længere væk fra testresultatet, man kommer, jo mindre sandsynligt er det, at elevens faktiske niveau befinder sig der.

I figuren har vi beregnet eksamensgennemsnit i dansk for elever i 9. klasse opdelt på deres resultater i de nationale læsetest i 8. klasse. Figuren viser, at elever, der scorede 20, havde et eksamensgennemsnit på cirka 4, dem, der scorede 50, havde et snit på godt 6, og dem, der scorede 80, havde et snit på lidt under 9. Der er altså en stærk sammenhæng mellem, hvordan folkeskoleelever klarer sig til læsetesten i 8. klasse, og hvordan de klarer sig til eksamen året efter.

Dette skulle man ikke forvente, hvis de nationale tests var så dårlige, som de beskyldes for. Endvidere viser figuren, at denne sammenhæng er cirka den samme, uanset om vi bruger de nationale test fra 2010 eller dem fra 2016, (den seneste årgang som har været til eksamen i 9. klasse) – selvom Bundsgaard og Kreiner hævder, at de nationale test er blevet meget dårligere i perioden. Resultaterne i figuren betyder ikke, at alle elevers færdigheder er målt præcist. Det er de ikke, og det er ønskværdigt at få nedbragt den usikkerhed gennem en forbedring af testen. Men sammenhængen med eksamenskaraktererne, som jo bedømmes af lærer og censor på et bredt udsnit af danskfaget, ville ikke være så stærk som i figuren, hvis elevernes resultater i de nationale test var mere eller mindre tilfældige, eller hvis de målte forhold, som kun har marginal relevans for dansk-faget generelt.

Så når Bundsgaard og Kreiner konkluderer, at de nationale test skal stoppes øjeblikkeligt, svarer det til at sige, at alle golfturneringer skal spilles om, fordi vi ikke kan være helt sikre på, at hver spiller altid har fået præcist den placering, der svarer til spillerens reelle niveau. Men ligesom i golf er sandsynligheden for at vinde turneringen markant højere, hvis du er dygtig til spillet, selvom den bedste spiller ikke vinder hver runde.



EN ANDEN KONKLUSION, som Bundsgaard og Kreiner drager, er: ”At resultater fra forskning og evaluering, der har taget udgangspunkt i data fra nationale test, tages op til fornyet undersøgelse.” Uden at angive præcist hvori problemet består, mistænkeliggør de dermed et stort antal forskningsartikler, som flere af os har bidraget til. Vel at mærke forskningsartikler, der er publicerede

i anerkendte, internationale tidsskrifter, hvor de har været igennem grundig bedømmelse fra redaktører og anonyme fagfæller, der hver især er udvalgt ud fra deres ekspertise til at vurdere kvaliteten af forskningen.

Forskningsresultaterne tager højde for, at der altid er en vis måleusikkerhed forbundet med de enkelte elevers resultater. Dette sker ved at teste, om forskelle mellem forskellige grupper af elever er så store i forhold til den usikkerhed, der er på testresultaterne, at det er meget usandsynligt, at forskellene skyldes tilfældigheder. Det kaldes statistisk signifikanstest og er en standardpraksis.

Ligeledes sammenligner mange forskningsartikler resultater for grupper af elever indenfor samme skoleår. Ændringer i testens måleegenskaber over tid, eller mislykkede testforløb, som rammer nogle elever tilfældigt, vil derfor ikke påvirke forskningsresultaterne. Derfor er det en urimeligt generaliserende konklusion, at alle forskningsresultater skal tages op til revision. Selv hvis Bundsgaard og Kreiner havde ret i kritikken af måleusikkerhed og fejl, er mange forskningsresultater upåvirkede af det, fordi de tager højde for måleusikkerheden og ikke studerer emner, der har med de mislykkede testforsøg at gøre.

EN TREDJE KONKLUSION i rapporten er, at opgavernes sværhedsgrader er forkerte i de nationale test. Det er en vigtig kritik, da opgavernes sværhedsgrader bruges til at udvælge passende opgaver til hver elev og dermed til at beregne, hvor dygtige eleverne er. Desværre fremgår det ikke klart af rapporten, hvordan forfatterne præcist har beregnet opgavernes sværhedsgrad, og det bliver let meget teknisk at forklare de nationale tests såkaldt adaptive princip i denne kronik. Men tilsyneladende tager Bundsgaard og Kreiner ikke højde for, at de dygtigste elever har fået de sværeste opgaver, og de mindre dygtige elever har fået de letteste opgaver. Hvis ikke man tager højde for dette, kommer de sværeste opgaver til at se lettere ud, end de er, og de letteste opgaver kommer til at se sværere ud, end de er. Det er netop den konklusion, som Bundsgaard og Kreiner når frem til, men det kan altså muligvis skyldes Bundsgaards og Kreiners beregninger og ikke fejl i testsystemets beregninger.

DET ER SOM NÆVNT PRISVÆRDIGT, at der til stadighed arbejdes på at optimere testene. Fra vores perspektiv er det væsentligste problem ved at bruge de nationale test imidlertid ikke, at der er måleusikkerhed på resultaterne. Det er der på alle test. Problemerne opstår, hvis de nationale test bliver tillagt alt for stor betydning. Eksempelvis fortæller Politiken d. 4. april meget følelsesladet om en dygtig pige, som fik 11,2 i snit i gymnasiet, men som desværre stadig er påvirket af en oplevelse i 7. klasse, hvor hun fik sit resultat fra den nationale test: ”Kroppen blev kold, øjnene slørede, hovedet blev tungt, og blodet suste ned til tæerne. (…) Har man lyst til at åbne sin skæbne foran andre mennesker?” Hvem har givet pigen det indtryk, at de nationale test afgør hendes skæbne?

