Tine Nielsen. Cand.psych, ph.d., lektor/seniorforsker i afdelingen for anvendt forskning i pædagogik og samfund, UCL

Jeg har på foranledning af Folkeskolen.dk haft lejlighed til at læse og udtale mig omkring Svend Kreiners (SK) artikel "om statistiske analyser af resultater fra pædagogiske test" bragt i Folkeskolen 24. august 2023. Jeg vil indledningsvist sige, at jeg i det store og hele er enig i SKs betragtninger og kritik. Nedenfor forholder jeg mig sammenfattende til, hvad jeg mener, der er det væsentligste i den fremsatte kritik.

Svend Kreiner (SK) starter med en beskrivelse af den proces, der gik i gang efter Bundsgaard og Kreiner (2019) viste at de estimater af sværhedsgraderne i læsning i 8.klasse, som blev benyttet, når opgaverne i de adaptive DNT blev valgt til den enkelte elev, afveg signifikant fra estimaterne af samme i 2017. Dette betød, at målingen af læsefærdighed i 8. klasse var behæftet med fejl i en eller anden grad, formentlig både usystematiske og systematiske. Når elevers dygtighed i læsning i 8. klasse således blev beregnet med en ukendt grad af fejl havde det helt naturligt den konsekvens at den enkelte elevs resultat ikke kunne bruges formativt, da det formative grundlag var usikkert – hvor usikkert var ikke kendt – og statistiske opgørelser og sammenligninger ville være usikre i ukendt grad, og måske også systematisk skævvredet? Analyserne lavet af Bundsgaard og Kreiner (2019) kunne ikke vise, hvad årsagen til fejlen var, ej heller om det var et problem blot for læsefærdighed i 8. klasse eller en generel problemstilling. Bundsgaard og Kreiner (2019) konkluderede derfor, at det var væsentligt at få undersøgt, om det var et lille problem isoleret til et enkelt testområde eller et større generelt problem i DNT, således at det kunne løses ift DNTs anvendelse i folkeskolen. En sådan undersøgelse ville afhængigt af resultatet også kunne give anledning til, at forskere kunne overveje, om og i hvilken grad deres forskningsresultater kunne være påvirket, hvis der var tale om et generelt problem.

At undersøge hvor stort et problem, der var tale om, tog Børne- og Undervisningsministeriet til sig, og i en rapport fra STIL (2020) blev det vist, at problemet, som blev identificeret af Bundsgaard og Kreiner, viste sig at være et problem i alle test i DNT og havde været det siden 2010, og det blev fundet nødvendigt at genberegne resultaterne på basis af korrekte sværhedsgrader for alle fagområder fra 2014 og frem. Det er efter min mening prisværdigt, at ministeriet har taget kritikken til sig og har valgt at undersøge omfanget meget grundigt. At de tillige fremlægger resultaterne, som viser, at der var fejl, og at det var en generel problematik, er også prisværdigt. Desværre er der udover genberegning af testresultaterne baseret på korrekte sværhedsgrader også blevet lagt delområder sammen inden for et fag/testområde. Delområder, som teoretisk set er adskilte og derfor ikke bør lægges sammen til et enkelt validt mål. Dermed har STIL i nogle testområder på sin vis begået den samme fejl som i rapporten af Andersen og Nielsen (2016).

At efterprøve forskningsresultater ved hjælp af de genberegnede testresultater, har der tilsyneladende ikke været interesse for at gøre. I hvert fald er det ifølge Svend Kreiner ikke sket for tre af de fire studier, som han forholder sig til, nemlig Andersen & Nielsen (2016), Andersen, Beuchert, Nielsen & Thomsen (2020) og Beuchert & Nielsen (2018).

Testresultaterne fra *før* 2015 er tilsyneladende ikke genberegnet. Denne oplysning fra STIL udvider, som jeg ser det, Svend Kreiners kritik af Holm, Fallesen & Heinesens (2023) rapport, idet forskerne bag denne rapport ikke vil kunne gentage deres analyser med genberegnede og korrekte DNT-data, da disse kun findes fra 2015 og frem, og ikke for perioden 2010-2014, som også indgår i den samlede analyse i rapporten.

