Udtalelse vedrørende rapporten om "Undersøgelse af De Nationale Tests måleegenskaber".

Claus Holm har bedt mig lave en udtalelse om læsning af kronikken "Drop kritikken af de nationale tests" ændrer eller ikke ændrer min vurdering af soliditeten af de konklusioner, som Jeppe Bundsgaard og Sven Kreiner kommer frem til i deres rapport.

Bundsgaard og Kreiner konkluderer side 71:

a) At de item sværhedsgrader, som de nationale test bruger, ikke svarer til sværhedsgraderne i 2017.
b) At dygtighedsgraderne for nogle elevers vedkommende estimeres forkert.
c) At nationale test for mange elevers vedkommende ikke måler så præcist som lovet.
d) At en for stor andel af elevernes testforløb ikke stemmer med forudsætningerne i Rasch-modellen.
e) At algoritmen udviser tegn på stiafhængighed (eller stopper for tidligt i processen).

Det korte svar på Claus Holms spørgsmål er, at kronikken overhovedet ikke ændrer min vurdering af Bundsgaard og Kreiners rapport eller konklusionerne i rapporten – af den simple grund, at kronikken kun meget sporadisk berører analyser og konklusioner i rapporten. I det omfang konklusionerne berøres argumenteres der så overfladisk, at man kommer i tvivl om det er bevidst fra de 31 forfatteres side eller et udtryk for begrænset kendskab til testområdet og psykometri (f. eks. henvisningen til, at enhver pædagogisk test er behæftet med usikkerhed).

De 31 forfattere synes først og fremmest optaget af at forsvare deres egen forskning baseret på nationale test – selv om det må være indlysende, at vurderingen af de nationale test primært må fokusere på deres kvalitet som instrumenter til at vurdere den enkelte elev (Bundsgaard & Kreiners rapport fokuserer da også på denne problemstilling). De 31 forfattere må være klar over, at argumentet om, at statistisk signifikanstest er standard ved undersøgelse af forskelle mellem grupper af elever, er irrelevant i forhold til problemstillinger, der vedrører usikkerheden ved den enkelte elevs resultat (faktisk ser man ikke så sjældent en sammenblanding af standard error of the mean og standard error of measurement, men man må gå ud fra de 31 forskere er opmærksomme på denne forskel). I øvrigt fremgår det af rapporten, at ikke alle målefejl kan betragtes som tilfældige, men at der kan være tale om systematiske fejl, som også kan medføre bias i sammenligning af grupper af elever – f. eks. grupper fra forskellige årgange.

Bundsgaard og Kreiner beskriver i rapporten pædagogisk og detaljeret en analyse af dansk/læsning i 8. klasse (det er urimeligt og vildledende, når der i kronikken påstås, at der er brug for en grundig gennemgang af analyserne i rapporten – påstanden antyder, at der i rapporten er tale om overfladiske analyser). Analysen er således begrænset i forhold til de nationale tests anvendelse på andre klassetrin og som test af matematik og andre fagområder. Forfatterne gør selv opmærksom på denne begrænsning, men henviser side 72 med rette til en tabel med tidligere undersøgelser af retest korrelationer for klassetrin mellem 2. og 8. klasse og dansk/læsning, matematik og andre fag (engelsk, fysisk biologi og geografi). Korrelationerne ligger mellem 0.41 og 0.81, og de laveste korrelationer er klart for lave for test, som anvendes til at vurdere den enkelte elev – problemet består altså ikke bare i, at "der altid er en vis usikkerhed knyttet til et testresultat" – det er et spørgsmål om hvor stor usikkerheden er, og det ved kronikkens forfattere formentligt godt. Da der er tale om nationale test, kan man også med god grund spørge, om de højeste korrelationer er optimale sammenlignet med andre pædagogiske test eller typer af psykologiske tests (for visse intelligenstest har man fundet retest korrelationer på op til 0.97). For mig at se antyder tabellen alvorlige generelle problemer med målepræcisionen, som sikkert kan henføres til det forhold, at den adaptive testalgoritme stopper ved en SEM på 0.55 og ikke som først planlagt 0.30.

I kronikken er fokus imidlertid ikke på den enkelte elev, og forfatterne viser i en figur gennemsnitlige sammenhænge mellem testscores i 8. klasse og karakterer i dansk ved eksamen i 9. klasse. Den gennemsnitlige sammenhæng anvendes som argument for, at de nationale test ikke kan være så dårlige som de beskyldes for. At der findes gennemsnitlige sammenhænge kan imidlertid ikke overraske, idet hverken Bundsgaard & Kreiner eller andre har påstået, at de nationale test giver forkerte testresultater for alle elever - tværtimod antyder retest korrelationerne og procenterne i tabellerne side 72-73, at de nationale test rammer nogenlunde korrekt for en ganske stor del af eleverne. At testen rammer rigtigt for nogle elever skal imidlertid ikke få os til at overse, at den rammer forkert for en uacceptabel stor del af eleverne.

De 31 forskere er i kronikken nærmest belærende omkring fortolkningen af konfidensintervaller (det virker nedladende i forhold til en kapacitet som Svend Kreiner). Rapporten illustrerer imidlertid i en tabel på side 73 den procentuelle andel af elever, hvor forskellen mellem estimaterne ved to testninger ligger uden for 95-procent konfidensintervallet. Procenterne er så høje, at et eller begge testresultater må være groft misvisende, og det må på landsbasis gælde tusindvis af elever, hvis disse test anvendes gennem mange år. Dette er et forhold, som de 31 forskere efter min mening burde forholde sig til, og som selvfølgeligt skal undersøges nøjere.

Jeg har gennem årene beskæftiget mig med mange forskellige typer af psykologiske test og er – i f.eks. kliniske sammenhænge – ofte stødt på test, der ikke er afprøvet på tilstrækkeligt store grupper til, at der gennemføres en systematisk psykometrisk analyse og validering. Dette må ofte accepteres som et vilkår i lille sprogområde som det danske, da omfattende afprøvning af en test kræver ganske store økonomiske ressourcer. Det er imidlertid min opfattelse, at der skal stilles helt andre krav til afprøvning, analyse og monitorering af test, der får status af obligatoriske nationale test. Bundsgaard & Kreiners rapport er et første skridt til sådanne analyserne af de nationale test, og da kronikken tydeligvis er inspireret af rapporten, må man spørge om kronikkens titel i virkeligheden betyder, at de helst ikke vil have, at der gennemføres relevante psykometriske analyser, som potentielt kan stille spørgsmål ved noget af grundlaget for deres egen forskning. En sådan holdning er sjælden hos forskere, og den bliver ikke mindre problematisk af, at de 31 forskere i den grad ser bort fra, at de fejl og mangler, som Bundsgård og Kreiner påpeger, faktisk kan medvirke til forkerte beslutninger vedrørende tusindvis af danske skolebørn.