I stedet bør man tage testene for hvad, de er: Én blandt mange forskellige måder at følge elevernes udvikling på, som giver brugbar viden om elevernes kompetencer inden for centrale områder som afkodning og sprogforståelse set i forhold til andre elever på samme klassetrin. De i alt cirka 10 timer, som en elev bruger på at tage de 10 obligatoriske nationale test fra 2. til 8. klasse, udgør en forsvindende lille del af de omkring 7.000 timer, som eleverne går i skole den periode.

VI MENER DERFOR, at det er forhastet og potentielt skadeligt, når Bundsgaard og Kreiner ønsker testene standset med øjeblikkelig virkning. Der er snarere brug for en grundig gennemgang af deres egen analyse, så det bliver tydeligt, hvilket forskningsmæssigt belæg der er for deres konklusioner, og hvilke slutninger det er rimeligt at drage på grundlag heraf. Derudover sidder Bundsgaard og Kreiner begge med i den rådgivningsgruppe, der støtter Undervisningsministeriet i den

igangværende evaluering af de nationale test, og det er oplagt af afvente resultaterne af dette arbejde, inden der konkluderes på de nationale tests fremtidige skæbne.

Hvis konklusionen på dette bliver, at der er behov for at forbedre de nationale test, anbefaler vi, at man afprøver et alternativ, før man sætter de nuværende test på standby. Der er nemlig så mange modstridende ønsker til de nationale test i den aktuelle debat, at man næppe kan udvikle et testredskab, der opfylder alle ønsker. For eksempel vil det koste på måleusikkerheden eller kræve længere prøver, hvis testen skal måle flere elementer af danskfaget, eller hvis man dropper den adaptive mekanisme, som nogle har ytret ønske om.

Det er meget muligt, at der kan findes gode og brugbare alternativer til de nuværende nationale test på sigt. I mellemtiden har vi med de nationale test faktisk et redskab, der giver udmærkede målinger af elevernes dygtighed, og samtidig bidrager til at skabe stærke danske forskningsresultater, der kommer samfundet og ikke mindst børn og unge til gavn.

Simon Calmar Andersen, professor, Institut for Statskundskab, Aarhus Universitet.

Dorthe Bleses, professor, Institut for Kommunikation og Kultur, Aarhus Universitet.

Anna Piil Damm, professor, Institut for Økonomi, Aarhus Universitet.

Miriam Gensowski, adjunkt, Økonomisk Institut, Københavns Universitet.

Maria Koch Gregersen, ph.d., Institut for Økonomi, Aarhus Universitet.

Mette Gørtz, lektor, Økonomisk Institut, Københavns Universitet.

Thorbjørn Sejr Guul, adjunkt, Institut for Statskundskab, Aarhus Universitet.

Ahmad Hassani, ph.d., Institut for Økonomi, Aarhus Universitet.

Eskil Heinesen, forskningsleder, ROCKWOOL Fonden.

Jakob Majlund Holm, adjunkt, Institut for Statskundskab, Aarhus Universitet.

Maria Humlum, lektor, Institut for Økonomi, Aarhus Universitet.

Ulrik Hvidman, adjunkt, Institut for Statskundskab, Aarhus Universitet.

Anders Højen, lektor, Institut for Kommunikation og Kultur, Aarhus Universitet.

Morten Jakobsen, lektor, Institut for Statskundskab, Aarhus Universitet.

Peter Jensen, professor, Institut for Økonomi, Aarhus Universitet.

Eva Rye Johansen, ph.d., Institut for Økonomi, Aarhus Universitet.

Sarah Yde Junge, ph.d., Institut for Statskundskab, Aarhus Universitet.

Rasmus Landersø, seniorforsker, ROCKWOOL Fonden.

Morten Hjortskov Larsen, adjunkt, Institut for Statskundskab, Aarhus Universitet.

Elena Mattana, adjunkt, Institut for Økonomi, Aarhus Universitet.

Anne Nandrup, postdoc, Institut for Økonomi, Aarhus Universitet.

Helena Skyt Nielsen, professor, Institut for Økonomi, Aarhus Universitet.

Søren Nielsen, ph.d., Institut for Økonomi, Aarhus Universitet.

Alexander Paul, adjunkt, Institut for Økonomi, Aarhus Universitet.

Lars Qvortrup, professor, Danmarks institut for Pædagogik og Uddannelse, Aarhus Universitet.

Astrid Würtz Rasmussen, lektor, Institut for Økonomi, Aarhus Universitet.

Michael Rosholm, professor, Institut for Økonomi, Aarhus Universitet.

Hans Henrik Sievertsen, adjunkt, Departement of Economics, University of Bristol.

Marianne Simonsen, professor, Institut for Økonomi, Aarhus Universitet.

Niels Skipper, lektor, Institut for Økonomi, Aarhus Universitet.

Nina Smith, professor, Institut for Økonomi, Aarhus Universitet.

Kim Sønderskov, professor, Institut for Statskundskab, Aarhus Universitet.