At der ikke er sket en genberegning af testresultaterne fra før 2015, er naturligvis uheldigt. Der er, for mig at se, dog ikke noget der forhindrer ministeriet i at råde bod på dette ved at undlade at udlevere data fra før 2015. Dette ville have kunnet forhindre, at der i rapporten af Holm, Fallesen & Heinesen (2023) er gennemført en analyse af DNT-data som ikke burde indgå i en og samme analyse. Dermed ville den nylige debat om disse resultater være undgået – eller i hvert fald en del af den.

SK fremsætter desuden en kritik af de analyser af DNT-data der gennemføres i førnævnte fire rapporter. Kritikken er, efter min faglige overbevisning, berettiget. Jeg mener i princippet, at forskere skal vælge de analysemetoder, de finder egnede til at besvare de forskningsspørgsmål eller hypoteser, som de undersøger. Arbejdes der med testresultater er det vigtigt validiteten og objektiviteten af testresultaterne ikke smides ud med badevandet, og at det er muligt at tolke resultaterne meningsfuldt og brugbart. Jeg er selv af uddannelse psykolog og specialiseret i pædagogisk psykologi og psykometri gennem både min forskning og min praksis. Jeg har, i mit eget arbejde, altid stræbt efter at vælge den bedst egnede/mest passende statistiske metode. Det er klart, at jeg qua min faglighed, har begrænsninger i den henseende. Dette har dog ført til god læring fra forskellige statistikere og ikke mindst gode samarbejder, når jeg har haft brug for at supplere min egen statistiske kunnen. At opdage at der var forhold, der kunne have betydning for resultaterne af min forskning, ville bestemt gøre mig nysgerrig på at undersøge, om det var tilfældet ved at genanalysere og evt. korrigere resultaterne. Jeg mener, at vi er nødt til at arbejde sådan – det er den måde vi udvider videnskaben og skaber reel viden. For at vende tilbage til kritikken omkring anvendte analysemetoder, så er er det dels helt fundamentale statistiske principper som SK mener ikke er opfyldt, dels psykometriske forhold der peger på at validiteten og objektiviteten i de ellers vel-validerede og objektive målinger af elev-dygtighed sættes over styr i de metoder, der anvendes. Det kan jeg kun være enig i.

Afslutningsvis, vil jeg gentage, at jeg mener at det mest alvorlige i Svend Kreiners kritik er, at der er publicerede forskningsresultater, som er baseret på, hvad vi ved er fejlagtige (både usystematisk og systematisk fejlagtige) målinger af elev-dygtigheder. Disse resultater bør undersøges for om genberegningerne gør en forskel, således at de ikke fortsat kan have politiske, bevillingsmæssige eller faglige konsekvenser i samfund og forskning, hvis det er tilfældet. Det bør slås fast om de kan stå til troende. Desuden er der tilsyneladende (mindst) en rapport, som bygger på usammenlignelige data ifølge STIL. Konsekvensen af dette bør stå klart for enhver.

Sidst vil jeg gerne sige åbent, at jeg har adskillige tidligere faglige samarbejder med Svend Kreiner. Dette betyder dog ikke, at jeg er inhabil ift at udtale mig om hans kritik eller at jeg blot vil være enig af den grund. Faglighed og fagfælle-kritik er alfa og omega i videnskab og forskning, og i den henseende behandler jeg alles arbejde ens.

Referencer

Andersen, S.C. & Nielsen, H.S. (2016). The Positive Effects of Nationwide Testing on Student Achievement in a Low-Stakes System.

Andersen S.C., Beuchert, L. Nielsen, H.S & Thomsen M.K (2020) The Effect of Teacher's Aides in the Classroom: Evidence from a Randomized Trial. *Journal of the European Economic Association*, 18, 469-505.https://doi.org/10.1093/jeea/jvy048

Beuchert, L.V & Nandrup, A. B. (2018): The Danish National Tests at a Glance. *Nationaløkonomisk Tidsskrift, 1,* 1-37.

Holm, M.L., Fallesen, P. & Heinesen, E. (2023) The Effect of parental Union Dissolution on Children's Test Scores. *ROCKWOOL Fondens Forskningsenhed*, Study Paper nr. 185

Kreiner, S. (2023). Om statistiske analyser af resultater fra pædagogiske test. Fagbladet *Folkeskolen*

STIL (2020) Evaluering af de statistiske aspekter ved de nationale test. STYRELSEN FOR IT OG LÆRING. Børne- og Undervisningsministeriet.